University of Cape Town

Masters Dissertation

Detection and down-weighting of outliers in non-normal data: Theory and Application

Author and supervisors

Author:
Tinashe Chatora

Supervisors:
Dr Freedom Gumedze
A/Prof Francesca Little
Professor Linda Haines

February 24, 2013

Contents

1 Introduction **1**

1.1 Datasets . . . 3

1.1.1 Fabric dataset . . . 4

1.1.2 Epilepsy dataset . . . 5

1.1.3 Leukemia rats dataset . . . 6

1.1.4 Seeds germination dataset . . . 7

1.1.5 Contagious bovine pleuropneumonia dataset . . . 8

2 Review of GLMMs and HGLMs **10**

2.1 Linear Mixed Models . . . 12

2.2 Generalized Linear Mixed Models . . . 13

2.2.1 Model Formulation . . . 13

2.2.2 Marginal Properties . . . 14

2.2.2.1 Mean of Y_{ij} . . . 14

2.2.2.2 Variance of Y_{ij} . . . 15

2.2.3 Estimation . . . 15

2.2.3.1 Maximum likelihood estimation . . . 15

2.2.3.2 Empirical Bayes estimation . . . 17

2.2.3.3 Bayesian methods . . . 17

2.2.4 Inference . . . 18

2.3 Hierarchical Generalized Linear Models . . . 18

3 Review of model diagnostics in non-normal data **22**

3.1 Aims of model diagnostics . . . 22

3.2 Model diagnostics for independent non-normal data . . . 23

3.3 Model diagnostics for longitudinal non-normal data 26
3.4 The variance shift outlier model (VSOM) 29

4 VSOM for normal data **31**
4.1 A VSOM for independent normally distributed data 31
4.2 A VSOM for longitudinal normally distributed data 34
4.3 Applying the VSOM using R 37

5 VSOM for count data **39**
5.1 Poisson regression models 39
5.1.1 Poisson GLM 40
5.1.2 Overdispersed count data 40
5.2 Variance shift outlier model (VSOM) for Poisson count data . 42
5.2.1 A VSOM for independent count data 43
5.2.2 A VSOM for outlying observations in longitudinal count data 44
5.2.3 A VSOM for outlying subjects in clustered count data 47
5.3 Poisson-Normal VSOM 49
5.4 Hypothesis tests on variance shift parameters 50
5.5 Examples 51
5.5.1 Fabric dataset 52
5.5.2 Epilepsy dataset 58
5.5.2.1 VSOM for individual observations 58
5.5.2.2 VSOM for subjects 61
5.5.3 Leukemia rats dataset 66
5.5.3.1 VSOM for individual observations 66
5.5.3.2 VSOM for subjects 68

6 VSOM for binomial data **73**
6.1 Binomial regression models 75
6.1.1 Binomial GLM 75
6.1.2 Quasi-binomial model 76
6.1.3 Beta-binomial model 77
6.2 Variance shift outlier model (VSOM) for binomial data . . . 79
6.2.1 A VSOM for independent binomial data 79

6.2.1.1 A VSOM with no covariates 79
6.2.1.2 A VSOM with covariates 81
6.2.2 A VSOM for longitudinal binomial data 82
6.2.2.1 A VSOM for outlying subjects in longitudinal binomial data 83
6.3 Examples 84
6.3.1 Seeds germination dataset 84
6.3.2 Contagious bovine pleuropneumonia dataset 88

7 Cytokine dataset **94**
7.1 Data exploration 96
7.2 Cytokine data analyzed as normally distributed responses . . 104
7.3 Cytokine dataset analyzed as counts 113
7.3.1 Quasi-Poisson-gamma HGLM 113
7.3.2 Negative binomial HGLM 119
7.4 Summary 125

8 Conclusion **127**

9 References **130**

10 Appendix **137**
10.1 VSOM code for GENSTAT and R 137
10.1.1 Cytokine linear mixed model R code 137
10.1.1.1 R code for subjects 140
10.1.2 Cytokine Count VSOM GENSTAT code 144
10.1.3 Updated seed germination dataset VSOM GENSTAT code 147

List of Tables

2.1 **Examples of conjugate HGLMs** 21

5.1 **Table of marginal variances for Y_i (independent data) and Y_{ij} (longitudinal data) for different models.** 49
5.2 **Parameter estimates of models fitted to the fabric dataset.** 56
5.3 **Parameter estimates of models fitted to the epilepsy dataset.** 60
5.4 **Parameter estimates of models fitted to the leukemia rats dataset.** 67

6.1 **Parameter estimates of models fitted to the seeds germination dataset.** 85
6.2 **Parameter estimates of models fitted to the adjusted seeds germination dataset.** 87
6.3 **Parameter estimates of models fitted to the CBPP dataset.** 90

7.1 **Parameter estimates of models fitted to the log-cytokine dataset assuming an underlying normal distribution.** . 109
7.2 **Parameter estimates of quasi-Poisson-gamma models fitted to the count cytokine dataset.** 116
7.3 **Parameter estimates of models fitted to the count cytokine data using a negative binomial HGLM null model.** 122

List of Figures

1.1 **Scatterplot of the number of fabric faults against the logarithm of the fabric length (x).** 4
1.2 **Scatterplot of the number of seizures against time by treatment group.** 6
1.3 **Scatterplot of the number of cancer colonies in rats against time by treatment group.** 7
1.4 **Scatterplot of the proportion of seeds which germinated grouped by species.** 8
1.5 **Scatterplot of the proportion of the herd infected over the periods of investigation.** 9

5.1 **Residual plots for model M_{F1} from the fabric dataset.** 54
5.2 **VSOM statistics plotted against observation number for the fabric dataset.** (a) Variance shift estimates, λ_i. (b) Dispersion parameter estimates, α_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th} and 97.5^{th} percentile cut-off values. 57
5.3 **Residual plots for model M_1 from the Epilepsy dataset.** 59
5.4 **VSOM statistics plotted against observation number for the epilepsy dataset.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, ϕ_k. (c) Likelihood ratio statistics, LRT_k, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values. 62

5.5 **VSOM statistics plotted against patient number for the epilepsy dataset.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, ϕ_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values. 64

5.6 **VSOM statistics plotted against observation number for the epilepsy dataset.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, ϕ_k. (c) Likelihood ratio statistics, LRT_k, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values. 65

5.7 **Residual plots for observations from model M_{R0} from the leukemia rats dataset.** 69

5.8 **VSOM statistics plotted against observation number for the leukemia rats dataset.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, ϕ_k. (c) Likelihood ratio statistics, LRT_k, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values. 70

5.9 **VSOM statistics plotted against subject number for the leukemia rat dataset.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, ϕ_i. (c) Likelihood ratio statistics, LRT_i, with $95^{th}, 97.5^{th}$ and 99^{th} percentile cut-off values. 72

6.1 **VSOM statistics plotted against observation number for the seeds dataset.** (a) Variance shift estimates, λ_i. (b) Dispersion parameter γ_i. (c) Likelihood ratio statistics, LRT_i, with a 95^{th} percentile cut-off value. 86

6.2 **Plot of the absolute residuals against fitted values for model M_{S0} from the seeds dataset.** 87

6.3 **VSOM statistics plotted against observation number for the adjusted seeds dataset.** (a) Variance shift estimates, λ_i. (b) Dispersion parameter γ_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values. 88

6.4 **VSOM statistics plotted against observation number for the CBPP dataset.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter for the herd random effect γ_k. (c) Likelihood ratio statistics, LRT_k, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values. 91
6.5 **VSOM statistics plotted against herd number for the CBPP dataset.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter for the herd random effect γ_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values. 92
6.6 **Plot of absolute residuals against fitted values for individual observations using model M_{H0} in the CBPP dataset.** 93

7.1 **Histograms of the cd4:inf+tnf+il2+ counts and the logarithm of the cd4:inf+tnf+il2+ counts.** 97
7.2 **Histograms of the cd4:inf+tnf+il2+ counts by treatment group**. 98
7.3 **Boxplots of the cd4:inf+tnf+il2+ counts by groups**. . 98
7.4 **Histograms of the logarithm of the cd4:inf+tnf+il2+ counts against time by treatment group**. 99
7.5 **Boxplots of the logarithm of the cd4:inf+tnf+il2+ counts by groups**. 100
7.6 **Mean profiles of the cd4:inf+tnf+il2+ counts and the logarithm of the cd4:inf+tnf+il2+ counts.** 101
7.7 **Boxplots of the cd4:inf+tnf+il2+ counts against treatment group by time.** 101
7.8 **Boxplots of the logarithm of the cd4:inf+tnf+il2+ counts against treatment group by time.** 102
7.9 **Scatterplot of the cd4:inf+tnf+il2+ counts against by treatment group.** 103
7.10 **Scatterplot of the logarithm of the cd4:inf+tnf+il2+ counts against time by treatment group.** 103

7.11 **VSOM statistics plotted against observation number for the log-cytokine dataset.** (a) Variance shift estimates, ω_k. (b) Residual variance estimates, σ^2. (c) Likelihood ratio statistics, LRT_k, with 95th percentile of the empirical distribution under the null hypothesis shown for the first r order statistics for each test: $r = 4$ (red line), $r = 5$ (green line) and $r = 6$ (blue line). 106

7.12 **VSOM statistics plotted against subject number for the log-cytokine dataset.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, σ^2. (c) Likelihood ratio statistics, LRT_i, with 95th percentile of the empirical distribution under the null hypothesis shown for the first r order statistics for each test: $r = 2$ (red line), $r = 3$ (green line) and $r = 4$ (blue line). 110

7.13 **Histogram of standardized error residuals for the log-cytokine dataset.** 111

7.14 **Scatterplot of standardized residuals for the log-cytokine dataset.** 111

7.15 **Histogram of standardized subject residuals for the log-cytokine dataset.** 112

7.16 **Scatterplot of standardized subject residuals for the log-cytokine dataset.** 112

7.17 **VSOM statistics plotted against observations for the cytokine count dataset analyzed using a quasi-Poisson-gamma model.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, ϕ_k. (c) Likelihood ratio statistics, LRT_k, with $95^{th}, 97.5^{th}$ and 99^{th} percentile cut-off values. . . 115

7.18 **VSOM statistics plotted against patient number for the cytokine count dataset analyzed using a quasi-Poisson-gamma model.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, ϕ_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values. 118

7.19 **Absolute residuals against fitted values plot for individual observations from the cytokine count dataset using the quasi-Poisson-gamma HGLM as the null model** 119
7.20 **VSOM statistics plotted against observations for the cytokine count dataset analyzed using a negative binomial HGLM.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, α_k. (c) Likelihood ratio statistics, LRT_k, with $95^{th}, 97.5^{th}$ and 99^{th} percentile cut-off values. . . 121
7.21 **VSOM statistics plotted against patient number for the cytokine count dataset analyzed using a negative binomial HGLM.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, α_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values. 124
7.22 **Absolute residuals against fitted values plot for individual observations from the cytokine count dataset using the negative binomial HGLM as the null model** 125

Acknowledgments

My special heartfelt thanks to Dr Freedom Gumedze of the University of Cape Town for his constant input, supervision and mentorship throughout all aspects of this research. I wish to also thank Professor Linda Haines and Associate Professor Francesca Little of the University of Cape Town for their guidance and input throughout the course of this study as well as for providing me with opportunities to progress in my academic career.

Thanks also go out to the staff and postgraduate students of the Statistical Science department at the University of Cape Town for all the assistance and knowledge that they passed on to me throughout my tenure at the University.

I would also like to thank the South African Tuberculosis Vaccine Initiative and the Institute of Infectious Diseases and Molecular Medicine for use of the cytokine dataset.

I owe a debt of gratitude to the National Research Foundation for their financial assistance which made this research possible.

Finally I would like to thank my family and friends for all the support that they provided me throughout the course of my studies. This especially applies to my father (Noah) and brothers (Godfrey and Tawanda) whose support both financially and emotionally was unwavering throughout my studies; for this and more I am eternally grateful. I dedicate this thesis to my mother, you will always be in my prayers.

PUBLICATION

I hereby grant the University free license to publish this dissertation in whole or part in any format the University deems fit.

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use anothers work and pretend that it is my own.

2. I have used the APA referencing guide for citation and referencing. Each contribution to, and quotation in this dissertation from the work(s) of other people has been contributed, and has been cited and referenced.

3. I know the meaning of plagiarism and declare that all of the work in the dissertation, save for that which is properly acknowledged, is my own.

Signature:

Date:

Abstract

This thesis introduces a variance shift outlier model (VSOM) for independent count and binomial data. The VSOM will be used for the detection and down-weighting of outliers. This model considers outliers as observations with inflated variances and uses random effects to model the overdispersion associated with a single observation or a group of observations.

For count data a VSOM is formulated assuming an underlying negative binomial distribution. This is done by assuming a Poisson distribution for all observations, with a gamma distributed random effect for the i^{th} observation. The status of the i^{th} observation as an outlier is indicated by the size of the associated shift in variance. This model is then extended to clustered count data.

The VSOM for binomial data is formulated by assuming a binomial distribution for all observations, with a beta distributed random effect for the i^{th} observation, resulting in an underlying beta-binomial distribution. It will be shown that the size of the associated shift in the variance of the i^{th} observation will determine the status of that observation as an outlier.

The variance shift outlier models (VSOMs) are illustrated using several published datasets, chosen for their specific underlying design and the presence of outlying observations. Finally all the different models are used to analyze a cytokine response to the BCG vaccine among a cohort of infants (Mansoor et al., 2009). The distribution of the cytokine response is typically very skew with a tail to the right. The cytokine response is treated as both a continuous (when a logarithmic transformation is applied to the response) and a count response (when the response is left in its original state). The treatment of outlying observations by the VSOM is then compared to other forms of model diagnostics.

Chapter 1

Introduction

In biological sciences research discrete data, including binary and count data, are very common. The Poisson model is generally used to model count data while the binomial model is used to model grouped binary data, which is also known as binomial data. These models are applied within the generalized linear modeling (GLM) framework. The GLM framework has been discussed by many authors including Wedderburn (1974), McCullagh and Nelder (1989) and Dobson and Barnett (2008).

The GLM framework has a key and highly restrictive feature when it comes to binomial and count data. The feature is that of the relationship between the mean and the variance of the data. This relationship implies that the variance of a random variable (Y) must be a deterministic function of its mean (μ). For Poisson data the relationship is $\text{var}(Y) = \mu$. When the data are overdispersed then $\text{var}(Y) = \phi\mu$ where ϕ, the dispersion parameter, is greater than 1. For binomial data the mean-variance relationship is given as $\text{var}(Y) = n\pi(1-\pi)$, where $\mu = n\pi$ and π is the probability of success, for example the probability of having a disease. Williams (1982) showed that when binomial data are overdispersed, $\text{var}(Y) = n\pi(1-\pi)[1+(n-1)\phi]$ where $\phi \geq 0$. Model diagnostics in GLMs have also been well studied by many authors including McCullagh and Nelder (1989).

When the data are clustered, generalized linear mixed models (GLMMs) are used. GLMMs are extensions of linear mixed models. They are used to analyze data with non-normal responses from the exponential family of dis-

tributions with repeated measurements and other forms of clustered data. GLMMs account for multiple sources of variation and address various correlation structures in the data by adding a random component to the linear predictor in a GLM. The random effects are specified to explicitly handle dependency and overdispersion found in longitudinal studies and in data involving hierarchical design structures. The random effects are also assumed to be normally distributed. The main methods of parameter estimation, used for GLMMs, are maximum likelihood (Schall, 1991), Bayesian and Empirical Bayes estimation (Molenberghs and Verbeke, 2005) methods. Since parameter estimation is based on maximum likelihood principles the estimates are asymptotically normally distributed, thus classical Wald-type tests can be used for inference on model parameters.

An extension of GLMMs is hierarchical generalized linear models (HGLMs) (Lee and Nelder, 1996). HGLMs can be considered as generalizations of GLMMs. They relax the normality assumption of the random effects in GLMMs, and thus allow the random effects to follow any distribution from the exponential family of distributions. The estimation of the random and fixed parameters is based on the joint maximization of the hierarchical log-likelihood (Lee and Nelder, 1996).

Model diagnostics in GLMMs and HGLMs have received limited attention in the literature. The main types of model diagnostics used in GLMMs are case deletion (Xu et al., 2006) and local influence analysis (Zhu and Lee, 2003) (Xiang et al., 2002).

Another approach to model diagnostics in linear regression is called the variance shift model (Cook and Weisberg, 1982). A variance shift model considers outliers as observations with an inflated variance, with the status of the i^{th} observation as an outlier indicated by the size of its associated shift in variance. Thompson (1985) used a restricted maximum likelihood approach in applying this model. Gumedze et al. (2010) extended the work on the variance shift model by formulating it as a linear mixed model and also proposed several tests for the detection and down-weighting of outliers. They called the model the variance shift outlier model (VSOM).

The aim of this thesis is to extend the variance shift outlier model (VSOM) to independent and clustered binomial and count data. These

extensions will be illustrated using datasets, chosen for their specific underlying design and the presence of outlying observations. The different VSOMs will also be compared using data relating to a cytokine response to the BCG vaccine (Mansoor et al., 2009). The key features of this dataset are that it is a longitudinal dataset whose responses are very skew and overdispersed. The data can be transformed, thus making the response normally distributed and allowing a linear mixed model to be used on the data. Another possible approach to analyzing the data is to consider it as counts, and thus analyze it using a quasi-Poisson-gamma HGLM or a negative binomial HGLM. In this thesis the VSOM will be applied to all these forms of the data.

The thesis is organized as follows. The various datasets are introduced, then the generalized linear mixed model (GLMM) and hierarchical generalized linear model (HGLM) are defined in chapter 2. Chapter 3 reviews existing methods of model diagnostics for non-normal data in both independent and longitudinal data. The variance shift outlier model (VSOM) methodology for normal data (Gumedze et al., 2010) is outlined in chapter 4. The VSOM is extended to accommodate count data in chapter 5 and further extended to binomial data in chapter 6. In chapter 7 the various forms of the VSOM are applied to the cytokine dataset. Finally chapter 8 gives a synthesis of the study, whereby the findings are summarized and conclusions are derived from the preceding chapters. Areas of further research using the the VSOM will also be discussed in chapter 8.

1.1 Datasets

The datasets used in this thesis are introduced in this section. These datasets were chosen primarily because of the presence of outlying observations/subjects in the data. They include either count or binomial responses and have either single or repeated measures per observation. The VSOM will be applied to these datasets in order to detect and down-weight these outliers.

1.1.1 Fabric dataset

This dataset is taken from Bissell (1972). It involves the number of faults (y) in a bolt of fabric of length l. The covariate of interest in this dataset is the logarithmic transformation of l, that is $\log(l) = x$. The sample mean of y is 8.875 which is considerably smaller than the sample variance of 33.79, thus indicating that the data are overdispersed.

A scatterplot of y against x (Figure 1.1) reveals that there are observations which are possibly outlying in the data. The potential outliers are observations 13 and 19 which have higher numbers of faults for the given values of x as compared to the rest of the observations. On the other end of the scale observations 30 and 32 have slightly low numbers of faults for the given values of x.

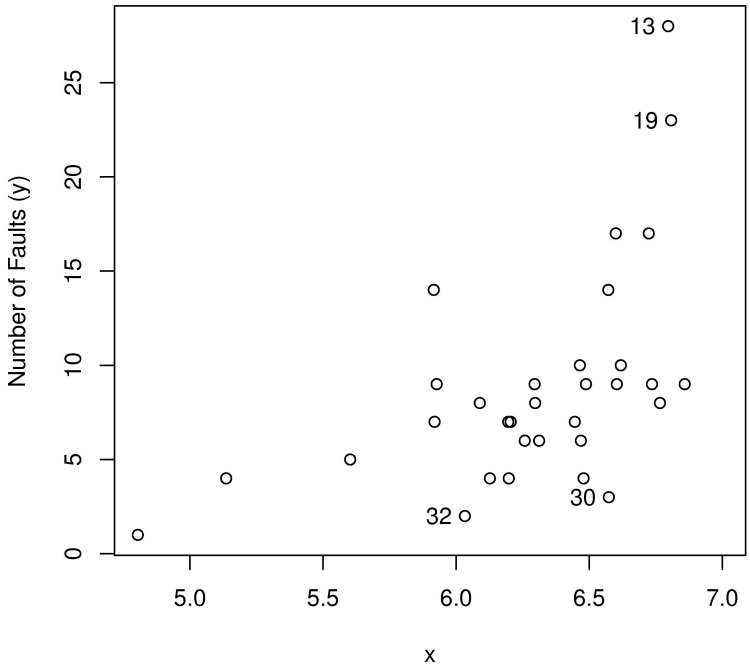


Figure 1.1: **Scatterplot of the number of fabric faults against the logarithm of the fabric length (x).**

This dataset will be used to illustrate the VSOM for overdispersed independent count data. The null model used for this dataset is the negative binomial model which is formulated as a Poisson-gamma HGLM with satu-

rated random effects (Lee et al., 2006).

1.1.2 Epilepsy dataset

Thall and Vail (1990) presented longitudinal data from a clinical trial of 59 epileptics who were randomized to a new drug or a placebo (Treatment = 0 or Treatment = 1). The data included the logarithm of a quarter of the number of epileptic seizures recorded in the 8 week period preceding the trial (lbase). Also the logarithm of age was recorded as well as a linear trend (visit) coded as (-0.3,-0.1,0.1,0.3). The multivariate response variable consisted of the seizure counts during 2-week periods before each of four visits to the clinic. This dataset was found to be overdispersed by some authors, including Breslow and Clayton (1993), Thall and Vail (1990) and Lee and Nelder (1996), who fitted various Poisson HGLMs to the data.

Figure 1.2 shows a scatterplot of the number of seizures for each subject over time, grouped by treatment type. It can be seen from this plot that the third observation of subject 25 stands out as a single outlier, while the entire profile for subject 49 is an example of a cluster of outlying values associated with a single subject.

This dataset is chosen as an example of longitudinal count data where the VSOM is used to identify and down-weight, both outlying observations and subjects. The quasi-Poisson-normal model is used as the null model for this dataset.

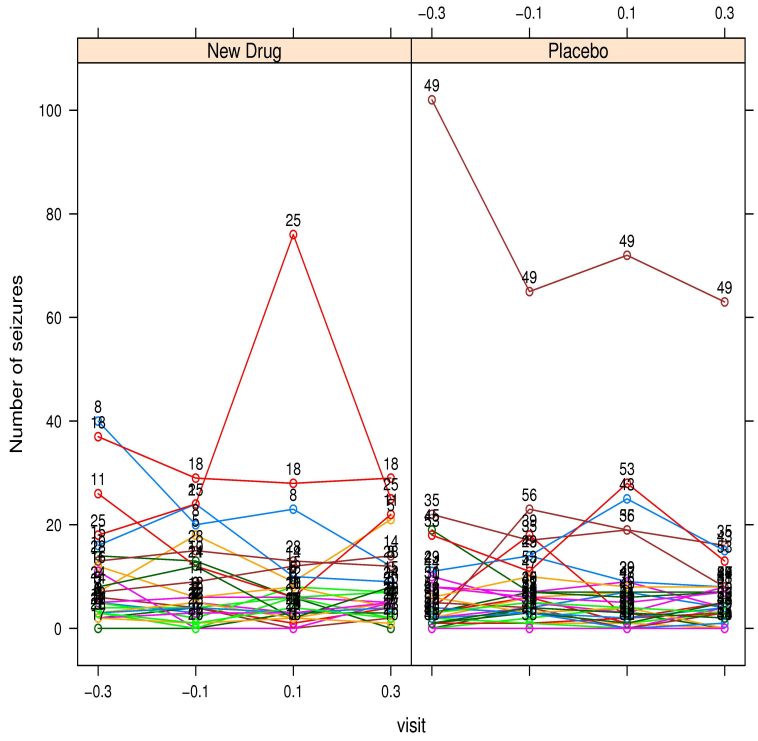


Figure 1.2: **Scatterplot of the number of seizures against time by treatment group.**

1.1.3 Leukemia rats dataset

This dataset is taken from Myers et al. (2002). In this dataset three chemotherapy drugs were given to 30 rats that had an induced leukemic condition. The response of interest was the number of cancer colonies in the rats. The covariates that were collected were the white and red blood cell counts. The data was collected from each rat at four different time periods. The choice of drug administered to the rats was also considered as a covariate, specifically it was a between-rat covariate while the blood cell counts were within-rat covariates.

From a scatterplot of the response profiles (Figure 1.3) it was identified that there were several subjects which could potentially be outliers.

This dataset was chosen as another example of longitudinal count data where the VSOM was used to identify and down-weight outlying obser-

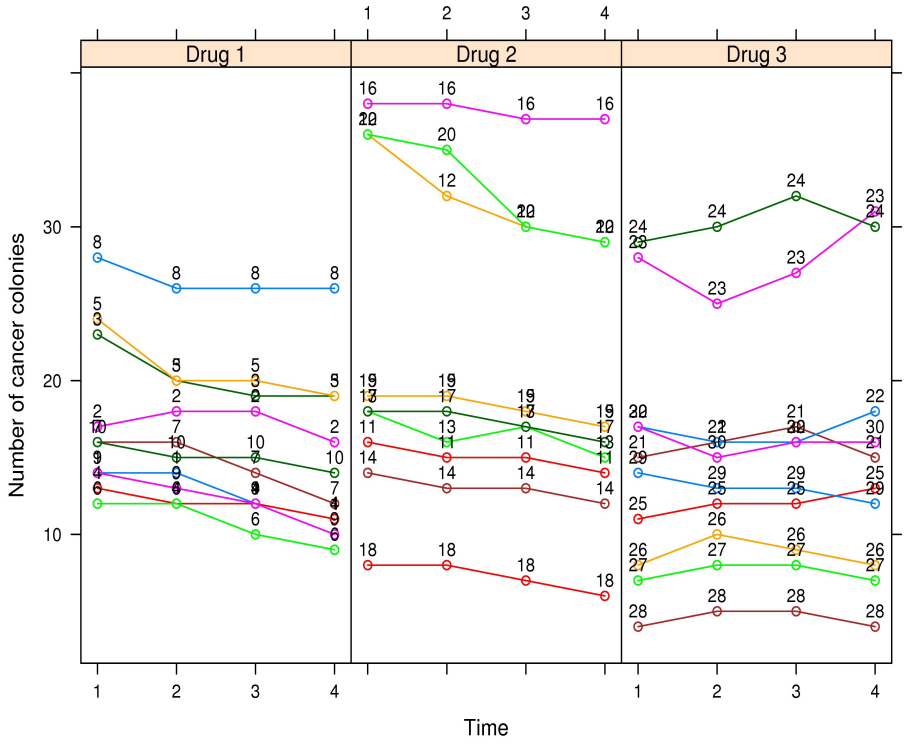


Figure 1.3: **Scatterplot of the number of cancer colonies in rats against time by treatment group.**

vations and subjects. In this example however, the quasi-Poisson-gamma model was used as the null model for this dataset. Thus the random effects were assumed to follow a gamma distribution, as opposed to the assumption of normally distributed random effects used in the epilepsy data analysis.

1.1.4 Seeds germination dataset

This dataset is taken from Crowder (1978). The data involves the number of seeds which germinated from two varieties of the Orobanche aegyptiaca species, that is the Orobanche aegyptiaca 75 (o75) and Orobanche aegyptiaca 73 (o73) varieties. These species varieties were brushed onto a plate containing a 1/125 dilution of an extract prepared from roots of either a bean or cucumber plant. The response of interest in this study is the proportion of seeds that germinate given certain conditions, thus this is an independent binomial study. A scatterplot of the proportion of seeds that

germinate grouped by species is shown in Figure 1.4. This plot shows that observations 6, 16 and 17 are potentially outlying. Collett (1996), page 203, showed that this data was overdispersed. As a result, the null model used for this dataset was the beta-binomial model which is able to accommodate overdispersion.

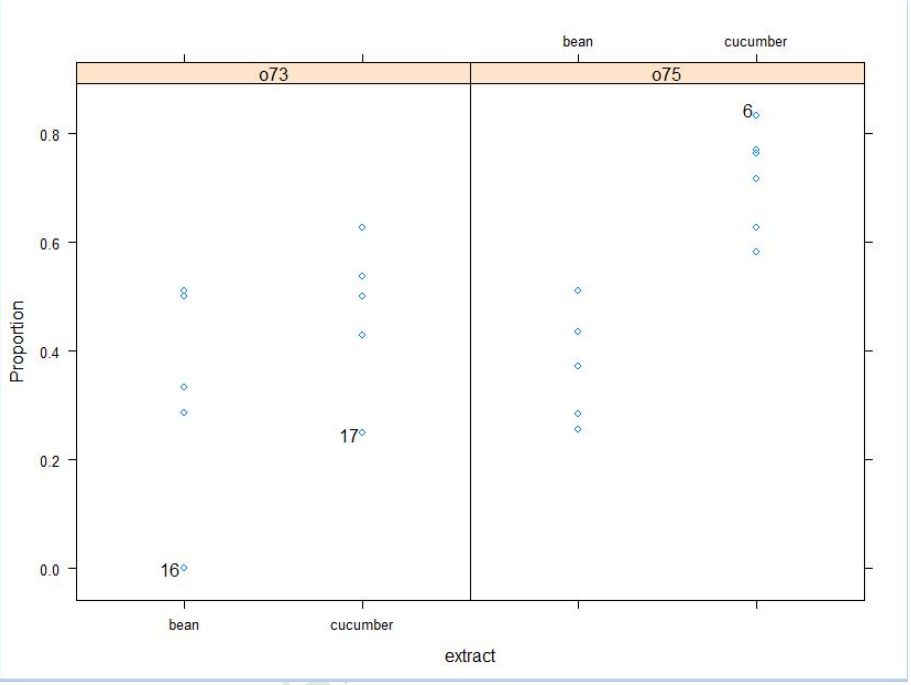


Figure 1.4: **Scatterplot of the proportion of seeds which germinated grouped by species.**

1.1.5 Contagious bovine pleuropneumonia dataset

Contagious bovine pleuropneumonia (CBPP) is a major disease which affects the health and production of cattle in Africa. This disease is caused by mycoides which are subspecies of a small colony type of mycoplasma mycoides. The transmission of CBPP occurs due to direct and repeated contact between infected and healthy cattle. This dataset is taken from Lesnoff et al. (2004) and it is a serological and clinical incidence study of CBPP in zebu cattle, which was implemented in 15 newly infected herds located in the Boji district of Ethiopia. The aim of the study was to in-

vestigate the within-herd spread of CBPP in newly infected herds. Blood samples were quarterly collected from all animals of these herds to determine their CBPP status. These data were used to compute the serological incidence of CBPP, that is new cases occurring during a given time period. This is thus a binomial longitudinal study.

A scatterplot of the proportion of the herd infected over the periods of investigation (Figure 1.5) shows that there are potentially outlying observations and subjects. Specifically subjects 1 and 14 seem to be potentially outlying.

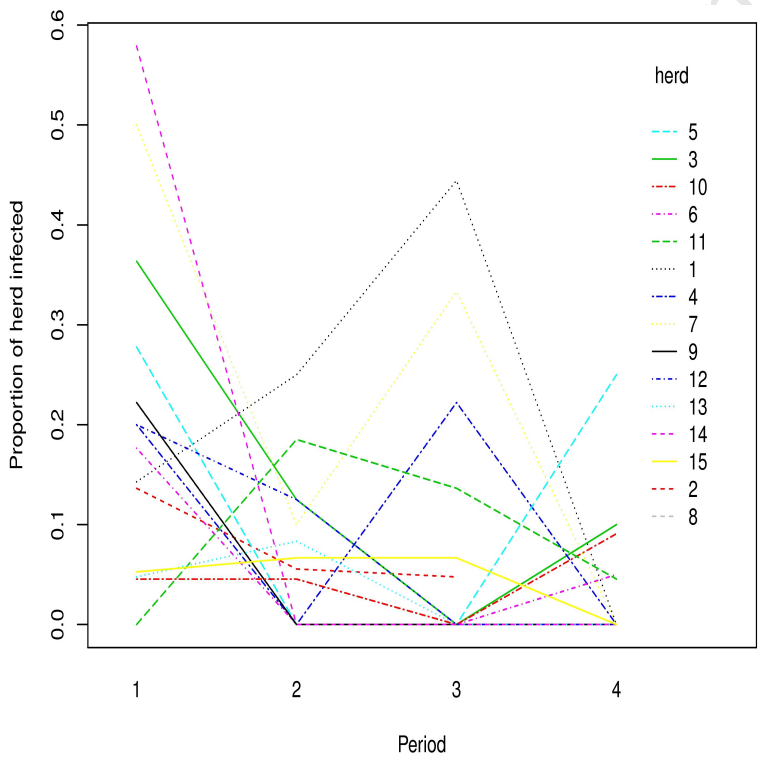


Figure 1.5: **Scatterplot of the proportion of the herd infected over the periods of investigation.**

This dataset is an example of longitudinal binomial data where proportions are clustered by herd. The VSOM was used to identify and downweight outlying observations and herds. The binomial-normal HGLM and beta-binomial HGLM were used as the null models for this example in order to show the versatility of the VSOM.

Chapter 2

Review of GLMMs and HGLMs

In biological sciences research both count and binomial data are common. The data are usually in a longitudinal/clustered [1] data structure, that is data which has been collected on a particular subject/cluster [2] multiple times. In clustered data the usual assumption about the independence of observations, which is used in classical statistical tests, is not upheld. In such data there are two sources of variation, which are the within-cluster variation and the between-cluster variation. It is typically assumed that observations within a cluster are correlated because they are more homogeneous than observations from different clusters, thus leading to within-cluster dependence. Also different clusters can be considered to be heterogeneous. This is because they differ systematically, thus leading to between-cluster heterogeneity (Collett, 1996). There are various complicated forms of within-cluster correlation structures which may exist, for example an AR1 correlation structure. In this study a simple constant correlation structure will be used.

A naive approach to analyzing clustered data would be to ignore the between-cluster heterogeneity and within-cluster dependence. This approach thus assumes that the data are independent and leads to the estimation of parameters that are common to all the clusters. There are some downfalls

[1] In this thesis longitudinal and clustered will be used interchangeably.

[2] In this thesis cluster and subject will be used interchangeably.

which arise from using this approach. For example the estimated parameters will most likely be biased and the standard errors will be incorrect, as the assumption of independence within the population will not be true. Another downfall is that no information is provided about the between-cluster heterogeneity (Tuerlinckx et al., 2006).

Another naive approach would be to analyze each cluster individually, thus upholding the assumption of independence of observations between clusters. A shortcoming of this approach is that the common effect of variables across the clusters is not analyzed. Also if the clusters have very few observations, the standard errors of the estimated parameters will be high and in some cases the parameters cannot be estimated.

These two approaches can be combined so that the statistical model which is built contains parameters common to all clusters, as well as parameters which are specific to a cluster.

Mixed models assume that cluster-specific effects are a random sample from the population distribution of such effects. These cluster-specific effects are called random effects and the effects which are common to all the clusters are fixed effects. When clustered data are assumed to be normally distributed, linear mixed models (LMMs) are used to analyze the data (Laird and Ware, 1982). A generalization of these models is generalized linear mixed models (GLMMs). GLMMs allow the distribution of the data to follow any distribution from the exponential family of distributions, whilst the random effects are assumed to follow a normal distribution (McCulloch and Searle, 2001). In cases where the random effects are assumed to follow non-normal distributions, hierarchical generalized linear models (HGLMs) may be used (Lee and Nelder, 1996).

Another approach to handling clustered data is the generalized estimating equation (GEE) approach. The GEE approach is used to analyze clustered data without introducing cluster-specific effects (Liang and Zeger, 1986). In GEEs a model is proposed for the expected values of observations and a working correlation matrix is defined for the observations within a cluster. The GEE approach is robust to the specification of the working correlation matrix, thus guaranteeing consistent and asymptotically normally distributed estimates for the regression parameters, regardless of whether

the correlation matrix has been misspecified or not. Another advantage is that this approach can be used on any data which follows the exponential family of distributions. There are also a variety of working correlation matrices which can be used in GEEs, thus allowing a researcher to have some freedom in model selection. The estimation procedure is simple to carry out and it is done using the iterative reweighted least squares (IRLS) method (Demidenko, 2004).

There are some disadvantages to using this method. Such as, there is no likelihood generated by GEEs thus making it impossible to assess the adequacy of the model during model selection (Tuerlinckx et al., 2006). Also GEEs do not allow researchers to assess cluster-specific effects, thus reducing the efficiency of subjects used in studies.

In this thesis I will be using GLMMs and HGLMs. As a result, for the remainder of this chapter I will proceed to outline the model formulation, estimation procedures and inference for GLMMs and HGLMs.

2.1 Linear Mixed Models

In order to provide some background I will briefly describe the linear mixed model (LMM) before proceeding to describe in more detail its extension, that is the generalized linear mixed model (GLMM).

In describing the LMM I will consider the simple case whereby subjects are involved in a longitudinal study, thus the only random effects considered are due to the subjects. Given that there are q subjects and the number of observations for the i^{th} subject is n_i (for $i = 1, ..., q$) with the total number of observations given as $n = \sum_{i=1}^{q} n_i$, the LMM can be written in two forms. Firstly, for the i^{th} subject the form of the LMM is given as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \tag{2.1}$$

for $i = 1, \ldots, q$, where $\mathbf{Y}_i$ is an n_i-dimensional vector of responses Y_{ij} for the i^{th} subject observed at the j^{th} occasion, $\mathbf{X}_i$ is an $n_i \times p$ design matrix of explanatory variables and $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effect coefficients, $\mathbf{Z}_i$ is a $n_i \times q$ design matrix describing the random effects $\mathbf{b}_i$ and $\mathbf{e}_i$ is a vector of residual errors. It is assumed that $\mathbf{e}_i \sim \mathrm{N}(\mathbf{0}, \sigma_e^2\mathbf{I}_{n_i})$

and $\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \sigma_b^2 \mathbf{I}_q)$, where $\mathbf{I}_{n_i}$ and $\mathbf{I}_q$ are identity matrices of order n_i and q respectively. The implication of $\mathbf{e}_i$ and $\mathbf{b}_i$ having their corresponding covariance matrices is that the residual errors and subject random effects, respectively, are independent for the i^{th} and k^{th} subject. It is also assumed that the residual error vector $\mathbf{e}_i$ is independent of the subject random effects $\mathbf{b}_i$. Overall observations due to the i^{th} subject will be independent of observations due to the k^{th} subject.

Another way of writing the model is by considering it at the j^{th} observation for the i^{th} subject. The model is then written as

$$Y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{b}_i + e_{ij}, \tag{2.2}$$

for $i = 1, ..., q$ and for $j = 1, ..., n_i$, where $\mathbf{x}_{ij}^{'}$ and $\mathbf{z}_{ij}^{'}$ are the j^{th} rows of the design matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ respectively. It is also assumed that $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$ and $\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \sigma_b^2 \mathbf{I}_q)$.

2.2 Generalized Linear Mixed Models

2.2.1 Model Formulation

Generalized linear mixed models (GLMMs) are an extension of linear mixed models (LMMs) as they allow the response variable, Y_{ij}, to follow any distribution from the exponential family of distributions. The conditional probability density function of members of the exponential family is dependent on $\mathbf{b}_i$ and is given as

$$f(y_{ij}|\mathbf{b}_i) = \exp[\phi^{-1}\{y_{ij}\theta_{ij} - c(\theta_{ij})\} + d(y_{ij}, \phi)] \tag{2.3}$$

where $c()$ and $d()$ are known functions, ϕ is the dispersion parameter and θ_{ij} is the canonical parameter which is defined implicitly by the mean of Y_{ij}, that is μ_{ij}, conditional on $\mathbf{b}_i$. Similarly to the LMM the random effects, $\mathbf{b}_i$, are assumed to follow a normal distribution.

The linear predictor for generalized linear mixed models is given by

$$g(\mathrm{E}(Y_{ij}|\mathbf{b}_i)) = g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{b}_i, \tag{2.4}$$

where $g()$ is the link function and η_{ij} is the linear predictor.

The expectation of Y_{ij} conditional on $\mathbf{b}_i$ is given by

$$\mathrm{E}(Y_{ij}|\mathbf{b}_i) = \mu_{ij} = g^{-1}\{\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{b}_i\}, \tag{2.5}$$

and the conditional variance is given by

$$\mathrm{var}(Y_{ij}|\mathbf{b}_i) = \phi V(\mu_{ij}), \tag{2.6}$$

where $V(\mu_{ij})$ is the variance function which depends on μ_{ij} and thus η_{ij}.

2.2.2 Marginal Properties

Given that the formulation of the GLMM is made conditionally on the value of $\mathbf{b}_i$ it is possible to derive the marginal properties of Y_{ij}. This has been done in great detail by McCulloch and Searle (2001), page 222. I will proceed to briefly outline the mean and variance marginal properties as well as providing examples of the marginal properties when a log link function is used in the GLMM, that is the function $g() = \log()$ and thus $g^{-1}() = \exp()$. This choice of link function is used when the response variable (Y_{ij}) follows the Poisson distribution.

2.2.2.1 Mean of Y_{ij}

Iterated expectations can be used to derive the marginal mean of Y_{ij} such that

$$\begin{aligned}\mathrm{E}(Y_{ij}) &= \mathrm{E}\{\mathrm{E}(Y_{ij}|\mathbf{b}_i)\} \\ &= \mathrm{E}\{g^{-1}(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{b}_i)\}.\end{aligned}$$

As an example assuming that the log link function is used with a single random effect for each individual, $b_i \sim N(0, \sigma_b^2)$, so that the linear predictor is given by

$$\eta_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + b_i.$$

Then the marginal mean is given by

$$\mathrm{E}(Y_{ij}) = \exp\{\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + (\sigma_b^2/2)\} = \mu_{ij}. \tag{2.7}$$

This derivation involves the use of the moment generating function (m.g.f) of the normal distribution (McCulloch and Searle, 2001).

2.2.2.2 Variance of Y_{ij}

Given that the variance of $Y_{ij}|\mathbf{b}_i$ is equal to $V(\mu_{ij})$, where μ_{ij} is the conditional mean of $Y_{ij}|\mathbf{b}_i$, the marginal variance is given by

$$\begin{aligned}\text{var}(Y_{ij}) &= \text{var}\{\text{E}(Y_{ij}|\mathbf{b}_i)\} + \text{E}\{\text{var}(Y_{ij}|\mathbf{b}_i)\} \\ &= \text{var}\{g^{-1}(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{b}_i)\} + \text{E}\{V[g^{-1}(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{b}_i)]\}\end{aligned}$$

As an example it is assumed that the log link function is used with a single random effect for each individual, $b_i \sim N(0, \sigma_b^2)$, and Y_{ij} follows a Poisson distribution with mean μ_{ij} such that the conditional variance of $Y_{ij}|b_i$ is given by μ_{ij} and the linear predictor is given by

$$\eta_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + b_i.$$

The result is that the marginal variance is given by

$$\begin{aligned}\text{var}(Y_{ij}) &= \text{E}(Y_{ij})\{\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta})[\exp(3\sigma_b^2/2) - \exp(\sigma_b^2/2)] + 1\} \\ &= \mu_{ij} + \mu_{ij}^2[\exp(\sigma_b^2) - 1] \qquad (2.8)\end{aligned}$$

It can be seen from equation (2.8) that the marginal variance of Y_{ij} will always be greater than the mean as the expression $(\exp(\sigma_b^2) - 1)$ will always be greater than 1. This shows that even though $Y_{ij}|b_i$ follows a Poisson distribution, the marginal distribution will not. As a result the marginal distribution will be overdispersed thereby highlighting the way in which random effects can be used to model overdispersion (McCulloch and Searle, 2001).

2.2.3 Estimation

In this section I will discuss 3 commonly used methods of estimation for GLMMs.

2.2.3.1 Maximum likelihood estimation

The estimation of the parameters in a GLMM involves maximization of marginal likelihoods, which are found by integrating out the random effects. The marginal likelihood contribution for the i^{th} subject is found by integration over $\mathbf{b}_i$, for $i = 1, ..., q$ with each subject having n_i observations such

that the total number of observations $n = \sum_{i=1}^{q} n_i$. The marginal likelihood contribution is thus given as

$$L_i(\boldsymbol{\beta}, \mathbf{D}, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i. \tag{2.9}$$

This contribution is used to get the likelihood for $\boldsymbol{\beta}, \mathbf{D}$ and $\boldsymbol{\phi}$, which is

$$\begin{aligned} L(\boldsymbol{\beta}, \mathbf{D}, \phi) &= \prod_{i=1}^{q} f_i(y_i|\boldsymbol{\beta}, \mathbf{D}, \phi) \\ &= \prod_{i=1}^{q} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i. \end{aligned} \tag{2.10}$$

This integration is usually analytically intractable and as a result various algorithms have been devised in order to deal with this problem. A numerical approximation to the integral is the Gauss-Hermite quadrature (GHQ) approach. This approach is feasible if the random effects are independent of each other so that only single integrals are being evaluated. This approach cannot be used if high dimensional integrals are required, for instance when the random effects have a crossed design (Lee et al., 2006).

Simulation methods can be used to overcome the shortcomings of the GHQ. Examples of simulation methods are the Monte Carlo EM method which McCulloch (1994) used to develop a framework for maximum likelihood and restricted maximum likelihood estimation of variance components from binary data. Other examples are the simulated maximum likelihood method (McCulloch, 1997) and the Gibbs sampling method (Karim and Zeger, 1992). It must be stated that the simulation methods have the shortcomings of being computationally intensive and also being inaccurate in parameter estimation on occasions (Hobert and Casella, 1996).

An alternative method of estimation is the use of approximation methods. Schall (1991) outlined an algorithm which yields approximate maximum likelihood or quasi-maximum likelihood estimates for the fixed effects and dispersion components, as well as approximate empirical Bayes estimates of the random effects.

2.2.3.2 Empirical Bayes estimation

This method is used to obtain best linear unbiased predictions (BLUPs) of the random effects. The random effects reflect between-subject variability which is helpful in detecting subjects or groups of subjects which evolve differently over time. The predictions are also required when the focus of the study is on prediction of subject-specific evolutions. The prediction of the random effects are based on the posterior distribution with the probability density function given as

$$f_i(\mathbf{b}_i|y_{ij}, \boldsymbol{\beta}, \mathbf{D}, \phi) = \frac{f_i(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\mathbf{D})}{\int f_i(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i}. \quad (2.11)$$

This posterior density usually does not have a normal distribution and thus the posterior mode is used as a point predictor for $\mathbf{b}_i$, instead of the posterior mean. As a result the predictor $\hat{\mathbf{b}_i}$ is the value of $\mathbf{b}_i$ that maximizes $f_i(\mathbf{b}_i|y_{ij}, \boldsymbol{\beta}, \mathbf{D}, \phi)$ where the unknown parameters are replaced by the estimates obtained from maximum likelihood estimation.

2.2.3.3 Bayesian methods

Bayesian methods of model fitting are widely used among researchers. These methods involve assigning prior distributions for $\boldsymbol{\beta}$, ϕ and $\mathbf{D}$ whilst usually assuming independence between them. The prior density function for $\boldsymbol{\beta}$, namely $f(\boldsymbol{\beta})$, is usually a normal distribution or a flat, non-informative prior (Molenberghs and Verbeke, 2005).

Jeffreys priors can be used for the $\mathbf{D}$ and ϕ prior densities, denoted $f(\mathbf{D})$ and $f(\phi)$ respectively (Gelman et al., 1995). However such a choice of priors can lead to improper posteriors (Fahemeir and Tutz, 2001). An alternative approach was proposed by Besag et al. (1995) who used proper but highly dispersed inverted Wishart (IW) priors for $\mathbf{D}$, such that $\mathbf{D} \sim IW(\xi, \psi)$, where ξ and ψ are hyper-parameters which have to be selected carefully.

When the prior distributions have been specified, the posterior distribution can be expressed as

$$f(\boldsymbol{\beta}, \mathbf{D}, \phi, \mathbf{b}_1, ..., \mathbf{b}_q) \propto \prod_{i=1}^{q} \prod_{j=1}^{n_i} f_i(y_{ij}|\boldsymbol{\beta}, \phi, \mathbf{b}_i) f(\mathbf{b}_i|\mathbf{D}) f(\mathbf{D}) f(\boldsymbol{\beta}) f(\phi)$$

Standard algorithms can then be used to draw samples from the posterior distribution which will be used to get estimates for the fixed and random effects for example, Karim and Zeger (1992) used Gibbs sampling.

2.2.4 Inference

Since the fitting of GLMMs is based on maximum likelihood principles, the inferences for the estimated parameters are obtained from classical maximum likelihood theory. As a result, assuming that the appropriate model has been fitted, the estimators of the parameters are asymptotically normally distributed with the true value of parameter as the mean and the inverse Fisher information matrix as the covariance matrix. Due to the estimators having an asymptotic normal distribution, Wald-type tests can be performed on estimates. Composite hypotheses can also be tested using a standardized quadratic formulation of the Wald statistic which is compared to the chi-squared distribution (Molenberghs and Verbeke, 2005). Other tests which can be used are the likelihood ratio and score tests.

Classical Wald, likelihood ratio and score tests can be used when interest is on inference for the variance components in $\mathbf{D}$, as long as the hypotheses being tested are not on the boundary of the parameter space. An example is when a researcher wants to test whether the variance σ_b^2 of a random effect is equal to zero, that is $H_0 : \sigma_b^2 = 0$ versus $H_1 : \sigma_b^2 > 0$. Under such a situation none of the classical Wald, likelihood ratio and score tests can be used as the regularity conditions are not met (Stram and Lee, 1995). The theory on tests of hypotheses on the boundary of the parameter space can be found in work by Self and Liang (1987).

2.3 Hierarchical Generalized Linear Models

Hierarchical generalized linear models (HGLMs) are a generalization of GLMMs in that the random effects can have any distribution in the exponential family, whereas GLMMs always have normal random effects. The fitting of HGLMs is not as computationally intensive as that of GLMMs. This is because instead of integrating out the random effects, which occurs in fitting

GLMMs, the fitting of HGLMs is based on a modified form of the likelihood function known as the hierarchical or h-likelihood.

According to Lee and Nelder (1996), a hierarchical generalized linear model (HGLM) is defined by the following properties:

1. Let Y_{ij} be the response variable for the j^{th}occurrence of the i^{th} subject and $\mathbf{b}_i$ be the unobserved random effect for the i^{th} subject. Conditional on $\mathbf{b}_i$, Y_{ij} has the following properties:

$$\mathrm{E}(Y_{ij}|\mathbf{b}_i) = \mu_{ij}$$

$$\mathrm{var}(Y_{ij}|\mathbf{b}_i) = \phi V(\mu_{ij}),$$

where $V()$ is a monotonic function of μ_{ij} and ϕ is the dispersion parameter. The linear predictor is of the form

$$g(\mathrm{E}(Y_{ij}|\mathbf{b}_i)) = g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\boldsymbol{\nu}_i, \tag{2.12}$$

where $\boldsymbol{\nu}_i$ is a monotonic function of $\mathbf{b}_i$.

2. The random component $\mathbf{b}_i$ follows a distribution conjugate to a GLM family of distributions with parameter λ.

Thus the probability density function (p.d.f) of $Y_{ij}|\mathbf{b}_i$ has parameters μ_{ij} and ϕ, and the p.d.f of $\boldsymbol{\nu}_i$ has parameter λ_i.

The estimation of parameters in HGLMs involves maximizing the hierarchical log likelihood (Lee and Nelder, 1996), an analogue of Henderson's mixed model likelihood equations (Henderson, 1975). The log h-likelihood has the form

$$h_i \equiv \log f_{\boldsymbol{\beta},\phi}(\mathbf{y}_i|\boldsymbol{\nu}_i) + \log f_{\lambda_i}(\boldsymbol{\nu_i}), \tag{2.13}$$

where $\log f_{\lambda_i}(\boldsymbol{\nu}_i)$ is the logarithm of the density function for $\boldsymbol{\nu}_i$ and $\log f_{\boldsymbol{\beta},\phi}(\mathbf{y}_i|\boldsymbol{\nu}_i)$ is the log likelihood for $\mathbf{y}_i|\boldsymbol{\nu}_i$ with unknown parameters ϕ and λ_i which are dispersion parameters. The h-likelihood is not a joint likelihood in the normal sense because the random effects are not observed.

The estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\nu}_i$ are found by joint maximization of h_i for parameters $\boldsymbol{\beta}$ and $\boldsymbol{\nu}_i$. This is done by solving the equations

$$\frac{\partial h_i}{\partial \boldsymbol{\beta}} = 0, \frac{\partial h_i}{\partial \boldsymbol{\nu_i}} = 0.$$

This estimation must be done before estimating the dispersion parameters. The dispersion parameters are estimated by maximizing the adjusted profile h-likelihood, which is an extension of the adjusted profile likelihood outlined by Lindsey (1996). Alternatively the dispersion parameters can be estimated by using an extended quasi h-likelihood derived from Wedderburn (1974) quasi-likelihood equations.

Lee and Nelder (1996) showed that the use of the h-likelihood gave reliable and useful estimators which shared the same properties as the estimators derived from the marginal likelihood, with the added advantage of not requiring the integrating out of random effects. They also showed that the fitting algorithm for the HGLMs can be reduced to fitting a two-dimensional set of generalized linear models, one dimension being the mean and dispersion parameters while the other is the fixed and random effects.

Once the model has been fitted, validity of its underlying assumptions must be assessed. Lee and Nelder (1996) provide residuals for the mean generalized linear model as well as for the dispersion model, thus allowing the model assumptions for both the mean and dispersion parameters to be tested. Deviance residuals are preferred because they provide a good approximation to normality for all generalized linear model distributions (Pierce and Schafer, 1986) excluding extreme cases like binary data.

The model checking plots which are used are the normal probability plots, a plot of residuals against fitted values and the plot of absolute residuals. A satisfactory model must have the plot of residuals against fitted values and the plot of absolute residuals showing running means that are approximately straight and flat. If the running mean of the plot of residuals against fitted values has a marked curvature the link function is deemed to be unsatisfactory or there are missing terms in the linear predictor, or both shortfalls are occurring. The plot of absolute residuals is used to check the choice of variance function. If for example there is a downward trend in this plot it would imply that the residuals are falling in absolute value as the mean increases, that is the assumed variance is increasing too rapidly with

the mean (Lee and Nelder, 2001).

Table 2.1: **Examples of conjugate HGLMs**

$y_{ij}\|\mathbf{b}_i$ distribution	$g(\mu_{ij})$	$\mathbf{b}_i$ distribution	$\boldsymbol{\nu}_i(\mathbf{b}_i)$	Model
Normal	identity	Normal	identity	linear mixed model
Binomial	logit	Beta	logit	beta-binomial model
Gamma	reciprocal	Inverse-gamma	recip	gamma-inverse-gamma model
Poisson	log	Gamma	log	Poisson-gamma model

$\boldsymbol{\nu}_i(\mathbf{b}_i)$ is the link function between $\boldsymbol{\nu_i}$ and $\mathbf{b}_i$

$g(\mu_{ij})$ is the link function

Different assumptions about the distribution of $y_{ij}|\mathbf{b}_i$ and that of $\boldsymbol{\nu}_i$, the random component, lead to different forms of HGLMs. Some of the possible conjugate HGLMs used are shown in Table 2.1 (Lee et al., 2006). In addition to these conjugate HGLMs there are other combinations of $y_{ij}|\mathbf{b}_i$ and $\boldsymbol{\nu}_i$ which lead to non-conjugate HGLMs for example the binomial-normal HGLM (binomial GLMM), binomial-gamma HGLM, gamma-gamma HGLM and Poisson-normal HGLM. For a conjugate HGLM joint maximization of the h-likelihood gives the same parameter estimates as the use of a marginal likelihood, this is not the case for the non-conjugate HGLMs as they are estimated using Laplace approximation.

Chapter 3

Review of model diagnostics in non-normal data

In this chapter I will review the existing model diagnostic techniques used for non-normal data. This chapter will begin with a brief outline of the aims of using model diagnostics. A review of the model diagnostic techniques used in independent and longitudinal non-normal data will then be provided. Finally a review of the variance shift outlier model will be given.

3.1 Aims of model diagnostics

Once a GLM/GLMM has been fit, it is important to validate the adequacy of the model. This validation process is called model diagnostics. There are three components which make up model diagnostics, the first being checking the validity of the underlying assumptions of the model. An example of an underlying assumption of a model is the choice of the distribution of the observed responses. The other two components are identifying isolated and systematic discrepancies in the model. A systematic discrepancy occurs when the model does not adequately fit either all or a large subset of data. There are many reasons why such a discrepancy might arise. For example, it might be due to an inappropriate choice of link function or the need for one of the covariates in the model to be transformed. An isolated discrepancy arises when a few individual observations do not fall in line with the pattern

of the rest of the observations (Davison and Tsai, 1992).

Picking up isolated discrepancies involves identifying outlying (outliers) and influential observations. The difference between outliers and influential observations is that outliers are observations which are not adequately fit by the model, while influential observations are observations which have a large effect on the parameter estimates of a model. It is important to note that influential observations can also be outliers. The parameter estimation methods used for GLMs and GLMMs are very sensitive to influential observations. These observations can lead to distorted results which lead to incorrect conclusions. As a result it is very important that these influential observations are identified before any conclusions are made. In this thesis I will be primarily looking at isolated discrepancies.

3.2 Model diagnostics for independent non-normal data

Model diagnostic tests are mainly based on model residuals and transformations of these residuals. Some examples of frequently used residuals are raw, Pearson's, standardized Pearson's, deviance and standardized deviance residuals. The i^{th} raw residual (r_i), for $i = 1, ..., n$, is simply the difference between the observed response (y_i) and the predicted response ($\hat{y}_i$). This residual does not take into account the precision involved in fitting y_i. Pearson's residuals are used to account for this by incorporating the standard error of the response, that is se(y_i). The i^{th} Pearson residual is given as

$$r_{Pi} = \frac{y_i - \hat{y}_i}{\text{se}(\hat{y}_i)}.$$

The sum of these residuals give the Pearson's χ^2 statistic which is a measure of the relative deviations between the observed and predicted responses, this is used as a measure of the goodness of fit of a particular model. A large value of this statistic indicates a poor model fit.

The Pearson's residuals do not take into account the inherent variation in the predicted responses thus the residuals do not have even approximate unit variance, standardized Pearson residuals are used to account for this.

Standardized Pearson residuals divide raw residuals by their standard error, $\text{se}(y_i - \hat{y}_i)$. The standard error is given as

$$\text{se}(y_i - \hat{y}_i) = \sqrt{[\hat{v}_i(1 - h_i)]},$$

where $\hat{v}_i = \text{se}(y_i)$ and h_i is the leverage statistic defined as the i^{th} diagonal element of the $n \times n$ matrix $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}}$, $\mathbf{W}$ is a $n \times n$ diagonal matrix of weights which were used in the model fitting process, $\mathbf{X}$ is a $n \times p$ design matrix of covariates where p is the number of unknown parameters used in the model (Collett, 1996). The i^{th} standardized Pearson residual is given as

$$r_{Pi} = \frac{y_i - \hat{y}_i}{\sqrt{[\hat{v}_i(1 - h_i)]}}.$$

Pregibon (1981) introduced deviance residuals, d_i, which are residuals which measure the deviance contributed from each response. Given a simple linear logistic regression the deviance residual for the i^{th} response is given as

$$d_i = \text{sgn}(y_i - \hat{y}_i)\left[2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right]^{\frac{1}{2}},$$

where $\text{sgn}(y_i - \hat{y}_i)$ is a function which is used to ensure that d_i is positive when $y_i \geq \hat{y}_i$ and negative otherwise. The overall deviance of the model, D, is the sum of the squared deviance residuals such that $D = \sum_{i=1}^{n} d_i^2$. These residuals can be standardized, resulting in the i^{th} standardized residual being given as

$$r_{Di} = \frac{d_i}{\sqrt{(1 - hi)}}.$$

Another residual commonly used is the likelihood residual, r_{Li}, which was defined by Williams (1987). This residual is a combination of the standardized Pearson's and deviance residuals such that

$$r_{Li} = \text{sgn}(y_i - \hat{y}_i)\sqrt{[h_i r_{Pi}^2 + (1 - h_i)r_{Di}^2]},$$

where h_i is the leverage statistic as defined previously. Pierce and Schafer (1986) used numerical studies to show that standardized deviance residuals and likelihood residuals are reasonably well approximated by the standard normal distribution, the result of this finding is that for fairly large samples the residuals which are not in the range -2 to 2 are potentially outlying;

for smaller samples the t-distribution should be considered when outlining a range for the residuals. The same study also revealed that standardized Pearson's residuals are not closely approximated by the normal distribution, thus it is preferable to use standardized deviance residuals for model checking.

Once the respective residuals have been calculated plots or tables of the residuals can be used to identify outlying observations. A plot of residuals against observation number or against values of the linear predictor can be used to identify outliers by considering relatively large residuals as outliers. The same plot can be used to test model adequacy as a systematic pattern in the plot indicates model inadequacy. Normal and half-normal plots of deviance residuals can also be used to test for model adequacy. Normal plots of deviance residuals plot the residuals in increasing order against their expected values assuming that the residuals follow a normal distribution, if the plot is a straight line then the model is adequate. A half normal plot considers absolute values of the deviance residuals, this plot must also be a straight line if the model is adequate.

I will now proceed to describe how to detect influential observations in GLMs. Influential observations are observations which have a large effect on the model parameter estimates, as a result it is very important that such observations are identified. It may not be possible to identify influential observations in a similar way as outliers, that is using residual plots, as these observations may have a large influence on the model thus their associated residuals are small.

Collett (1996) showed that an observation which is very different in value, in terms of explanatory variables, may be a possible influential observation. A statistic used to measure this difference is the leverage statistic, h_i. An observation is said to be influential if the value of h_i is greater than $2p/n$ where p is the number of unknown parameters and n is the number of observations. A useful plot which can be used is a plot of Pearson's residuals against the leverage values, as this plot allows a researcher to identify both outliers and influential observations in the same plot.

Certain observations may have an influence on the overall goodness of fit of a model. The influence of an observation on the goodness of fit can

be measured by deleting it from the dataset and assessing the change in the value of the χ^2 statistic and the deviance statistic. The change in the χ^2 statistic can be approximated by the square of the standardized Pearson's residual (r_{Pi}^2) while the change in the deviance statistic is approximated by the square of the likelihood residual (r_{Li}^2).

It is possible to measure the influence of the i^{th} observation on a set of parameter estimates by assessing the change in the parameter estimates when the i^{th} observation is deleted from the dataset. Collett (1996) showed that the change can be approximated by using the statistic

$$D_i = \frac{h_i r_{Pi}^2}{p(1-h_i)},$$

where h_i is the leverage statistic for the i^{th} observation, r_{Pi} is the i^{th} Pearson's residual and p is the number of unknown parameters. An index plot of this statistic can be used to identify influential observations as relatively large values of D_i are indicative of influential observations.

If a researcher wants to measure the influence of the i^{th} observation on the j^{th} parameter, $\hat{\beta}_j$ for $j = 1, ..., p$, then the researcher can delete the observation and assess the change in parameter estimate. This change can be approximated by the delta-beta statistic given as

$$\triangle_i \hat{\beta}_j = \frac{(\mathbf{X}'\mathbf{W}\mathbf{X})_{j+1}^{-1}\mathbf{x}_i(y_i - \hat{y}_i)}{(1-h_i)se(\hat{\beta}_j)},$$

where $(\mathbf{X}'\mathbf{W}\mathbf{X})_{j+1}^{-1}$ is the $(j+1)^{th}$ row of the variance covariance matrix of parameter estimates and $\mathbf{x}_i$ the p-dimensional vector of covariates for the i^{th} observation. Relatively large values of delta-beta are indicative of influential observations on the j^{th} parameter estimate and this can be shown by use of index plots.

3.3 Model diagnostics for longitudinal non-normal data

The estimation methods used when fitting GLMMs are quite complicated which makes the calculation of diagnostic measures difficult. As a result

very little work has been done on model diagnostics for GLMMs. There has been some work done on deletion measures by Xu et al. (2006). Xu et al. (2006) extended the work of Cook (1977) and developed deletion measures for GLMMs when the stochastic approximation (SA) and the Markov chain Monte Carlo (MCMC) methods (Gu and Kong, 1998) are used to calculate the maximum likelihood parameter estimates. This method of estimation considers random effects as hypothetical missing values. The by-products of the maximum likelihood estimation are used to compute the diagnostic measures outlined in the paper. The diagnostic measures assess the influence of an observation or cluster of observations, given as M_i, on the parameter estimates when they are excluded from the dataset. Given that the observed vector of observations is $\mathbf{Y}$ and the estimated parameters are given as $\hat{\boldsymbol{\beta}}$, with the set of observations excluded from the analysis given as M_i; the measure of influence is given as the distance between the parameter estimates from the full dataset ($\hat{\boldsymbol{\beta}}$) and the parameter estimates when the M_i set of data is excluded ($\hat{\boldsymbol{\beta}}_{[M_i]}$). The diagnostic measures developed were

$$GD_{[M_i]} = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[M_i]})'\{-Q(\hat{\boldsymbol{\beta}})\}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[M_i]}),$$

where $Q(\hat{\boldsymbol{\beta}}) = \delta^2 Q(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}})/\delta\boldsymbol{\beta}\delta\boldsymbol{\beta}'|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ and $Q(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}) = \mathrm{E}\{\mathrm{L}(\boldsymbol{\beta}|\mathbf{Y})|\mathbf{Y}, \hat{\boldsymbol{\beta}}\}$. Another measure of influence which was developed was the $QD_{[M_i]}$ statistic which is similar to the likelihood distance measure developed by Cook and Weisberg (1982). This was given by

$$QD_{[M_i]} = 2\{Q(\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\beta}}) - Q(\hat{\boldsymbol{\beta}}_{[M_i]}|\hat{\boldsymbol{\beta}})\}.$$

Given that the observed data likelihood is given by $p(\mathbf{Y}|\hat{\boldsymbol{\beta}})$, another measure of influence can be developed based on determining the distance between the likelihoods of the full dataset and the dataset with the set M_i excluded. This measure is given by

$$LD_{M_i} = -2\log\frac{p(\mathbf{Y}|\hat{\boldsymbol{\beta}}_{[M_i]})}{p(\mathbf{Y}|\hat{\boldsymbol{\beta}})}.$$

Since the by-products of estimation procedure were used in developing these diagnostic measures, there is very little additional computation that is required thus making this method easy to apply. Xu et al. (2006) highlighted

the fact that deletion type measures can suffer from the masking or swamping effects that are caused by an influential group, rather that a single observation, in the data. This can be solved by deleting a set of observations at a time instead of a single observation.

Xiang et al. (2002) investigated the Cook's distance for identifying influential observations for GLMMs with applications to clustered data. They showed that their method was efficient in identifying influential clusters. The Cook's distance statistic developed for the k^{th} cluster is given by

$$CD_k(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_k)'(-\ddot{l}(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_k)/(p.a^{-1}(\hat{\phi})),$$

where $\ddot{l}(\hat{\boldsymbol{\beta}})$ is the second order derivative of the log-likelihood function with respect to the unknown parameters $\boldsymbol{\beta}$, p is the number of unknown parameters and a is a function acting on the dispersion parameter (ϕ) which is defined in the linear predictor of the GLMM which is being investigated. This paper did not attempt to perform individual observation deletions. This was because individual observations are either correlated within a cluster or present as a series of repeated measurements, thus making case-deletion impractical (Banerjee and Frees, 1997).

The generalization of the local influence measures from normally distributed responses to that for generalized linear mixed models was discussed by Ouwens et al. (2001). It was shown in this paper that the subject-oriented influence measure is a special case of the proposed observation-oriented influence measure. The statistic, for the set of observations M_i, used in the paper was initially developed by Cook (1986) and it is given as

$$C_d = 2\left|\mathbf{d}' \triangle' \left(\frac{\delta^2 L(\mathbf{y}|\boldsymbol{\varsigma})}{\delta \boldsymbol{\varsigma}^2}\right)^{-1} \triangle \mathbf{d}\right|,$$

where ς is the maximum likelihood estimate of the full dataset, $\mathbf{d}$ is a vector with values of one in the positions corresponding to observations in M_i and zeros everywhere else, $\mathbf{y}$ is a vector of observed responses, $L()$ is the likelihood function and $\triangle = \delta^2 L(\mathbf{y}|\boldsymbol{\varsigma}, \boldsymbol{\omega})/\delta\boldsymbol{\varsigma}\delta\boldsymbol{\omega}$ with $\boldsymbol{\omega}$ being a vector of weights. The computation of $\triangle$ is described is in full by Ouwens et al. (2001). Ouwens et al. (2001) went on to propose a two-step diagnostic procedure which involved firstly searching for influential subjects, and then finally searching

for influential observations. The article then goes on to illustrate through an example, of a two-treatment multiple-period crossover trial, that there is practical importance in the detection of influential observations in addition to the detection of influential subjects.

3.4 The variance shift outlier model (VSOM)

For the remainder of this chapter I will aim to briefly describe the development of another form of model diagnostics which is called the "Variance Shift Outlier Model". The actual model will be discussed in greater depth in Chapter 4.

This model was initially developed for use in linear regression by Cook and Weisberg (1982) and it was called the "Variance inflation single outlier model". The model was used for identifying a single outlier as being an observation with an inflated variance. In this model maximum likelihood estimates were used and they were characterized in terms of standard least square statistics. It is noted in Cook and Weisberg (1982) that the position of the outlier differed from when normal case-deletion methods were used. However it was shown that if the largest absolute studentized residual corresponded to the largest absolute residual, then the position of the outlier would be the same. An example of where this occurs is when the variance of all residuals is the same, like in balanced design experiments. The main difference between the case-deletion methods and the variance inflation single outlier model is that the case deletion methods are based on deleting outlying observations whilst the variance inflation single outlier model down-weights the outlier in the analysis, thus preserving the data in the analysis. When Thompson (1985) used residual maximum likelihood (REML) estimation, he noted that the outliers were in the same position under both the variance inflation single outlier model and case deletion methods.

Zewotir (2007) considered the model underlying the VSOM in the context of local influence. He observed the effect of using known weights in the error variance of a single observation on changes in the estimates of the fixed and random effects. This was done using a Cook's distance measure (Cook, 1977). Gumedze et al. (2010) extended the work on variance shift models,

by Cook et al. (1982) and Thompson (1985). They did this by formulating it as a linear mixed model and also proposing several tests for the detection and down-weighting of outliers. They called the model the variance shift outlier model (VSOM).

This thesis aims to extend the variance shift outlier model (VSOM) by Gumedze et al. (2010) to count and binomial data. These forms of data are very common in biological areas. Since the collection of data is very expensive it is very appealing for researchers to be able to preserve observations, which are not clearly outlying, whilst down-weighting their effect on the estimation of parameters. The VSOM will be applied to several types datasets in this thesis.

Chapter 4

VSOM for normal data

In this chapter I will introduce the VSOM methodology that will be implemented in this study. In order to get a better understanding of the methodology, I will outline how it was used in Gumedze et al. (2010). Its use in the paper was restricted to normal data. After outlining the VSOM methodology for normal data in a simple linear model and linear mixed model setting, I will proceed to outline the methodology for count and binomial data in the subsequent chapters. The VSOM methodology in this section is derived from Gumedze et al. (2010).

4.1 A VSOM for independent normally distributed data

In this section I will review the VSOM in linear regression in order to get a better understanding of the VSOM methodology in the simple linear case. Consider a simple linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{4.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{X}$ is an $n \times p$ design matrix of full column rank, $\boldsymbol{\beta}$ is a vector of p unknown coefficients and $\mathbf{e} = (e_1, \ldots, e_n)'$ is a vector of residual errors which follow a normal distribution such that $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. The variance-covariance matrix of the data under this model (model (4.1)) is given as $\text{var}(\mathbf{y}) = \sigma^2\mathbf{I}_n$. The REML log-likelihood function (RL) under

this model takes the form

$$-2RL(\sigma^2;\mathbf{y}) = c(\mathbf{X}) + (n-p)\log(\sigma^2) + (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/\sigma^2, \quad (4.2)$$

where $c(\mathbf{X})$ is a constant term, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the best linear unbiased estimate (BLUE) of $\boldsymbol{\beta}$ under the null model (model (4.1)).

According to Gumedze et al. (2010) the variance shift outlier model (VSOM) for the i^{th} observation takes the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \delta_\mathbf{i}\mathbf{d_i} + \mathbf{e}. \quad (4.3)$$

It can be seen that model (4.3) differs from model (4.1) by the addition of the term $\delta_i\mathbf{d}_i$, where $\mathbf{d}_i$ is an i^{th} unit vector of length n, which has a value of 1 in its i^{th} position and zeros in all other positions. δ_i is an unknown random effect with $\delta_i \sim N(0,\alpha_i)$ for $\alpha_i \geq 0$. It becomes apparent that model (4.3) is just a simple linear mixed model with random effect δ_i which has a variance of α_i. This model can be easily fitted using any statistical software which is capable of fitting linear mixed models. The variance-covariance matrix of the data under model (4.3) is given by

$$\text{var}(\mathbf{y}) = \alpha_i\mathbf{d}_i\mathbf{d}_i' + \sigma^2\mathbf{I}_n.$$

This can be parameterized in another form which uses the ratio of the variance components to the residual variance, σ^2

$$\text{var}(\mathbf{y}) = \sigma^2(\omega_i\mathbf{d}_i\mathbf{d}_i' + \mathbf{I}_n) = \sigma^2\mathbf{H}_{(i)},$$

where $\omega_i = \alpha_i/\sigma^2$. The advantage of this parameterization is that the size of the variance shift for the i^{th} unit relative to the residual variance is given by ω_i (Gumedze et al., 2010). The REML log-likelihood function for this model is

$$\begin{aligned}-2RL_{(i)}(\omega_i,\sigma^2;\mathbf{y}) =& c(\mathbf{X}) + (n-p)\log(\sigma^2) + \log|\mathbf{H}_{(i)}| + \log|\mathbf{X}'\mathbf{H}_{(i)}^{-1}\mathbf{X}| \\ &+ (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)})\mathbf{H}_{(i)}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)})/\sigma^2, \quad (4.4)\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'\mathbf{H}_{(i)}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}_{(i)}^{-1}\mathbf{y}$.

In order to identify outliers Thompson (1985) suggested the calculation of $-2RL_{(i)}(\hat{\omega}_i,\hat{\sigma}^2;\mathbf{y})$, where the estimates $\hat{\omega}_i$ and $\hat{\sigma}^2$ are calculated during

the model fitting process. This statistic is to be calculated for all observations such that the observations with large log-likelihood values are deemed to be outliers. The change in the log-likelihood value from the null model is given as

$$LRT_i = -2\{RL(\hat{\sigma}^2; \mathbf{y}) - RL_{(i)}(\hat{\omega}_i, \hat{\sigma}^2; \mathbf{y})\}.$$

Since the LRT_i statistic is calculated for all observations in turn, the issue of multiple testing becomes a problem. In order to address this problem Gumedze et al. (2010) used a parametric bootstrap procedure to identify observations which had significantly large LRT_i values. Observations with significantly large LRT_i values were considered to be potential outliers.

Assuming that the overall type I error rate to be used in the bootstrapping procedure is given as ξ. The observation with the r^{th} largest LRT_i statistic is compared to the $(100 - \xi)^{th}$ percentile of the simulated distribution of the r^{th} largest LRT_i statistic. The parametric bootstrap procedure used by Gumedze et al. (2010) is as follows

1. Fit the null model (model (4.1)) to the data to obtain estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$.

2. Create a new response vector

$$\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}^*$$

   where $\mathbf{e}^*$ is simulated following $N(\mathbf{0}, \hat{\sigma}^2\mathbf{I}_n)$. Then fit the null model (model (4.1)) to $\mathbf{y}^*$ and obtain the set of change in log-likelihood statistics $\{LRT_i; \forall i = 1, ..., n\}$. These statistics are then ordered from largest to smallest and saved.

3. Repeat Step 2 for a reasonably large number of times, R, for example $R = 5000$. This generates an empirical distribution of size R for the change in log-likelihood statistic.

4. Calculate the $(100 - \xi)^{th}$ percentile for the change in log-likelihood statistic.

If the r largest change in log-likelihood statistics all exceed their respective thresholds, then it is concluded that all these corresponding observations

are outliers and a revised model is fitted including a separate variance shift for each of the corresponding observations.

Gumedze et al. (2010) went on further to develop additional test statistics which included squared standardized residuals and score statistics. I will not be considering these statistics in this study.

4.2 A VSOM for longitudinal normally distributed data

Consider the linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}, \tag{4.5}$$

where $\mathbf{Z}$ is a $n \times q$ design matrix for a vector of unknown random effects $\mathbf{b}$, and $\mathbf{e}$ represents the vector of residual errors. The residual errors and random effects are assumed to be independent such that $\text{cov}(\mathbf{b}, \mathbf{e}) = 0$. It is possible to partition the random component of the model, $\mathbf{Zb}$, into c model terms with $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_c]$ and $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_c')'$ with $\text{cov}(\mathbf{b}_h, \mathbf{b}_l) = 0$ for $h \neq l$. Another assumption is that $\mathbf{e} \sim N(0, \sigma^2\mathbf{I}_n)$ and $\mathbf{b_h} \sim N(0, \sigma^2\gamma_h\mathbf{I})$, $\forall h = 1, \ldots, c$; the variance of $\mathbf{b}$ can also be written as $\text{var}(\mathbf{b}) = \sigma^2\mathbf{G}$.

The covariance-variance matrix of model (4.5) can be written in terms of the overall scale parameter, σ^2, and the ratios derived from it. As a result the variance-covariance matrix can be written as

$$\text{var}(\mathbf{y}) = \sigma^2[\Sigma_{h=1}^{c}\gamma_h\mathbf{Z}_h\mathbf{Z}_h' + \mathbf{I}_n] = \sigma^2(\mathbf{G}\mathbf{Z}\mathbf{Z}^{'} + \mathbf{I}_n) = \sigma^2\mathbf{H}. \tag{4.6}$$

The REML log-likelihood function for this model is given as

$$\begin{aligned} -2RL(\sigma^2, \boldsymbol{\gamma}; \mathbf{y}) =& c(\mathbf{X}) + (n-p)\log(\sigma^2) + \log|\mathbf{H}| + \log|\mathbf{X}^{'}\mathbf{H}^{-1}\mathbf{X}| \\ &+ \mathbf{y}^{'}\mathbf{P}\mathbf{y}/\sigma^2, \end{aligned} \tag{4.7}$$

where $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^{'}\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{H}^{-1}$. The REML estimates of the variance parameters are given as $\hat{\sigma}^2$ and $\hat{\boldsymbol{\gamma}}$.

A VSOM for the k^{th} observation in the linear mixed model takes the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \delta_k\mathbf{d}_k + \mathbf{e}, \tag{4.8}$$

where $\delta_k \sim N(0, \sigma^2\omega_k)$ for $\omega_k \geq 0$ and $\mathbf{d}_k$ is a k^{th} unit vector of length n, which has a value of 1 in its k^{th} position and zeros in all other positions. It can be seen that this is just model (4.5) with the addition of the term $\delta_k\mathbf{d}_k$. The variance-covariance matrix of this model takes the form

$$\text{var}(\mathbf{y}) = \sigma^2(\mathbf{ZGZ}^{'} + \omega_k\mathbf{d}_k\mathbf{d}_k^{'} + \mathbf{I}_n). \tag{4.9}$$

It is thus evident that the variance of the k^{th} observation will be inflated by a quantity of $\sigma^2\omega_k$ with the covariances and variances of the other observations remaining the same.

The REML log-likelihood function for model (4.8) is given as

$$\begin{aligned} -2RL_{(k)}(\omega_k, \sigma^2, \boldsymbol{\gamma}; \mathbf{y}) =& c(\mathbf{X}) + (n-p)\log(\sigma^2) + \log|\mathbf{H}_{(k)}| + \log|\mathbf{X}^{'}\mathbf{H}_{(k)}^{-1}\mathbf{X}| \\ &+ \mathbf{y}^{'}\mathbf{P}_{(k)}\mathbf{y}/\sigma^2, \end{aligned} \tag{4.10}$$

where $\mathbf{P}_{(k)} = \mathbf{H}_{(k)}^{-1} - \mathbf{H}_{(k)}^{-1}\mathbf{X}(\mathbf{X}^{'}\mathbf{H}_{(k)}^{-1}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{H}_{(k)}^{-1}$ and $\mathbf{H}_{(k)} = \mathbf{H} + \omega_k\mathbf{d}_k\mathbf{d}_k^{'}$.

Gumedze et al. (2010) used the change in the log-likelihood value from the null model as a test statistic for identifying outlying observations. This statistic is given by

$$LRT_k = -2\{RL(\hat{\sigma}^2, \hat{\boldsymbol{\gamma}}; \mathbf{y}) - RL_{(k)}(\hat{\omega}_k, \hat{\sigma}^2, \hat{\boldsymbol{\gamma}}; \mathbf{y})\}.$$

Once again the issue of multiple testing becomes a problem as the LRT_k statistic is calculated for all observations in turn. As a result the parametric bootstrap procedure was used to identify observations which had significantly large LRT_k values. Observations with significantly large LRT_k values were considered to be potential outliers.

Assuming that the overall type I error rate to be used in the bootstrapping procedure is given as ξ. The observation with the r^{th} largest LRT_k statistic is compared to the $(100-\xi)^{th}$ percentile of the simulated distribution of the r^{th} largest LRT_k statistic. The parametric bootstrap procedure used by Gumedze et al. (2010), under the null hypothesis that no outliers are present in the data, is as follows

1. Fit the null model (model (4.5)) to the data to obtain estimates $\hat{\boldsymbol{\beta}}$, $\hat{\gamma_h}, \forall h = 1, ..., c$, and $\hat{\sigma}^2$.

2. Create a new response vector

$$\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{b}^* + \mathbf{e}^*$$

where $\mathbf{b}_h^*$ is randomly generated following $N(\mathbf{0}, \hat{\sigma}^2\gamma_h\mathbf{I})$ and $\mathbf{e}^*$ is randomly simulated following $N(\mathbf{0}, \hat{\sigma}^2\mathbf{I}_n)$. Then fit the null model (model (4.5)) to $\mathbf{y}^*$ and obtain the set of change in log-likelihood statistics $\{LRT_k; \forall k = 1, ..., n\}$. These statistics are then ordered from largest to smallest and saved.

3. Repeat Step 2 for a reasonably large number of times, R, for example $R = 5000$. This generates an empirical distribution of size R for the change in log-likelihood statistic.

4. Calculate the $(100 - \xi)^{th}$ percentile for the change in log-likelihood statistic.

If the r largest change in log-likelihood statistics all exceed their respective thresholds, then it is concluded that all these corresponding observations are outliers and a revised model is fitted including a separate variance shift for each of the corresponding observations.

This methodology can be extended to the identification of outlying subjects. The VSOM for the i^{th} subject effect $\forall i = 1, ..., q$ is given as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \varsigma_i\mathbf{d}_i + \mathbf{e}, \tag{4.11}$$

where $\mathbf{d}_i$ is a vector of length n with values of 1 for the units which correspond to the observations due to the i^{th} subject, and ς_i is an unknown random coefficient which follows a normal distribution such that $\varsigma_i \sim \mathrm{N}(0, \sigma^2\psi_i)$ for $\psi_i \geq 0$. The variance of model (4.11) is given as

$$\mathrm{var}(\mathbf{y}) = \sigma^2\{\gamma\mathbf{Z}\mathbf{Z}^{'} + \psi_i\mathbf{d}_i\mathbf{d}_i^{'} + \mathbf{I}\} = \sigma^2\mathbf{H}_{(i)}. \tag{4.12}$$

The REML log-likelihood function for model (4.11) is given as

$$-2RL_{(i)}(\psi_i, \sigma^2, \boldsymbol{\gamma}; \mathbf{y}) = -2RL(\sigma^2, \boldsymbol{\gamma}; \mathbf{y}) + \log(c_{ii}\psi_i + 1) - \psi_i c_i^2/\{\sigma^2(c_{ii}\psi_i + 1)\}, \tag{4.13}$$

where $c_i = \mathbf{d}_i^{'}\mathbf{P}\mathbf{y}$ and $c_{ii} = \mathbf{d}_i^{'}\mathbf{P}\mathbf{d}_i$. The change in log-likelihood statistic was used to identify outlying subjects and this was given as

$$LRT_i = -2\{RL(\hat{\sigma}^2, \hat{\boldsymbol{\gamma}}; \mathbf{y}) - RL_{(i)}(\hat{\psi}_i, \hat{\sigma}^2, \hat{\boldsymbol{\gamma}}; \mathbf{y})\}.$$

The parametric bootstrapping procedure similar to the one used for identifying outlying observations was then used to identify outlying subjects.

The variance shift outlier models for observations and subjects can be combined by using models (4.11) and (4.8). The resulting model would be of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \delta_k\mathbf{d}_k + \varsigma_i\mathbf{d}_{sub(i)} + \mathbf{e}, \tag{4.14}$$

where $\mathbf{d}_k$ is an k^{th} unit vector of length n, which has a value of 1 in its k^{th} position and zeros in all other positions, and $\mathbf{d}_{sub(i)}$ is a vector of length n with values of 1 for the units which correspond to the observations due to the i^{th} subject. The variance of model (4.14) is given by

$$\text{var}(\mathbf{y}) = \sigma^2\{\gamma\mathbf{Z}\mathbf{Z}^{'} + \omega_k\mathbf{d}_k\mathbf{d}_k^{'} + \psi_i\mathbf{d}_{sub(i)}\mathbf{d}_{sub(i)}^{'} + \mathbf{I}\}. \tag{4.15}$$

4.3 Applying the VSOM using R

The VSOM introduced by Gumedze et al. (2010) for simple linear and linear mixed models was implemented using the GENSTAT statistical system (Payne et al., 2011). In order to fully understand the work they did, I went about translating the GENSTAT code they used into R (R Core Team, 2012) code. The full code used is given in the appendix, chapter 10.1.1. In order to analyze linear mixed models in R, the **lme4** package (Bates et al., 2012) can be used. This package is an extension of the previously used **nlme** package (Pinheiro et al., 2012). The advantages of using the **lme4** package over the **nlme** package are that firstly, the **lme4** package can handle models with crossed and nested random effects unlike the **nlme** package which can only handle nested random effects. Secondly the **lme4** package has improved computational algorithms, thus it can process large datasets at a faster rate than the **nlme** package.

The general structure of a model fit using the **lme4** package is given by:

$$lmer(response \sim covariate(s) + (covariates|group), data, family),$$

where *group* is the grouping factor and $family$ is the distributional family of the response. The **lme4** package can also be used to analyze GLMMs upon specification of the distributional family of the responses.

Gumedze et al. (2010) stated that residual maximum likelihood (REML) estimation was the preferred method to use when implementing a VSOM because it does not give biased estimates of the variance parameters unlike maximum likelihood estimation (MLE). Maximum likelihood estimation produces biased estimates because it does not take into account the degrees of freedom involved in estimating the fixed effects, when it is used to estimate the variance parameters. The REML method was first used in random effect models by Patterson and Thompson (1971).

In this study it was found that for normally distributed responses, the **lme4** package uses REML estimation by default. The results given by using the R code in the appendix gave similar results to those when using the GENSTAT statistical system. There was, however, a computational difference when it came to calculating likelihood statistics as the **lme4** package calculates the full likelihood statistic, unlike GENSTAT which does not calculate the constant part of the statistic. This difference becomes irrelevant in practice as we are only concerned with the deviance statistic. This statistic is approximately the difference between the likelihood statistics of two models, thus the constant values would cancel out resulting in similar deviance results when using both statistical systems.

The aim of this study is to extend the work of Gumedze et al. (2010) to incorporate non-normal responses. GENSTAT is able to give REML estimates regardless of the distributional family of the response. However **lme4** package in R uses Laplace approximation in its estimation procedure for non-normal responses. This estimation procedure only gives maximum likelihood estimates which would give biased variance estimates for GLMMs, including the VSOM. As a result R will not be used for fitting the VSOM for non-normal responses in this thesis.

Chapter 5

VSOM for count data

The Poisson model is generally used to model count data with covariates, and it is applied within the generalized linear modeling (GLM) framework. There is a key highly restrictive feature for Poisson data, that is the relationship between the mean and the variance of the data. This relationship implies that the variance must be a deterministic function of the mean with $\mathrm{V}(\mu) = \mu$, where μ is the mean and $\mathrm{V}(\mu)$ is the variance of the counts. Thus it is clear that there will be problems if the data are overdispersed. Overdispersion is generally due to missing covariates, inadequate link functions or outlying observations. This thesis will be restricted to overdispersion due to outliers.

I will proceed in this section by outlining how the VSOM framework, as a model for outliers, can be applied to Poisson data using the Poisson-normal, Poisson-gamma and negative-binomial models. It will thus be shown that it is possible to link the VSOM for normally distributed data to that for Poisson data by considering the overdispersion associated with the i^{th} observation or subject (for longitudinal data) as a proxy for the variance inflation for that observation or subject.

5.1 Poisson regression models

Poisson regression models are used to fit a model to independent Poisson count data with covariates. In this section the Poisson GLM and the negative binomial distribution, which is used to model overdispersed count data, will

be described.

5.1.1 Poisson GLM

If $Y_i \sim$ Poisson (λ_i) then the probability mass function is given by

$$\Pr(Y_i = y_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!},$$

which can be written in exponential form as

$$\Pr(Y_i = y_i) = \exp\{y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\}.$$

The linear predictor of the Poisson distribution is of the form

$$\log(\mu_i) = \mathbf{X}_i \boldsymbol{\beta}, \tag{5.1}$$

where $\mathbf{X}_i$ is the i^{th} row of the $n \times p$ design matrix of covariates with $\boldsymbol{\beta}$ being a p-dimensional vector of fixed coefficients. The Poisson distribution is a member of the exponential family of distributions with the natural parameter $\theta = \log(\lambda)$, dispersion parameter $\phi = 1$ and variance function

$$\text{var}(Y_i) = \mu_i = \lambda_i. \tag{5.2}$$

The property of equality between the variance and the mean of the Poisson distribution, may not be met for a particular dataset involving independent count data. That is the variance may be smaller (underdispersed count data) or larger that the mean (overdispersed count data). Underdispersion rarely occurs in practice (Collett, 1996), thus only overdispersion will be considered in this study. Overdispersion is generally due to missing covariates (either fixed or random terms), inadequate link functions or outlying observations. In the next section the negative binomial distribution which is usually used to model overdispersed count data will be discussed. The negative binomial distribution is introduced because it will be used to formulate a VSOM, for count data, in the subsequent section.

5.1.2 Overdispersed count data

In situations where count data are overdispersed, that is $\text{var}(Y_i) = \phi\mu_i$ with $\phi > 1$, either a quasi-likelihood approach (McCullagh and Nelder, 1989)

or a negative binomial distribution can be used to model the data. Lee and Nelder (2000) showed that the negative binomial distribution can be formulated as a Poisson-gamma HGLM. I will discuss these different ways of modeling overdispersion, in count data, below.

The quasi-likelihood approach (McCullagh and Nelder, 1989) allows the variance of the response variable to have an adjustable dispersion parameter $\phi > 1$ such that

$$\text{var}(Y_i) = \phi\mu_i. \tag{5.3}$$

A traditional approach for deriving the negative binomial distribution is to assume that $Y_i|\lambda_i$ follows a Poisson distribution with conditional mean $\text{E}(Y_i|\lambda_i) = \mu_i$ and that the parameter, λ_i, follows a gamma distribution with mean $\text{E}(\lambda_i) = \mu_i$ and variance, $\text{var}(\lambda_i) = \mu_i^2\nu_i^{-1}$. The marginal distribution of Y_i can be shown to follow a negative binomial distribution, by integrating out λ_i, with probability mass function

$$\begin{aligned} Pr(Y_i = y_i) &= \int Pr(Y_i = y_i|\lambda_i)f(\lambda_i)d\lambda_i \\ &= \frac{\Gamma(y_i + \nu_i)}{\Gamma(y_i + 1)\Gamma(\nu_i)} \left(\frac{\nu_i}{\nu_i + \mu_i}\right)^{\nu_i} \left(\frac{\mu_i}{\nu_i + \mu_i}\right)^{y_i}. \end{aligned} \tag{5.4}$$

The marginal mean is then given by $\text{E}(Y_i) = \mu_i$ and the marginal variance is given by $\text{var}(Y_i) = \mu_i + \mu_i^2\nu_i^{-1}$. This can be re-parameterized by letting $\nu_i^{-1} = \alpha_i$, such that the marginal mean is given by $\text{E}(Y_i) = \mu_i$ and the marginal variance is given by

$$\text{var}(Y_i) = \mu_i + \alpha_i\mu_i^2 = \mu_i(1 + \alpha_i\mu_i),$$

where α_i represents the dispersion parameter with $\alpha_i \geq 0$ (Bissell, 1972). This dispersion parameter has the same interpretation as ϕ in (5.3). In practice the dispersion parameter α_i is assumed to be constant $\forall i = 1, \ldots, n$ thus $\alpha_i = \alpha$ and the marginal variance is given by

$$\text{var}(Y_i) = \mu_i + \alpha\mu_i^2 = \mu_i(1 + \alpha\mu_i).$$

Lee and Nelder (1996) showed that it is possible to get a similar parameterization by using a HGLM approach with saturated random effects, that is assuming that each observation is a random effect following a gamma

distribution. The linear predictor for this HGLM is given by

$$\log[\mathrm{E}(Y_i|s_i)] = \mathbf{x}_i\boldsymbol{\beta} + \nu_{s_i},$$

where $\mathbf{x}_i$ is the i^{th} row of the $n \times p$ design matrix of covariates with $\boldsymbol{\beta}$ being a p-dimensional vector of fixed coefficients, and s_i is the random effect for the i^{th} observation. Following Lee and Nelder (1996) $\nu_{s_i} = \log(s_i)$, with s_i following a gamma distribution with a mean of one and variance of α_i. During the model fitting process it is assumed that α_i is constant, thus $\alpha_i = \alpha \ \forall i = 1, \ldots, n$. As a result the linear predictor can be written as

$$\log[\mathrm{E}(Y_i|s)] = \mathbf{x}_i\boldsymbol{\beta} + \nu_s,$$

where s is the random effect for each individual observation, this will be referred to as the observation random effect throughout this thesis. Using iterated expectations the marginal variance of Y_i is found to be

$$\mathrm{var}(Y_i) = \mu_i + \alpha\mu_i^2 = \mu_i(1 + \alpha\mu_i). \tag{5.5}$$

Lee and Nelder (2000) combined the variance functions (5.3) and (5.5) to get $\mathrm{var}(Y_i) = \phi\mu_i + \alpha\mu_i^2$ by using a quasi-likelihood model for the distribution of $Y_i|s$. It can thus be seen that the measure of the variance shift arising from overdispersion is quantified by the term $\alpha\mu_i^2$. In this thesis I will consider the shift in variance as being the variance inflation caused by the i^{th} observation.

5.2 Variance shift outlier model (VSOM) for Poisson count data

Once a Poisson GLM is fitted to the data, a residual analysis is often used to identify outliers and/or influential observations in a dataset. Often outliers can be corrected or removed from the data and the analysis redone. However, in most cases they are anomalous and we may want to include them in the analysis. In this section a VSOM for both Poisson independent and longitudinal count data will be introduced. This model can be viewed as a model for overdispersion associated with the i^{th} observation, with observations having large overdispersion considered as potential outliers. These models will be fit using the HGLM approach of Lee and Nelder (1996).

5.2.1 A VSOM for independent count data

The VSOM for Poisson independent count data will be formulated as a Poisson-gamma HGLM. HGLMs were described in chapter 2.3. The null model, that is a model with no outliers, is given by

$$\log[\mathrm{E}(Y_i)] = \log(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}, \tag{5.6}$$

where $\mathbf{X}_i$ is the i^{th} row of the $n \times p$ design matrix of covariates with $\boldsymbol{\beta}$ being a p-dimensional vector of fixed coefficients. This model can also be written as

$$\mathrm{E}(Y_i) = \exp(\mathbf{X}_i\boldsymbol{\beta}) = \mu_i.$$

A VSOM for the i^{th} count, Y_i, is given by

$$\log[\mathrm{E}(Y_i|\delta_i)] = \mathbf{X}_i\boldsymbol{\beta} + \nu_{\delta_i}, \tag{5.7}$$

where δ_i is a random effect for the i^{th} count. Following Lee and Nelder (1996) it is assumed that $\nu_{\delta_i} = \log(\delta_i)$, with δ_i following a gamma distribution with a mean of one and variance of λ_i; hence model (5.7) is a Poisson-gamma HGLM (Lee and Nelder, 1996). This model is an extension of model (5.6) with the addition of the term ν_{δ_i}. It can be seen that model (5.6) is a marginal model and model (5.7) is a conditional model. It is assumed that conditional on δ_i, the outcome Y_i follows a Poisson distribution with a variance of $\phi\mathrm{E}(Y_i|\delta_i)$. The marginal properties of Y_i are derived using iterated expectations. The marginal mean of Y_i is given by

$$\begin{aligned}
\mathrm{E}(Y_i) &= \mathrm{E}[\mathrm{E}(Y_i|\delta_i)] \\
&= \mathrm{E}[\exp(\mathbf{X}_i\boldsymbol{\beta} + \nu_{\delta_i})] \\
&= \exp(\mathbf{X}_i\boldsymbol{\beta})\mathrm{E}[\exp(\nu_{\delta_i})] \\
&= \exp(\mathbf{X}_i\boldsymbol{\beta})\mathrm{E}(\delta_i) \\
&= \exp(\mathbf{X}_i\boldsymbol{\beta}) \\
&= \mu_i.
\end{aligned}$$

The marginal variance is given by

$$\begin{aligned}
\text{var}(Y_i) &= \text{E}[\text{var}(Y_i|\delta_i)] + \text{var}[\text{E}(Y_i|\delta_i)] \\
&= \phi\text{E}[\text{E}(Y_i|\delta_i)] + \text{var}[\text{E}(Y_i|\delta_i)] \\
&= \phi\text{E}[\exp(\mathbf{x}_i\boldsymbol{\beta} + \nu_{\delta_i})] + \text{var}[\exp(\mathbf{x}_i\boldsymbol{\beta} + \nu_{\delta_i})] \\
&= \phi\exp(\mathbf{x}_i\boldsymbol{\beta})\text{E}[\exp(\nu_{\delta_i})] + \exp(2\mathbf{x}_i\boldsymbol{\beta})\text{var}[\exp(\nu_{\delta_i})] \\
&= \phi\mu_i\text{E}(\delta_i) + \mu_i^2\text{var}(\delta_i) \\
&= \phi\mu_i + \lambda_i\mu_i^2. \qquad (5.8)
\end{aligned}$$

It can thus be seen that if $\phi = 1$, expression (5.8) resembles the marginal variance for the negative binomial model, as given in (5.5); and if $\phi = 1$ and $\lambda_i = 0$ then expression (5.8) is the marginal variance for the Poisson GLM. The size of the variance shift of the i^{th} observation is $\lambda_i\mu_i^2$, which is proportional to the dispersion parameter of the random effect due to the i^{th} observation (λ_i). The difference between (5.5) and (5.8) is that the dispersion parameter in (5.5), α, accommodates the overdispersion for all the observations in the dataset while λ_i accounts for the overdispersion due the i^{th} observation only. Observations with relatively large values of λ_i are indicative of potential outliers.

5.2.2 A VSOM for outlying observations in longitudinal count data

The linear predictor for a quasi-Poisson-gamma model, which has subjects as the only random effects, is given by the HGLM described in chapter 2.3. It takes the form

$$\eta_{ij} = \log[\text{E}(Y_{ij}|b_i)] = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i}, \qquad (5.9)$$

for $i = 1, \ldots, q$ and for $j = 1, \ldots, n_i$, where q is the number of subjects in the study, $\mathbf{x}_{ij}^{'}$ is the j^{th} row of the design matrix $\mathbf{X}_i$, where $\mathbf{X}_i$ is a $n_i \times p$ design matrix of covariates, with $\boldsymbol{\beta}$ being a p-dimensional vector of fixed effect coefficients. Following Lee and Nelder (1996) $\nu_{b_i} = \log(b_i)$ where b_i follows a gamma distribution with a mean of one and a variance of γ_i. In the estimation procedure γ_i is assumed to be constant $\forall i = 1, \ldots, q$, thus $\gamma_i = \gamma$. The conditional variance of $Y_{ij}|b_i$ is given by $\phi\text{E}(Y_{ij}|b_i)$. The

marginal properties of the j^{th} observation of the i^{th} subject are derived using iterated expectations. The marginal mean of Y_{ij} is given by

$$\begin{aligned}
\mathrm{E}(Y_{ij}) &= \mathrm{E}[\mathrm{E}(Y_{ij}|b_i)] \\
&= \mathrm{E}[\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i})] \\
&= \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta})\mathrm{E}(b_i) \\
&= \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta}) \\
&= \mu_{ij}.
\end{aligned}$$

The marginal variance is given by

$$\begin{aligned}
\mathrm{var}(Y_{ij}) &= \mathrm{E}[\mathrm{var}(Y_{ij}|b_i)] + \mathrm{var}[\mathrm{E}(Y_{ij}|b_i)] \\
&= \phi\mathrm{E}[\mathrm{E}(Y_{ij}|b_i)] + \mathrm{var}[\mathrm{E}(Y_{ij}|b_i)] \\
&= \phi\mathrm{E}[\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i})] + \mathrm{var}[\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i})] \\
&= \phi\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta})\mathrm{E}[\exp(\nu_{b_i})] + \exp(2\mathbf{x}_{ij}^{'}\boldsymbol{\beta})\mathrm{var}[\exp(\nu_{b_i})] \\
&= \phi\mu_{ij}\mathrm{E}(b_i) + \mu_{ij}^2\mathrm{var}(b_i) \\
&= \phi\mu_{ij} + \gamma\mu_{ij}^2, \qquad (5.10)
\end{aligned}$$

if $\phi = 1$ the marginal variance will be for a Poisson-gamma model.

A VSOM for the j^{th} observation of the i^{th} subject takes the form

$$\eta_{ij} = \log[\mathrm{E}(Y_{ij}|b_i, \delta_{ij})] = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i} + \nu_{\delta_{ij}}, \qquad (5.11)$$

where δ_{ij} is a random effect for the j^{th} observation of the i^{th} subject. It is assumed that $\nu_{\delta_{ij}} = \log(\delta_{ij})$ where δ_{ij} follows a gamma distribution with a mean of one and a variance of λ_{ij} (Lee and Nelder, 1996). The random effects are also assumed to be independent of each other. The conditional variance of $Y_{ij}|b_i, \delta_{ij}$ is given by $\phi\mathrm{E}(Y_{ij}|b_i, \delta_{ij})$. The marginal properties of the j^{th} observation of the i^{th} subject are derived using iterated expectations. In order to derive the marginal properties of Y_{ij} it is easier to find the conditional properties of $Y_{ij}|\delta_{ij}$ first, then to proceed to use iterated expectations

to find the marginal properties. The conditional mean of $Y_{ij}|\delta_i$ is given by

$$\begin{aligned}
\mathrm{E}(Y_{ij}|\delta_{ij}) &= \mathrm{E}[\mathrm{E}(Y_{ij}|b_i, \delta_{ij})] \\
&= \mathrm{E}[\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i} + \nu_{\delta_{ij}})] \\
&= \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{\delta_{ij}})\mathrm{E}(b_i) \\
&= \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{\delta_{ij}}),
\end{aligned}$$

the marginal mean of Y_{ij} is then given by

$$\begin{aligned}
\mathrm{E}(Y_{ij}) &= \mathrm{E}[\mathrm{E}(Y_{ij}|\delta_{ij})] \\
&= \mathrm{E}[\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{\delta_{ij}})] \\
&= \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta})\mathrm{E}(\delta_{ij}) \\
&= \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta}) \\
&= \mu_{ij}.
\end{aligned}$$

The conditional variance of $Y_{ij}|\delta_{ij}$ is given by

$$\begin{aligned}
\mathrm{var}(Y_{ij}|\delta_{ij}) &= \mathrm{E}[\mathrm{var}(Y_{ij}|b_i, \delta_{ij})] + \mathrm{var}[\mathrm{E}(Y_i|b_i, \delta_{ij})] \\
&= \phi\mathrm{E}[\mathrm{E}(Y_{ij}|b_i, \delta_{ij})] + \mathrm{var}[\mathrm{E}(Y_{ij}|b_i, \delta_{ij})] \\
&= \phi\mathrm{E}[\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i} + \nu_{\delta_{ij}})] + \mathrm{var}[\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i} + \nu_{\delta_{ij}})] \\
&= \phi\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{\delta_{ij}})\mathrm{E}[\exp(\nu_{b_i})] + \exp(2\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + 2\nu_{\delta_{ij}})\mathrm{var}[\exp(\nu_{b_i})] \\
&= \phi\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{\delta_{ij}})\mathrm{E}(b_i) + \exp(2\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + 2\nu_{\delta_{ij}})\mathrm{var}(b_i) \\
&= \phi\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{\delta_{ij}}) + \gamma\exp(2\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + 2\nu_{\delta_{ij}}),
\end{aligned}$$

with the marginal variance of Y_{ij} given by

$$\begin{aligned}
\mathrm{var}(Y_{ij}) &= \mathrm{E}[\mathrm{var}(Y_{ij}|\delta_{ij})] + \mathrm{var}[\mathrm{E}(Y_i|\delta_{ij})] \\
&= \mathrm{E}\{\phi\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{\delta_{ij}}) + \gamma\exp(2\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + 2\nu_{\delta_{ij}})\} + \mathrm{var}[\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{\delta_{ij}})] \\
&= \phi\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta})\mathrm{E}[\exp(\nu_{\delta_{ij}})] + \gamma\exp(2\mathbf{x}_{ij}^{'}\boldsymbol{\beta})\mathrm{E}[\exp(2\nu_{\delta_{ij}})] + \exp(2\mathbf{x}_{ij}^{'}\boldsymbol{\beta})\mathrm{var}[\exp(\nu_{\delta_{ij}})] \\
&= \phi\mu_{ij}\mathrm{E}(\delta_{ij}) + \gamma\mu_{ij}^2\{\mathrm{var}(\delta_{ij}) + [\mathrm{E}(\delta_{ij})]^2\} + \mu_{ij}^2\mathrm{var}(\delta_{ij}) \\
&= \phi\mu_{ij} + \gamma\mu_{ij}^2[\lambda_{ij} + 1] + \mu_{ij}^2\lambda_{ij} \\
&= \phi\mu_{ij} + \gamma\mu_{ij}^2 + \mu_{ij}^2\lambda_{ij}[\gamma + 1]. \qquad (5.12)
\end{aligned}$$

Once again the size of the variance inflation of the j^{th} observation of the i^{th} subject depends on the dispersion parameter due to the δ_{ij} random effect. If $\phi = 1$ expression (5.12) is the marginal variance of a Poisson-gamma VSOM. Observations with relatively large values of λ_{ij} are indicative of outliers. If $\lambda_{ij} = 0$ then the marginal variance of (5.12) is just the marginal variance of a quasi-Poisson-gamma HGLM as shown in expression (5.10).

In section 5.5 when presenting the results of the VSOM for individual observations in longitudinal data, I will consider the j^{th} observation of the i^{th} subject as being the k^{th} unit of the n-dimensional vector of counts $\mathbf{Y}$. Thus I will present my results in terms of the k^{th} unit. This implies that units with relatively large values of λ_k are indicative of outliers.

5.2.3 A VSOM for outlying subjects in clustered count data

Considering model (5.9) as the null model once again, with all assumptions previously stated being applied, the linear predictor is given by

$$\eta_{ij} = \log[\mathrm{E}(Y_{ij}|b_i)] = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i}.$$

The marginal properties will be

$$\mathrm{E}(Y_{ij}) = \mu_{ij},$$

and

$$\mathrm{var}(Y_{ij}) = \phi\mu_{ij} + \gamma\mu_{ij}^2,$$

as shown section 5.2.2.

A VSOM for detecting outlying subjects can be formulated as

$$\eta_{ij} = \log[\mathrm{E}(Y_{ij}|b_i)] = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i} + \nu_{\zeta_i}, \tag{5.13}$$

for $i = 1, \ldots, q$, where q is the number of subjects in the study and ζ_i is a random effect specific to all observations belonging to the i^{th} subject. It is assumed that $\nu_{\zeta_i} = \log(\zeta_i)$ where ζ_i follows a gamma distribution with a mean of one and a variance of ψ_i (Lee and Nelder, 1996). The random effects are also assumed to be independent of each other. The conditional variance of $Y_{ij}|b_i, \zeta_i$ is given by $\phi\mathrm{E}(Y_{ij}|b_i, \zeta_i)$. The marginal properties of Y_{ij} are derived using iterated expectations in the same fashion as in section

5.2.2. I will not go into as much detail in deriving the marginal properties of Y_{ij} in this section. The conditional mean of $Y_{ij}|\zeta_i$ is given by

$$\begin{aligned}\mathrm{E}(Y_{ij}|\zeta_i) &= \mathrm{E}[\mathrm{E}(Y_{ij}|b_i,\zeta_i)] \\ &= \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta}+\nu_{\zeta_i}),\end{aligned}$$

the marginal mean of Y_{ij} is then given by

$$\begin{aligned}\mathrm{E}(Y_{ij}) &= \mathrm{E}[\mathrm{E}(Y_{ij}|\zeta_i)] \\ &= \mu_{ij}.\end{aligned}$$

The conditional variance of $Y_{ij}|\zeta_i$ is given by

$$\begin{aligned}\mathrm{var}(Y_{ij}|\zeta_i) &= \mathrm{E}[\mathrm{var}(Y_{ij}|b_i,\zeta_i)] + \mathrm{var}[\mathrm{E}(Y_i|b_i,\zeta_i)] \\ &= \phi\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta}+\nu_{\zeta_i}) + \gamma\exp(2\mathbf{x}_{ij}^{'}\boldsymbol{\beta}+2\nu_{\zeta_i}),\end{aligned}$$

with the marginal variance of Y_{ij} given by

$$\begin{aligned}\mathrm{var}(Y_{ij}) &= \mathrm{E}[\mathrm{var}(Y_{ij}|\zeta_i)] + \mathrm{var}[\mathrm{E}(Y_i|\zeta_i)] \\ &= \phi\mu_{ij} + \gamma\mu_{ij}^2 + \mu_{ij}^2\psi_i[\gamma+1]. \qquad (5.14)\end{aligned}$$

Subjects with relatively large values of ψ_i are indicative of potentially outlying subjects. The difference between expressions (5.12) and (5.14) is that in expression (5.12), λ_{ij} accounts for the overdispersion due to a specific j^{th} observation belonging to the i^{th} subject; while in expression (5.14) ψ_i accounts for the overdispersion of all the observations belonging to the i^{th} subject.

The formulation of the VSOM when applied to count data in both the independent and clustered cases can be summarized by Table 5.1. From this table the size of the variance for the i^{th} observation or subject is shown to depend on the dispersion parameters of the δ_{ij} and ζ_i random effects, thus allowing these dispersion parameters to be used as the size of the variance shift. It is thus evident that the variance of the i^{th} observation or subject will be inflated by a quantity dependent on λ_i (independent data) and ψ_i (longitudinal data) with the covariances and variances of the other observations remaining the same.

Table 5.1: **Table of marginal variances for Y_i (independent data) and Y_{ij} (longitudinal data) for different models.**

Type of data	Null model	VSOM
Independent data	μ_i	$\mu_i + \lambda_i\mu_i^2$
	$\phi\mu_i$	$\phi\mu_i + \lambda_i\mu_i^2$
	$* \ \mu_i + \alpha\mu_i^2$	$\mu_i + \alpha\mu_i^2 + \mu_i^2\lambda_i[\alpha+1]$
Clustered data (observations)	$\mu_{ij} + \gamma\mu_{ij}^2$	$\mu_{ij} + \gamma\mu_{ij}^2 + \mu_{ij}^2\lambda_{ij}[\gamma+1]$
	$\phi\mu_{ij} + \gamma\mu_{ij}^2$	$\phi\mu_{ij} + \gamma\mu_{ij}^2 + \mu_{ij}^2\lambda_{ij}[\gamma+1]$
	$\star \ \mu_{ij} + \alpha\mu_{ij}^2 + \mu_{ij}^2\gamma[\alpha+1]$	$\mu_{ij} + \alpha\mu_{ij}^2 + \mu_{ij}^2\gamma[\alpha+1] + \mu_{ij}^2\lambda_{ij}[\alpha\gamma+\alpha+\gamma+1]$
Clustered data (subjects)	$\mu_{ij} + \gamma\mu_{ij}^2$	$\mu_{ij} + \gamma\mu_{ij}^2 + \mu_{ij}^2\psi_i[\gamma+1]$
	$\phi\mu_{ij} + \gamma\mu_{ij}^2$	$\phi\mu_{ij} + \gamma\mu_{ij}^2 + \mu_{ij}^2\psi_i[\gamma+1]$
	$\star \ \mu_{ij} + \alpha\mu_{ij}^2 + \mu_{ij}^2\gamma[\alpha+1]$	$\mu_{ij} + \alpha\mu_{ij}^2 + \mu_{ij}^2\gamma[\alpha+1] + \mu_{ij}^2\psi_i[\alpha\gamma+\alpha+\gamma+1]$

$*$ denotes the marginal variance considering a negative binomial null model

$\star$ denotes the marginal variance for a longitudinal negative binomial null model

5.3 Poisson-Normal VSOM

Lee and Nelder (1996) stated that the Poisson-Normal HGLM is equivalent to a Poisson GLMM. The VSOM can be extended to a Poisson GLMM where the random effects are assumed to be normally distributed, as opposed to the gamma distribution assumed in section 5.2.

I will consider the case of applying the VSOM to independent responses and thus adding only one random effect, δ_i, to the linear predictor of the Poisson GLM (as given by expression (5.6)) with $\delta_i \sim N(0, \lambda_i)$. The VSOM will be of the form

$$\log[\mathrm{E}(Y_i|\delta_i)] = \mathbf{x}_i\boldsymbol{\beta} + \delta_i. \tag{5.15}$$

Using iterated expectations the marginal properties of Y_i are given by

$$\mathrm{E}(Y_i) = \mu_i = \exp\left(\mathbf{x}_i'\boldsymbol{\beta} + \frac{1}{2}\lambda_i\right), \tag{5.16}$$

and

$$\mathrm{var}(Y_i) = \mu_i + \mu_i^2(e^{\lambda_i} - 1). \tag{5.17}$$

As a result, the variance inflation of the i^{th} observation is quantified by the term $\mu_i^2(e^{\lambda_i} - 1)$ which is proportional to the size of the variance of the random effect (λ_i).

In the case of repeated measurements the size of the variance inflation is once again linked to the variance of the random effect for the i^{th} subject and this can easily be found after model fitting using an appropriate statistical program.

In conclusion it has been shown that for both the cases where the δ_i random effect has followed a normal and a gamma distribution the size of the variance inflation depends on the size of the λ_i dispersion parameter. As a result it would make sense to use the dispersion parameter of the random effect for the i^{th} observation or subject as a proxy for the variance inflation.

5.4 Hypothesis tests on variance shift parameters

When analyzing longitudinal data, I will consider the j^{th} observation of the i^{th} subject as being the k^{th} observation of the n-dimensional vector of responses $\mathbf{Y}$. Thus I will present my results in terms of the k^{th} unit. In order to test whether the effect of adding a random effect for the k^{th} observation is statistically significant, I will use the likelihood ratio test (LRT) statistic given as

$$\text{LRT}_k = -2[p_{\boldsymbol{\beta\nu}}(h) - p_{\boldsymbol{\beta\nu}\lambda_k}(h)], \tag{5.18}$$

where $p_{\boldsymbol{\beta\nu}}(h)$ is the adjusted profile likelihood of both fixed ($\boldsymbol{\beta}$) and random effects ($\boldsymbol{\nu}$) simultaneously, that is both the fixed and random parameter estimates are being profiled out, while the term $p_{\boldsymbol{\beta\nu}\lambda_k}(h)$ includes the estimation of the additional random effect from the use of the VSOM (λ_k) and h is the h-likelihood. The likelihood ratio test (LRT) statistic given in (5.18) is used to test the hypothesis $H_0 : \lambda_k = 0$ against $H_1 : \lambda_k > 0$. For longitudinal data we could test the additional hypothesis $H_0 : \psi_i = 0$ against $H_1 : \psi_i > 0$, using a modified version of (5.18) given as

$$\text{LRT}_i = -2[p_{\boldsymbol{\beta\nu}}(h) - p_{\boldsymbol{\beta\nu}\psi_i}(h)], \tag{5.19}$$

Lee et al. (2006) recommended the use of the adjusted profile likelihood when testing for variance components. In the case of independent data, the test statistic in equation (5.18) does not involve ν, the subject random effect variance, since there are no subject random effects. Thus the likelihood ratio

test statistic for the i^{th} observation is given as

$$\text{LRT}_i = -2[p_{\boldsymbol{\beta}}(h) - p_{\boldsymbol{\beta}\lambda_i}(h)], \tag{5.20}$$

It is not possible to use standard asymptotic theory for the distribution of LRT statistics due to the fact that the null hypothesis of the test falls on the boundary of the parameter space, that is $H_0 : \lambda_i = 0$; as a result the regularity conditions are not met (Self and Liang, 1987). If the data values or subsets of the data can be assumed to be independent and identically distributed then Stram and Lee (1995) showed that the asymptotic null distribution, of the LRT statistic for testing the null hypothesis, was a 50:50 mixture of two chi-squared distributions with zero and one degrees of freedom respectively. Simulation experiments by Pinheiro and Bates (2000) showed that the 65:35 chi-squared distribution mixture was more appropriate in linear mixed models. Gumedze et al. (2010) pointed out that this is not applicable for the VSOM as the variance shift depends on a single observation. In this thesis I will use a 68:32 mixture of chi-squared distributions of zero and one degrees of freedom respectively, as an approximation for the asymptotic distribution; as this was found to be the optimal mixture after simulation experiments for the VSOM by Gumedze et al. (2010). This approach does not account for the problem of multiple testing involved in the VSOM methodology. A parametric bootstrap can be used to get an empirical distribution of the LRT statistic and also to account for multiple testing. This is an area of further research which is not covered in this thesis.

5.5 Examples

In order to fit the data in this section I will use the **HGLM** procedure in the GENSTAT statistical system (Payne et al., 2011). The variance components are estimated on a log-scale which is based on the extension of the quasi-likelihood theory developed by (Nelder and Pregibon, 1987). The delta method is then used to obtain appropriate variances for the random components on the scale of the data.

In presenting the results of the use of the VSOM I will initially perform standard residual plot analysis on the original model in order to identify

potential outliers. The deviance residuals are used because they provide a good approximation to normality for all general linear distributions, except for extreme cases like binary data (Pierce and Schafer, 1986), as a result normal plots can be used for model checking. The residual plots also consist of a plot of residuals against fitted values and a plot of absolute residuals against fitted values. If the model is adequate these two plots should show running means that are approximately straight and flat. If there is marked curvature in the plot of residuals against fitted values, there is either an unsatisfactory link function or missing terms in the linear predictor, or both. The choice of variance function is checked by the plot of absolute residuals against fitted values. If for example, this plot showed a marked downward trend, this would imply that the residuals are falling in absolute value as the mean increases, that is the assumed variance function is increasing too rapidly with the mean (Lee and Nelder, 2001).

After the residual analysis, I will apply the VSOM to the data and identify the outliers using the likelihood ratio test statistic compared to a 68:32 mixture of chi-squared distributions of zero and one degrees of freedom respectively. Index plots of the dispersion parameter ϕ, random effect λ_i (or ψ_i for longitudinal data) and LRT_i statistic under the VSOM will be provided.

I will use examples in this section which show how the VSOM can be applied to independent count data and longitudinal count data. The dataset used for independent count data is the fabric faults dataset (Bissell, 1972), the null model is fit with a negative binomial model. The longitudinal count datasets are the leukemia rats dataset (Myers et al., 2002) and the epilepsy dataset (Thall and Vail, 1990). The null models which were fit to the datasets were the quasi-Poisson-gamma and quasi-Poisson-normal HGLMs.

5.5.1 Fabric dataset

I applied the VSOM to the dataset from Bissell (1972), involving the number of faults in a bolt of fabric of length denoted as **l**. The total number of observations (n) in the dataset were 32. Lee et al. (2006) fitted the following

Poisson model to the data

$$\log[\mathrm{E}(Y_i)] = \beta_0 + x_i\beta_1,$$

where Y_i is the number of fabric faults in a bolt of fabric of length l_i, with $x_i = \log(l_i)$, for $i = 1, \ldots, 32$. This model gave a deviance of 64.5 with 30 degrees of freedom which shows that there is overdispersion. In order to account for the overdispersion either a quasi-Poisson (M_{F0}) model or a negative binomial (M_{F1}) model can be fitted to the data. These models were fitted using the **HGLM** procedure in GENSTAT (Payne et al., 2011) and the deviance statistics were 182.78 and 179.94 for the M_{F0} model and M_{F1} models, respectively. As a result the negative binomial model will be used when applying the VSOM. The negative binomial model is formulated as a Poisson-gamma HGLM with saturated random effects (Lee and Nelder, 2000), that is all observations are considered as random effects. The linear predictor of model M_{F1} is given as

$$\log[\mathrm{E}(Y_i|s)] = \beta_0 + x_i\beta_1 + \nu_s,$$

where s is the observation random effect, as defined in section 5.1.2. Following Lee and Nelder (1996) $\nu_s = \log(s)$, with s following a gamma distribution with a mean of one and variance of α. The marginal variance of Y_i from this model is given by

$$\mathrm{var}(Y_i) = \mu_i + \alpha\mu_i^2.$$

Standard residual analysis identified observations 13, 26 and 30 as being potential outliers as shown in Figure 5.1. The plot of residuals against fitted values has a marked curvature, thus implying that the link function is inadequate or there are missing terms in the linear predictor; or both errors have occurred. The histogram of residuals does not have any long tails, however it is not symmetrical thus the assumption of normality of residuals may be questionable. The choice of variance function was checked by the plot of absolute residuals against fitted values. This plot showed a marked upward trend, which implied that the residuals are rising in absolute value as the mean increases, thus the variance function is not increasing at the same rate as the mean. The normal qq plot is fairly linear though, thus the null model can be considered to be acceptable.

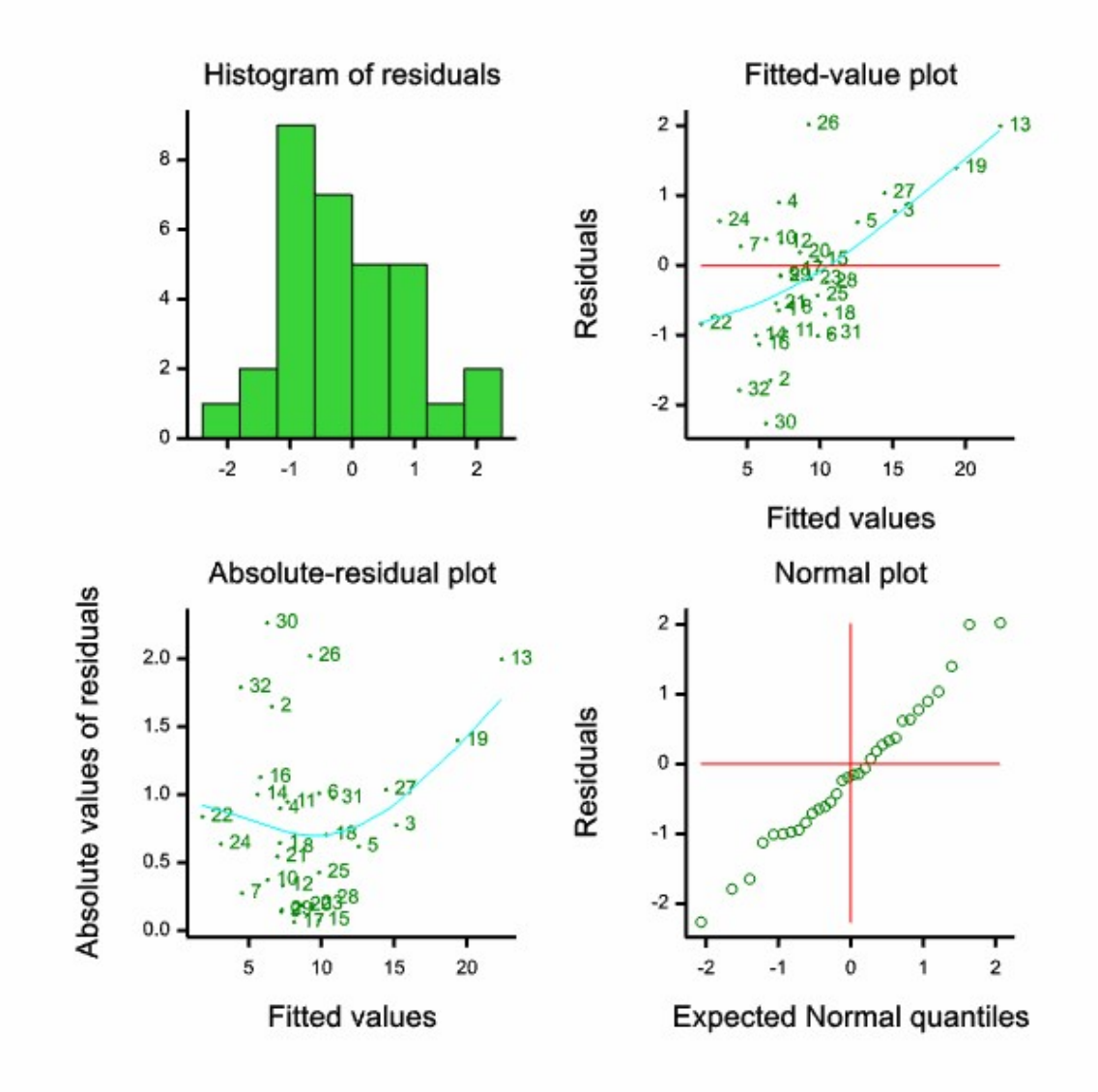


Figure 5.1: **Residual plots for model M_{F1} from the fabric dataset.**

The VSOM was applied to all observations in turn in the dataset. The VSOM for the i^{th} observation is given by

$$\log[\mathrm{E}(Y_i|s)] = \beta_0 + x_i\beta_1 + \nu_s + \nu_{\delta_i},$$

for $i = 1, \ldots, 32$ where δ_i is the random effect specific to the i^{th} observation, with $\nu_{\delta_i} = \log(\delta_i)$ and δ_i follows a gamma distribution with a mean of one and variance of λ_i. The marginal variance for the i^{th} observation is given by

$$\mathrm{var}(Y_i) = \mu_i + \alpha\mu_i^2 + \mu_i^2\lambda_i[\alpha + 1],$$

where the size of the variance shift for the i^{th} observation depends on the size of λ_i. This process identified observations 13 and 30 as outlying observations, this is shown in Figure 5.2. It must be noted that observation 26 also has a relatively large LRT_i value thus indicating that it is possibly an outlier, this is also supported by the large value of α_i corresponding to this observation. Figure 5.2 also shows that as the size of λ_i increases the size of α_i decreases and this coincides with observations with relatively large LRT_i values. For illustration I will only consider observations 13 and 30 as potential outliers. Model M_{F2} was fit using these outlying observations as random effects. Model M_{F2} can be written in matrix form as

$$\log(\mathrm{E}[\mathbf{Y}|s, \delta_{13}, \delta_{30}]) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu_s} + \nu_{\delta_{13}}\mathbf{d}_{13} + \nu_{\delta_{30}}\mathbf{d}_{30},$$

where $\mathbf{Y}$ is the vector of responses with length 32, δ_i is the random effect for the i^{th} observation and $\mathbf{d}_i$ is a vector with 1 in the i^{th} position and zeros in the other positions, $\mathbf{X}$ is a 32×2 design matrix of covariates with $\boldsymbol{\beta}$ being a vector of unknown fixed coefficients with length 2, $\mathbf{Z}$ is a 32×32 design matrix of the observation random effects (s). The parameter estimates and their standard errors from the fitted models are shown in Table 5.2.

From the results it can be seen that the estimates of the fixed effects for the negative binomial null model (M_{F1}) and the VSOM, applied to observations 13 and 30 only (M_{F2}), are similar thus the outlying observations are not influential observations. The only estimate which changes is the dispersion parameter (α) for the negative binomial model which decreases in size when the VSOM is used. Model M_{F3} is the model fitted when observations 13 and 30, which had been identified as potentially outlying, are deleted. It

Table 5.2: **Parameter estimates of models fitted to the fabric dataset.**

Parameter	M_{F0}	M_{F1}	M_{F2}	M_{F3}
constant	-4.173 (1.665)	-3.784 (1.440)	-3.349 (1.323)	-3.210 (1.333)
x	0.997 (0.258)	0.936 (0.225)	0.865 (0.207)	0.843 (0.209)
ϕ	2.151 (0.555)			
α		0.125 (0.384)	0.066 (0.030)	0.067 (0.030)
λ_{13}			0.591(0.887)	
λ_{30}			0.854 (1.332)	
deviance	182.78	179.94	175.46	160.49

where λ_i is the dispersion parameter of the i^{th} observation.

can be seen that the fixed effect estimates are similar when both models, M_{F2} and M_{F3}, are used.

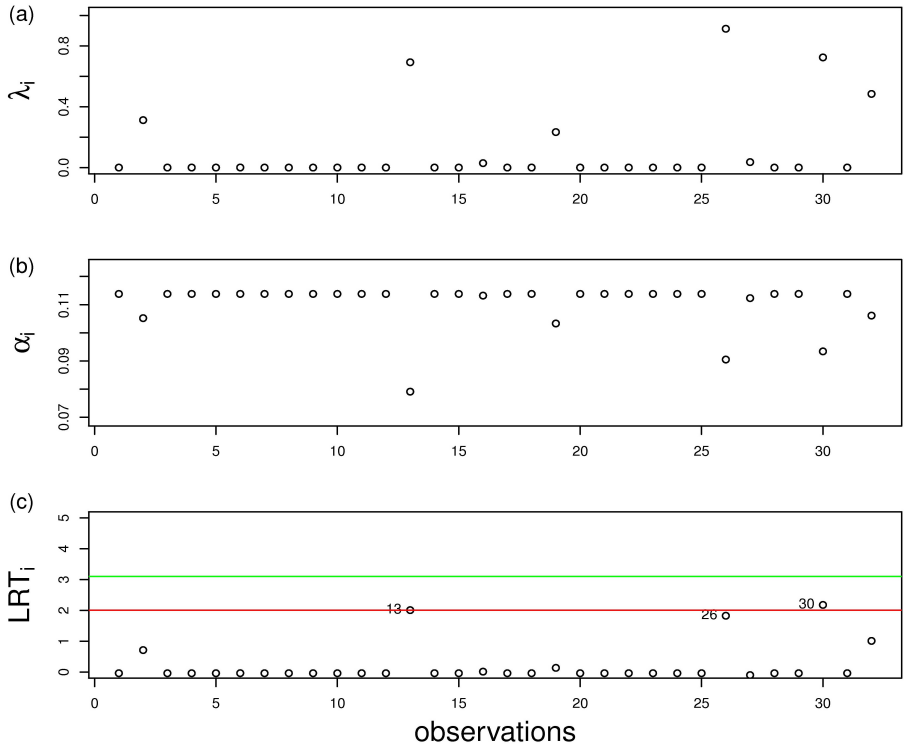


Figure 5.2: **VSOM statistics plotted against observation number for the fabric dataset.** (a) Variance shift estimates, λ_i. (b) Dispersion parameter estimates, α_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th} and 97.5^{th} percentile cut-off values.

5.5.2 Epilepsy dataset

In this example the VSOM will be applied to count data whilst assuming that the δ_k observation and ζ_i subject specific random effects follow a normal distribution.

5.5.2.1 VSOM for individual observations

In order to find the null model to use with the VSOM an ordinary Poisson GLMM (Poisson-normal HGLM with dispersion parameter $\phi = 1$) was compared to a quasi-Poisson-normal HGLM whose dispersion parameter was allowed to vary; these are models M_{E0} and M_{E1} respectively. From Table 5.3 it can be seen that model M_{E1} fits the data better than model M_{E0} as shown by its lower deviance of 1297.106 compared to 1343.859. The linear predictor of model M_{E1} is given by

$$\log[\mathrm{E}(Y_{ij}|b_i)] = \beta_0 + \beta_1 l_i + \beta_2 t_i + \beta_3 (t_i * l_i) + \beta_4 a_i + \beta_5 v_{ij} + b_i, \quad (5.21)$$

where Y_{ij} is the number of seizures experienced by the i^{th} subject at the j^{th} visit, for $i = 1, \ldots, 59$ and $j = 1, \ldots, 4$, l_i is the logarithm of a quarter of the number of epileptic seizures for the i^{th} subject recorded in the 8 week period preceding the trial, t_i is the type of treatment received by the i^{th} subject and this takes on values of 0 for the placebo and 1 for the new drug, a_i is the logarithm of the age of the i^{th} subject, v_{ij} is the linear trend for the visits with values of -0.3,-0.1,0.1 and 0.3, b_i is the subject random effect with $b_i \sim \mathrm{N}(0, \sigma_i^2)$ for $i = 1, \ldots, 59$.

Residual analysis to identify potentially outlying observations was performed after fitting the null model (M_{E1}) and the results are shown in Figure 5.3 . The histogram of residuals if symmetrical but it has a few outlying observations which give it slightly long tails. From the plot of residuals against fitted values it can be seen that observations 99, 221 and 97 are potentially outlying, however the running mean is fairly straight in the middle of the plot thus indicating that the model fitted is adequate. The normal plot is linear except for observation 99 which makes me assume that this observation might be an influential observation. The plot of absolute residuals against fitted values has an increasing running mean in the section where most of the

residuals are located in the plot, thus indicating that the assumed variance is not increasing at an adequate rate with the mean.

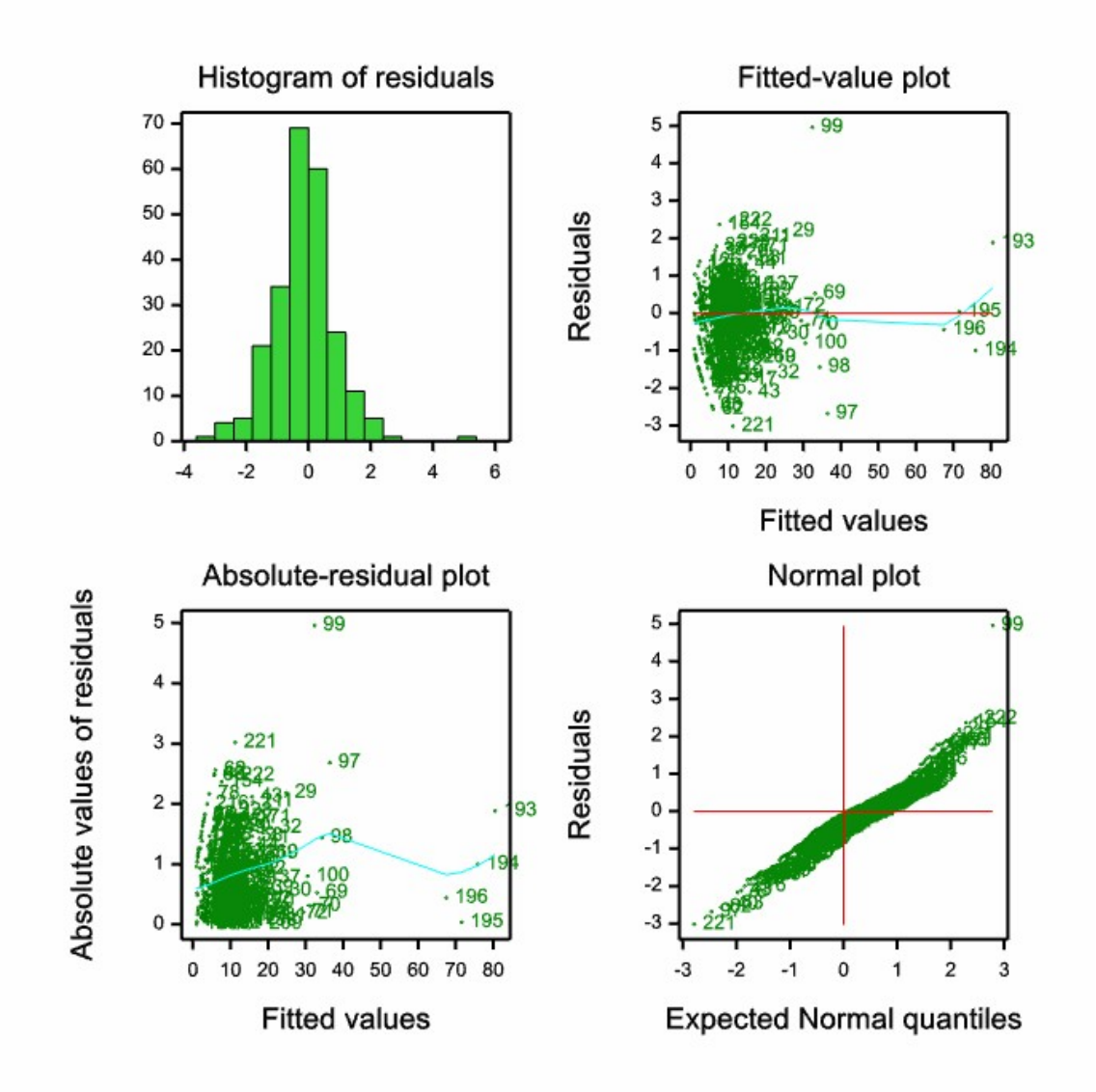


Figure 5.3: **Residual plots for model M_1 from the Epilepsy dataset.**

The VSOM was then applied to all the observations in turn. The VSOM for the j^{th} observation of the i^{th} subject takes the form

$$\log[\mathrm{E}(Y_{ij}|b_i)] = \beta_0 + \beta_1 l_i + \beta_2 t_i + \beta_3 (t_i * l_i) + \beta_4 a_i + \beta_5 v_{ij} + b_i + \delta_{ij},$$

where $\delta_{ij} \sim \mathrm{N}(\mathbf{0}, \lambda_{ij})$. As stated in section 5.2.2, when presenting the results

Table 5.3: **Parameter estimates of models fitted to the epilepsy dataset.**

Parameter	M_{E0}	M_{E1}	M_{E2}	M_{E3}	M_{E4}
constant	-1.296 (1.221)	-1.529 (1.225)	-1.418 (1.223)	-1.734 (1.203)	-1.662 (1.193)
lbase	0.872 (0.136)	0.884 (0.134)	0.838 (0.135)	0.883 (0.131)	0.838 (0.131)
treatment	-0.917 (0.413)	-0.914 (0.430)	-0.960 (0.422)	-0.809 (0.423)	-0.846 (0.413)
treatment*lbase	0.331 (0.210)	0.342 (0.211)	0.362 (0.210)	0.306 (0.207)	0.323 (0.205)
log(age)	0.472 (0.359)	0.512 (0.356)	0.527 (0.359)	0.603 (0.353)	0.600 (0.350)
visit	-0.294 (0.101)	-0.294 (0.152)	-0.357 (0.134)	-0.294 (0.150)	-0.357 (0.133)
γ	0.275 (0.058)	0.214 (0.051)	0.241 (0.055)	0.204 (0.049)	0.226 (0.052)
ϕ	1	2.236 (0.226)	1.641 (0.169)	2.189 (0.221)	1.599 (0.166)
λ_{40}			11.120 (18.348)		11.28 (18.500)
λ_{97}			0.033 (0.097)		0.033 (0.097)
λ_{99}			1.641 (2.237)		1.644 (2.351)
λ_{154}			1.494 (2.226)		1.480 (2.205)
λ_{221}			5.001 (7.75)		5.031 (7.798)
λ_{222}			0.418 (0.681)		0.419 (0.679)
ψ_{58}				11.94 (19.701)	14.90 (24.138)
deviance	1343.859	1297.106	1239.652	1291.594	1232.469

where γ is the dispersion parameter of the subject random effect, λ_k is the dispersion parameter of the k^{th} observation and ψ_i is the dispersion parameter of the i^{th} subject.

of the VSOM for individual observations in longitudinal data, I will consider the j^{th} observation of the i^{th} subject as being the k^{th} unit of the vector of counts $\mathbf{Y}$ with length 236. Thus I will present my results in terms of the k^{th} unit. This implies that units with relatively large values of λ_k are indicative of outliers for $k = 1, \ldots, 236$. The results of using the VSOM are shown in Figure 5.4. To illustrate the effect of applying the VSOM I will only consider observations which are potentially outlying at the 99^{th} percentile level and these are observations 40, 97, 99, 154, 221 and 222 (model M_{E2}). From this model it can be seen that the fixed effect estimates have changed, thus the effects of the outlying observations on the parameter estimates have been down-weighted. The deviance of this model is also substantially lower than the null model (M_{E1}). It can also be seen that the size of the dispersion parameter has decreased due to use of the VSOM (from 2.236 to 1.641) thus showing that the VSOM reduces the overall overdispersion of the model.

5.5.2.2 VSOM for subjects

The VSOM was then applied to all the subjects in turn. The VSOM for observations belonging to the i^{th} subject takes the form

$$\log[\mathrm{E}(Y_{ij}|b_i)] = \beta_0 + \beta_1 l_{ij} + \beta_2 t_{ij} + \beta_3 (t_{ij} * l_{ij}) + \beta_4 a_{ij} + \beta_5 v_{ij} + b_i + \zeta_i,$$

where $\zeta_i \sim \mathrm{N}(\mathbf{0}, \psi_i)$. The results from using the VSOM are shown in Figure 5.5. From this figure it can be seen that subjects 10, 25, 35, 56 and 58 are potentially outlying. For illustration purposes model M_{E3} was fit using subject 58 as a random effect and the results are shown in Table 5.3. This model was not better than model M_{E2} thus a model which had the individual observations and subject 58 as random effects (model M_{E4}) was fitted. This model fitted the data better than models M_{E1}, M_{E2} and M_{E3} with a deviance of 1232.469. The linear predictor for model M_{E4} in matrix form is given by

$$\begin{aligned}\log[\mathrm{E}(\mathbf{Y}|\mathbf{b}, \delta_k, \zeta_i)] =& \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \delta_{40}\mathbf{d}_{40} + \delta_{97}\mathbf{d}_{97} + \delta_{99}\mathbf{d}_{99} + \delta_{154}\mathbf{d}_{154} + \delta_{221}\mathbf{d}_{221} \\ &+ \delta_{222}\mathbf{d}_{222} + \zeta_{58}\mathbf{d}_{sub(58)},\end{aligned}$$

where δ_k is the random effect for the k^{th} observation and $\mathbf{d}_k$ is a vector of length 232 with 1 in the k^{th} position and zeros in the other positions; ψ_i

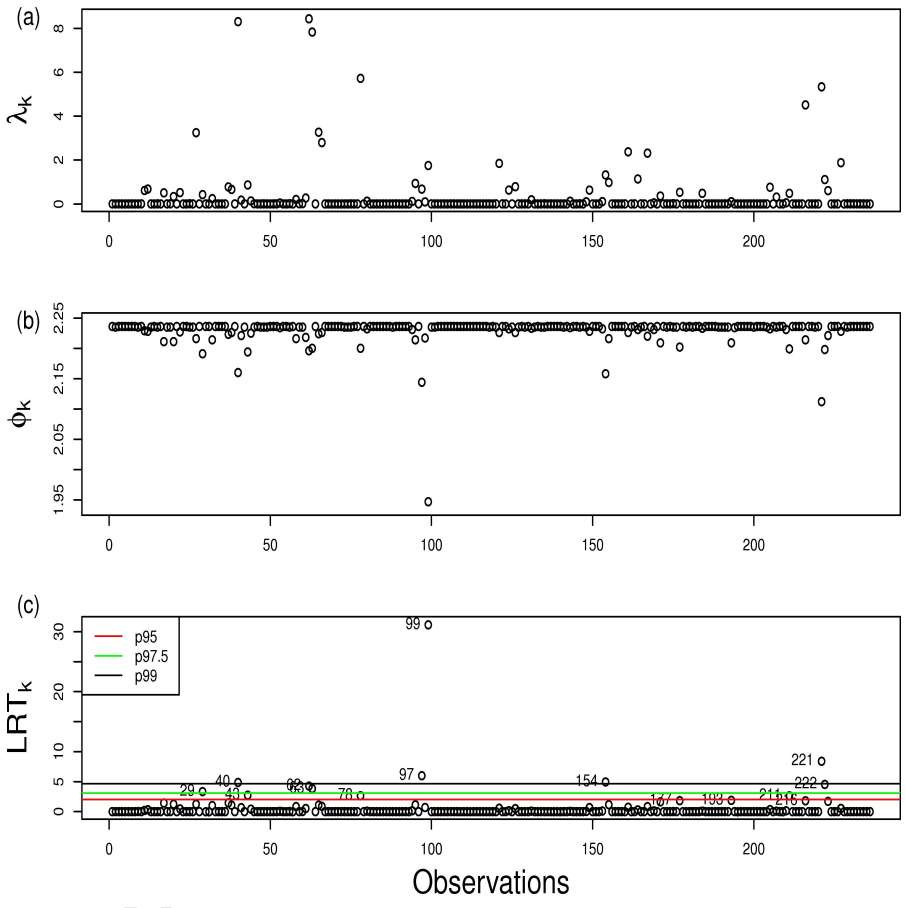


Figure 5.4: **VSOM statistics plotted against observation number for the epilepsy dataset.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, ϕ_k. (c) Likelihood ratio statistics, LRT_k, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values.

is the random effect for the i^{th} subject and $\mathbf{d}_{sub(i)}$ is a vector of length 232 with values of 1 in the positions corresponding to the i^{th} subject and zeros in the other positions, $\mathbf{Y}$ is the vector of responses with length 236, $\mathbf{X}$ is a 236×6 design matrix of covariates with $\boldsymbol{\beta}$ being a vector of unknown fixed coefficients with length 6, $\mathbf{Z}$ is a 59×59 design matrix of the subject random effects ($\mathbf{b}$). The fixed effects estimates for model M_{E4} can be seen to have changed significantly from the null model thus highlighting the down-weighting effect brought about by the VSOM. The use of the VSOM is also shown to reduce the overall overdispersion in the data, with a dispersion parameter of 1.599 compared to that of the null model which was 2.236.

An investigation into the effect of down-weighting a subject and then subsequently using the VSOM to identify additional outlying observations was also conducted. In this case the null model fit had subject 58 included as a random effect. Thus the null model was

$$\log[\mathrm{E}(\mathbf{Y}|\mathbf{b}, \delta_k, \zeta_i)] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \zeta_{58}\mathbf{d}_{sub(58)}.$$

Application of the VSOM identified observations 40, 97, 99, 154, 221 and 222 as potential outliers at the 99^{th} percentile cut off as shown in Figure 5.6. These were the same observations identified when the VSOM was applied using the null model, (5.21), which did not contain subject 58 as a random effect. It can thus be concluded that the results of the application of the VSOM to individual observations is not affected by the down-weighting of a potentially outlying subject.

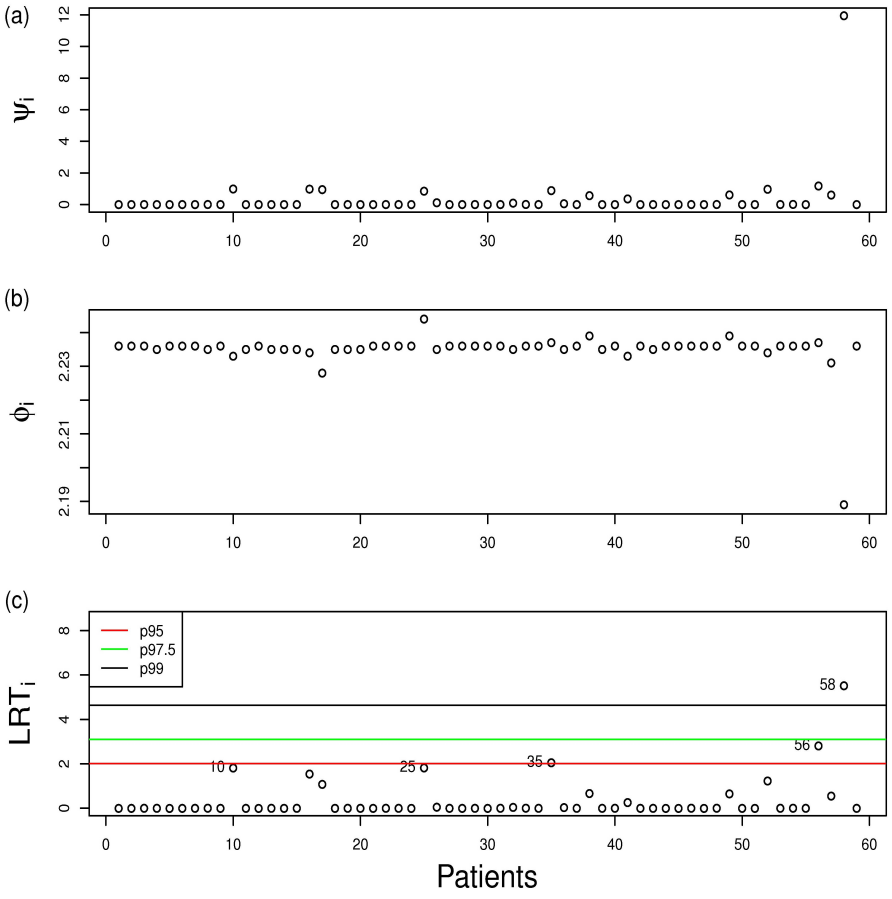


Figure 5.5: **VSOM statistics plotted against patient number for the epilepsy dataset.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, ϕ_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values.

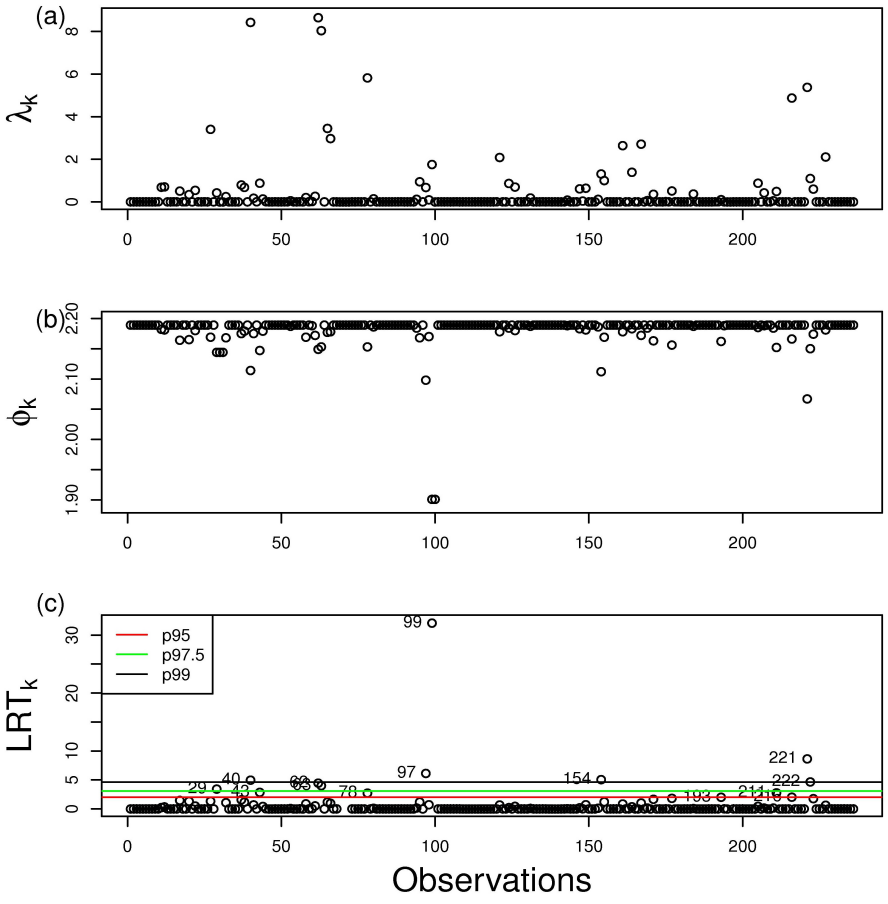


Figure 5.6: **VSOM statistics plotted against observation number for the epilepsy dataset.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, ϕ_k. (c) Likelihood ratio statistics, LRT_k, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values.

5.5.3 Leukemia rats dataset

The VSOM fitted to this dataset assumed that the δ_k observation and ζ_i subject specific random effects follow a gamma distribution, as opposed to the assumption of normally distributed random effects in the epilepsy data analysis.

5.5.3.1 VSOM for individual observations

The quasi-Poisson-gamma (model M_{R0}) HGLM was initially fitted to the data and the results are shown in Table 5.4. The linear predictor of model M_{R0} is given

$$\log[\mathrm{E}(Y_{ij}|b_i)] = \beta_0 + \beta_1 W_{ij} + \beta_2 R_{ij} + \beta_3 Drug2_{ij} + \beta_4 Drug3_{ij} + \nu_{b_i},$$

where Y_{ij} is the number of cancer colonies in the i^{th} subject at the j^{th} time, for $i = 1, \ldots, 30$ and $j = 1, \ldots, 4$, R_{ij} and W_{ij} are the red and white blood cell counts, respectively, for the i^{th} subject at the j^{th} time , $Drug2_{ij}$ is the effect of drug 2 relative to drug 1 for the i^{th} subject at the j^{th} time with $Drug2_{ij}$ given a value of 1 if the i^{th} subject at the j^{th} time is receiving drug 2 and it is coded as 0 otherwise, $Drug3_{ij}$ is the effect of drug 3 relative to drug 1 for the i^{th} subject at the j^{th} time with $Drug3_{ij}$ given a value of 1 if the i^{th} subject at the j^{th} time is receiving drug 3 and it is coded as 0 otherwise, b_i is the subject random effect with $\nu_{b_i} = \log(b_i)$ assuming b_i follows a gamma distribution with a mean of one and variance of γ for $i = 1, \ldots, 30$.

Model checking was then performed using residual analysis. The plot of residuals against fitted values and the absolute residuals against fitted values plot (Figure 5.7) showed that observations 17, 24, 28, 33, 36, 45, 77 and 80 were potential outliers. These plots also have a fairly straight and flat running mean which shows that this is an adequate model. The histogram of residuals was fairly symmetric and normally distributed, thus indicating that it is a decent model. The residual plots are shown in Figure 5.7.

The VSOM was then applied to all the observations in turn. The VSOM for the j^{th} observation of the i^{th} subject takes the form

$$\log[\mathrm{E}(Y_{ij}|b_i, \delta_{ij})] = \beta_0 + \beta_1 W_{ij} + \beta_2 R_{ij} + \beta_3 Drug2_{ij} + \beta_4 Drug3_{ij} + \nu_{b_i} + \nu_{\delta_{ij}},$$

Table 5.4: **Parameter estimates of models fitted to the leukemia rats dataset.**

Parameter	M_{R0}	M_{R1}	M_{R2}	M_{R3}	M_{R4}
constant	2.864 (0.123)	2.864 (0.123)	2.868 (0.112)	2.918 (0.103)	2.934 (0.102)
WBC	-0.019 (0.005)	-0.019 (0.005)	-0.020 (0.005)	-0.024 (0.005)	-0.025 (0.005)
RBC	0.030 (0.007)	0.030(0.007)	0.031(0.007)	0.032 (0.006)	0.032 (0.006)
Drug2	0.184 (0.152)	0.184 (0.152)	0.182 (0.135)	0.170 (0.124)	0.169 (0.121)
Drug3	0.0231 (0.153)	0.0231(0.153)	0.083(0.139)	0.089 (0.128)	0.102 (0.125)
ϕ	0.110 (0.016)	0.081 (0.012)	0.110 (0.017)	0.081 (0.013)	0.083 (0.013)
γ	0.113 (0.031)	0.10 (0.026)	0.088 (0.025)	0.075 (0.021)	0.072 (0.020)
λ_{24}		0.052 (0.079)		0.052 (0.079)	
λ_{33}		0.030 (0.049)		0.030 (0.049)	
λ_{36}		0.083 (0.122)		0.085 (0.124)	
λ_{45}		0.030 (0.045)		0.031 (0.046)	
λ_{80}		0.023 (0.034)		0.023 (0.034)	
ψ_{28}			0.088 (0.025)	0.564 (0.789)	
deviance	551.71	530.47	547.46	526.01	487.35

where γ is the dispersion parameter of the subject random effect, λ_k is the dispersion parameter of the k^{th} observation and ψ_i is the dispersion parameter of the i^{th} subject.

where $\nu_{\delta_{ij}} = \log(\delta_{ij})$ with δ_{ij} following a gamma distribution with a mean of 1 and variance of λ_{ij}. As previously stated, I will consider the j^{th} observation of the i^{th} subject as being the k^{th} unit of the n-dimensional vector of counts $\mathbf{Y}$. Thus my results will be presented in terms of the k^{th} unit. The results are shown in Figure 5.8. It can be seen from this illustration that as the size of λ_k increases, the dispersion parameter ϕ_k decreases. This implies that the VSOM is reducing the effect of the outlying observations on the overall overdispersion of the dataset.

The VSOM identified observations 24, 33, 36, 45 and 80 as being outliers. Model M_{R1}, which included these outlying observations as random effects, was then fit to the data and the results are shown in table 5.4.

From the results it can be seen that the estimates of the fixed effects in model M_{R1} remain the same as in the null model model M_{R0} in Table 5.4. The only estimates which change are the random effect estimates which show that ϕ and γ decrease in size. Since there is no change in the fixed estimates it can be seen that the identified outlying observations are not influential observations.

5.5.3.2 VSOM for subjects

The VSOM was then applied to all the subjects in turn. The VSOM for all observations belonging to the i^{th} subject takes the form

$$\log[\mathrm{E}(Y_{ij}|b_i, \zeta_i)] = \beta_0 + \beta_1 W_{ij} + \beta_2 R_{ij} + \beta_3 Drug2_{ij} + \beta_4 Drug3_{ij} + \nu_{b_i} + \nu_{\zeta_i},$$

where $\nu_{\zeta_i} = \log(\zeta_i)$ with ζ_i following a gamma distribution with a mean of 1 and variance of ψ_i. The results are shown in Figure 5.9. From this figure it can be seen that subject 28 is potentially outlying. Model M_{R2} was then fitted using subject 28 as a random effect and the results are shown in table 5.4. From this table it can be seen that the fixed effect estimates are unchanged thus indicating that subject 28 is not an influential subject.

This model was not better than model M_{R1}, thus a model which had the individual observations and subject 28 as being random effects (model M_{R3}) was fitted. This model fitted the data better than models M_{R1} and M_{R2} with a deviance of 526.01. The linear predictor for model M_{R3} is given

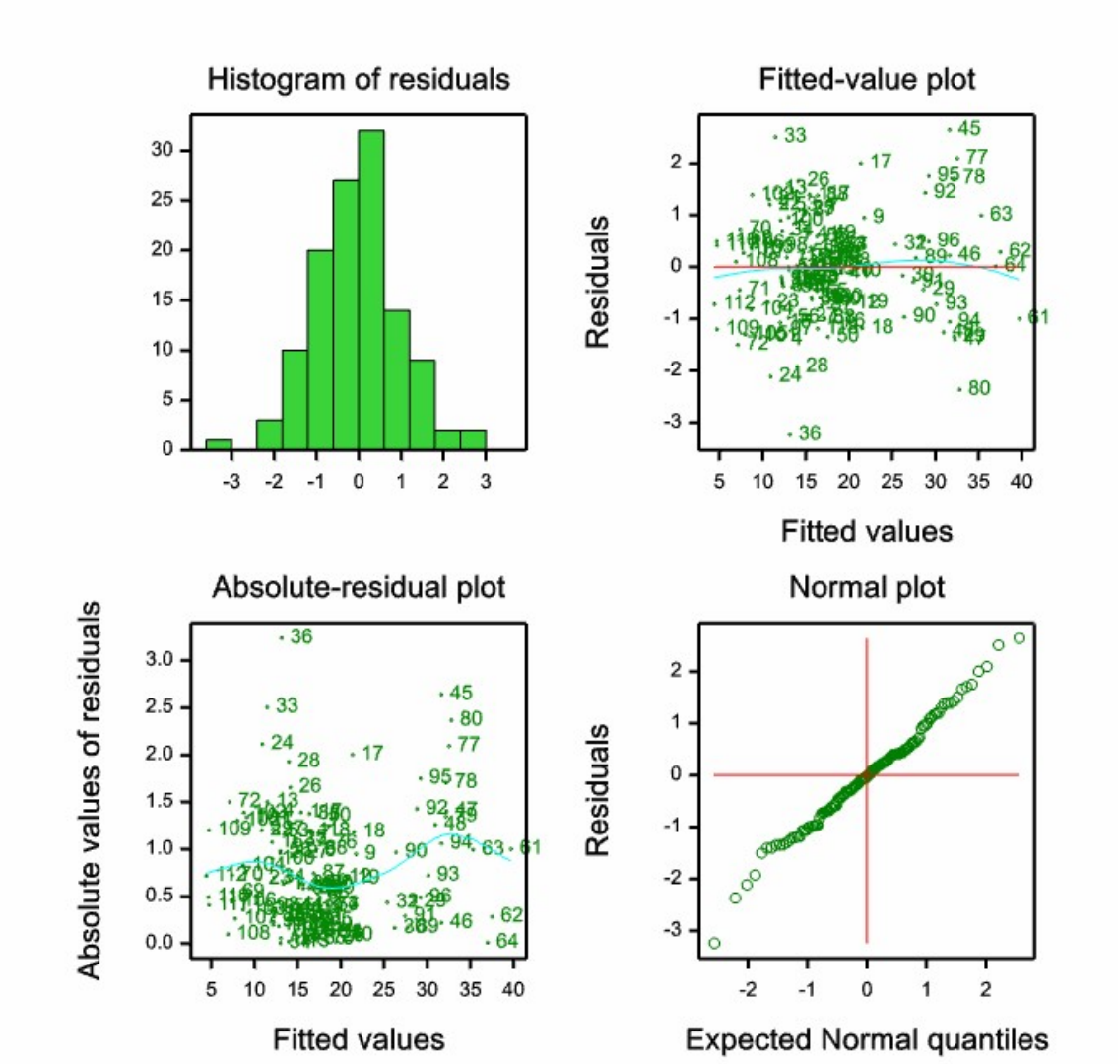


Figure 5.7: **Residual plots for observations from model M_{R0} from the leukemia rats dataset.**

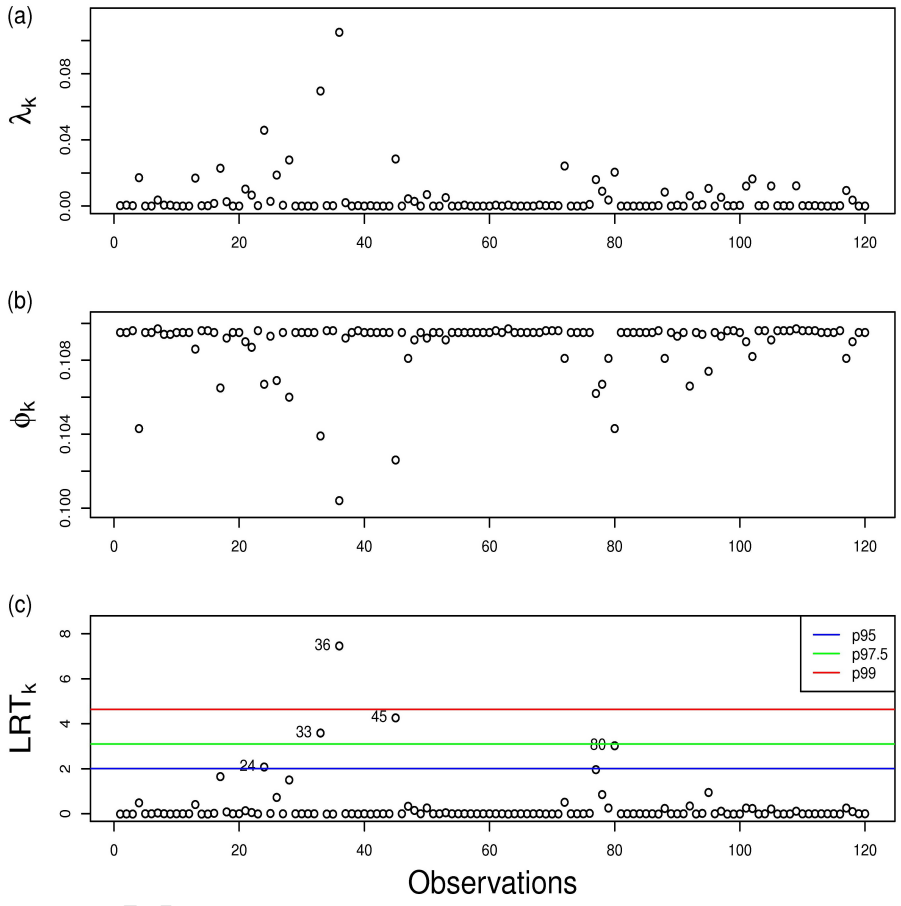


Figure 5.8: **VSOM statistics plotted against observation number for the leukemia rats dataset.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, ϕ_k. (c) Likelihood ratio statistics, LRT_k, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values.

by

$$\log(\mathrm{E}(Y_{ij}|\mathbf{b},\delta_k,\zeta_i)) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu_b} + \nu_{\delta_{24}}\mathbf{d}_{24} + \nu_{\delta_{33}}\mathbf{d}_{33} + \nu_{\delta_{36}}\mathbf{d}_{36} + \nu_{\delta_{45}}\mathbf{d}_{45} + \nu_{\delta_{80}}\mathbf{d}_{80} + \nu_{\zeta_{28}}\mathbf{d}_{sub(28)},$$

where δ_k is the random effect for the k^{th} observation and $\mathbf{d}_k$ is a vector of length 120 with 1 in the k^{th} position and zeros in the other positions; ψ_i is the random effect for the i^{th} subject and $\mathbf{d}_{sub(i)}$ is a vector of length 120 with values of 1 in the positions corresponding to the i^{th} subject and zeros in the other positions.

The potentially outlying observations and subjects were deleted from the study and model M_{R4} was fitted to the data. It can be seen that the fixed effects estimates for model M_{R3} and M_{R4} are similar with the VSOM having the added advantage that the data is retained.

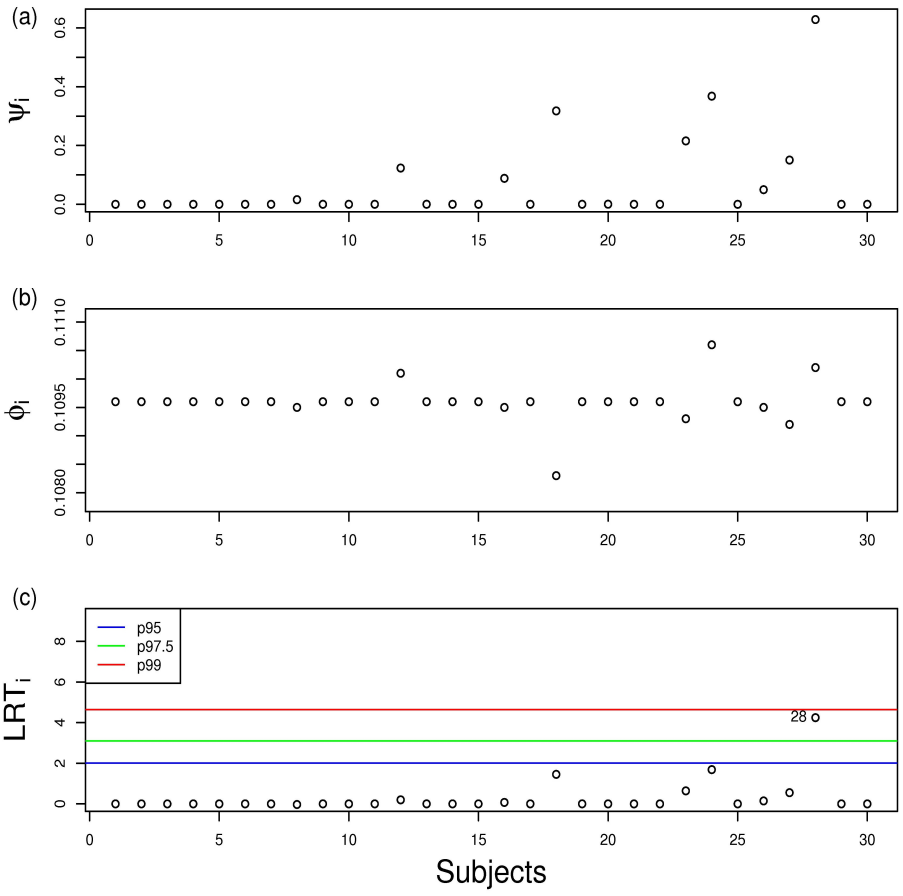


Figure 5.9: **VSOM statistics plotted against subject number for the leukemia rat dataset.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, ϕ_i. (c) Likelihood ratio statistics, LRT_i, with $95^{th}, 97.5^{th}$ and 99^{th} percentile cut-off values.

Chapter 6

VSOM for binomial data

In a biological research study the researcher maybe interested in a response which is either positive or negative, i.e. success or false. For instance in a medical trial a patient can either have a disease or not have it. Such data are referred to as binary data. In such data the response variable of interest R_i can take on values of 1 and 0 only, with $\Pr(R_i = 1) = \pi_i$, where values of 1 correspond to successes and π_i is the unknown probability of success.

Another form of data arises when the researcher is interested primarily in how a particular treatment affects a group of subjects with the same characteristics. For an example a researcher may be interested in modeling the proportion of insects killed by a particular dose of insecticide. Given that there are n groups with the size of each group given as n_i, there will be random variables associated with the observations in the group. These random variables will be $R_{i1}, \ldots, R_{in_i}$, where R_{ij} can take on values of 1 for a success and 0 for a failure, for $j = 1, \ldots, n_i$. The random variable of interest is $Y_i = \sum_{j=1}^{n_i} Rij$, the number of successes for the i^{th} group. In the context of an example in which the researcher is interested in the proportion of insects killed by a particular dose of insecticide, Y_i will be the number of insects killed out of a group of n_i insects at the i^{th} dose level. Such data are called grouped binary data or binomial data. This is a cross-sectional study as the groups are only being observed at a single time point and the groups are independent of each other.

The Bernoulli and binomial models are generally used to model binary

and binomial data respectively. These models are applied within the generalized linear modeling (GLM) framework. As mentioned in Chapter 5, the GLM framework has been discussed by many authors such as Wedderburn (1974), McCullagh and Nelder (1989) and Dobson and Barnett (2008). Linear logistic models are used in the analysis of these data. If the variance observed in the data is considerably greater than the variance reported by the model, the data can be said to be overdispersed. Collett (1996) page 201, state that neither overdispersion nor underdispersion arise in independent binary data.

In some situations the binary or binomial data are recorded repeatedly overtime. For binary data this would involve observing a particular subject at multiple occasions and recording whether a success or a failure has occurred at that occasion. Given that there are q subjects, the response variable would be R_{ij} for $i = 1, \ldots, q$ with each subject being observed n_i times for $j = 1, \ldots, n_i$, where $R_{ij} = 1$ for a success and 0 for a failure.

In the case of binomial data a group of subjects with similar characteristics are clustered together and observed at multiple occasions with the proportion of subjects with a successful response recorded at each occasion. Given that there are q groups of binary responses for $i = 1, \ldots, q$ with each group being observed n_i times for $j = 1, \ldots, n_i$ and the binary responses within each group at each time point is given as R_{ijk}, where $R_{ijk} = 1$ for a success and 0 for a failure for $k = 1, \ldots, n_{ij}$, where n_{ij} is the number of binary responses for a given group i at the j^{th} time. The number of successes for the i^{th} group at the j^{th} time is given by $Y_{ij} = \sum_{k=1}^{n_{ij}} R_{ijk}$ and the associated proportion of successes is $p_{ij} = Y_{ij}/n_{ij}$.

I will proceed in this chapter by outlining the binomial model. I will then proceed to describe the models which can be used to accommodate overdispersion in binomial data. The VSOM framework will then be applied using the beta-binomial HGLM. It will thus be shown that it is possible to link the VSOM for normally distributed data to that for binomial data. This will be done by considering the overdispersion associated with the i^{th} observation as a proxy for the variance shift for that observation. This approach will then be extended to longitudinal binomial data. Independent binary data cannot be overdispersed (Collett, 1996), thus I will not explore

overdispersion or the VSOM framework for this form of data.

6.1 Binomial regression models

Binomial regression models are used to fit independent binomial data with covariates. In this section I will firstly introduce a binomial GLM and then proceed to review the quasi-binomial and beta-binomial models which are used to model overdispersed binomial data.

6.1.1 Binomial GLM

Suppose the response variable Y_i is the number of successes out of n_i independent trials with unknown probability of success π_i. This implies that $Y_i \sim \text{Binomial}(n_i, \pi_i)$ with the probability mass function given by

$$Pr(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

The observed response variable is transformed into a proportion (p_i) when conducting linear logistic regression such that $p_i = Y_i/n_i$. The moments of p_i are $\text{E}(p_i) = \pi_i$ and $\text{var}(p_i) = \pi_i(1 - \pi_i)/n_i$.

The properties of Y_i are given by

$$\text{E}(Y_i) = \text{E}(n_i p_i) = n_i \pi_i,$$

and

$$\text{var}(Y_i) = \text{var}(n_i p_i) = n_i \pi_i (1 - \pi_i).$$

The natural link function for the binomial distribution is the logit function. The linear predictor is given by

$$\log\left(\frac{\text{E}(p_i)}{1 - \text{E}(p_i)}\right) = \mathbf{X}_i \boldsymbol{\beta}, \tag{6.1}$$

where $\mathbf{X}_i$ is the i^{th} row of the $n \times p$ design matrix of covariates and $\boldsymbol{\beta}$ is the p-dimensional vector of unknown coefficients. There may be cases where there is overdispersion in the data, such that $\text{var}(p_i) > \pi_i(1-\pi_i)/n_i$. In such situations two approaches can be considered as models for the data, and these approaches are the quasi-binomial and beta-binomial models (Collett, 1996).

6.1.2 Quasi-binomial model

Williams (1982) proposed the quasi-binomial model for independent binomial data with varying response probabilities. Suppose that the data consist of n proportions, $y_i/n_i, \forall i = 1, \ldots, n$. Also suppose the corresponding response probability for the i^{th} observation depends on its values for p explanatory variables $X_1, X_2, \ldots, X_p$ through a linear logistic model. The actual response probability, q_i, is assumed to vary about a mean of π_i in order to introduce variability in the response probabilities (Collett, 1996). Thus the response probability is a random variable, in the range (0,1), where $\mathrm{E}(q_i) = \pi_i$. The variance function of q_i is given as

$$\mathrm{var}(q_i) = \phi\pi_i(1 - \pi_i), \tag{6.2}$$

where $\phi \geq 0$ is the unknown scale parameter. q_i is an unobserved random variable, but given a particular value of q_i the random variable Y_i will have a binomial distribution with mean $n_i q_i$ and variance $n_i q_i(1 - q_i)$. The conditional properties of $Y_i|q_i$ are

$$\mathrm{E}(Y_i|q_i) = n_i q_i,$$

and

$$\mathrm{var}(Y_i|q_i) = n_i q_i(1 - q_i).$$

The marginal properties of Y_i are derived by using standard conditional probability theory, these results are given in detail by Collett (1996) on page 200. Briefly the marginal properties of Y_i are

$$\mathrm{E}(Y_i) = n_i\pi_i,$$

and

$$\mathrm{var}(Y_i) = n_i\pi_i(1 - \pi_i)\sigma_i^2,$$

where $\sigma_i^2 = 1 + (n_i - 1)\phi$. It can be seen that the term σ_i^2 accounts for the overdispersion in the data. The derivation of the marginal variance can be found on page 200 of Collett (1996). If $\phi = 0$ then $\sigma_i^2 = 1$, thus there will be no overdispersion and the model reverts to the binomial distribution with $\mathrm{var}(Y_i) = n_i\pi_i(1 - \pi_i)$.

6.1.3 Beta-binomial model

This model was proposed for use in modeling overdispersed binomial data by Williams (1975) and Crowder (1978). Consider that there are n binary groups, which implies that the data are independent binomial data, and that for each group $i = 1, \ldots, n$ with observations per group given as $j = 1, \ldots, n_i$. Also the number of successes per group is $Y_i = \sum_{j=1}^{n_i} R_{ij}$ where R_{ij} are binary responses, and the proportions of successes are Y_i/n_i. It is assumed that q_i, the actual response probability is allowed to vary around a mean of π_i. As a result q_i is a random variable. It is also assumed that q_i follows a beta distribution with parameters (α_i, θ_i), such that $\pi_i = \alpha_i/(\alpha_i + \theta_i)$. Conditional on a given value of q_i, Y_i has properties

$$Y_i|q_i \sim \text{Binomial}(n_i, q_i).$$

The density function for q_i is given by

$$q_i^{\alpha_i - 1}(1 - q_i)^{\theta_i - 1}/[B(\alpha_i, \theta_i)],$$

where $0 < q_i < 1$ and

$$B(\alpha_i, \theta_i) = \Gamma(\alpha_i + \theta_i)/[\Gamma(\alpha_i)\Gamma(\theta_i)],$$

with Γ being the gamma-function. The properties of q_i are given by

$$\text{E}(q_i) = \pi_i = \alpha_i/(\alpha_i + \theta_i),$$

and

$$\text{var}(q_i) = \rho_i \pi_i (1 - \pi_i),$$

where ρ_i is the correlation coefficient within each group which measures the correlation between the individual binary responses within each group. The inter-cluster correlation coefficient is given by $\rho_i = 1/(1 + \alpha_i + \theta_i)$ which can be reparameterized as $\rho_i = \gamma_i/(1 + \gamma_i)$, where $\gamma_i = 1/(\alpha_i + \theta_i)$. The marginal properties of Y_i are derived using iterated expectations. Thus the marginal properties are given by

$$\text{E}(Y_i) = \text{E}[\text{E}(Y_i|q_i)] = n_i \pi_i,$$

and

$$\begin{aligned} \text{var}(Y_i) &= \text{E}[\text{var}(Y_i|q_i)] + \text{var}[\text{E}(Y_i|q_i)] \\ &= n_i\pi_i(1-\pi_i) + \frac{\gamma_i}{(1+\gamma_i)} n_i(n_i-1)\pi_i(1-\pi_i). \end{aligned} \tag{6.3}$$

In fitting this model the parameters α_i and θ_i, and thus γ_i and ρ_i, are assumed to be constant $\forall i = 1, \ldots, n$ (Collett, 1996). This implies that $\alpha_i = \alpha$, $\theta_i = \theta$, $\gamma_i = \gamma$,$\rho_i = \rho$ and $\pi_i = \pi$. This simplifies the marginal variance of Y_i to

$$\text{var}(Y_i) = n_i\pi(1-\pi) + \frac{\gamma}{(1+\gamma)} n_i(n_i-1)\pi(1-\pi). \tag{6.4}$$

Thus the beta-binomial distribution accommodates overdispersion using the term $\frac{\gamma}{(1+\gamma)} n_i(n_i-1)\pi(1-\pi)$. This term is dependent on the size of the dispersion parameter of q_i, that is γ.

Guimarães (2005) used an existing Stata (StataCorp, 2009) command, i.e the **xtnbreg** command, for overdispersed count panel data in order to estimate the parameters for the beta-binomial distribution and its multivariate generalization the Dirichlet-multinomial distribution. This command also allows for the inclusion of covariates in the regression analysis. The beta-binomial model can also be fitted in the R statistical system (R Core Team, 2012) using the **aod** package (Lesnoff and Lancelot, 2012). Lee and Nelder (1996) fitted this model using saturated random effects, that is assuming that every observation is a random effect which follows a beta distribution. This is a similar approach to fitting a negative binomial model using a Poisson-gamma HGLM with saturated random effects. This approach can be used in the GENSTAT statistical system (Payne et al., 2011). Lee and Nelder (1996) proposed the use of the constraint $\alpha = \theta$, with the inverse of α being used as the unknown parameter. These constraints were imposed in order to improve convergence. This implies that

$$q_i \sim \text{Beta}(1/\alpha, 1/\alpha),$$

with the properties of q_i being $E(q_i) = 1/2$ and $\gamma = \alpha/2$.

6.2 Variance shift outlier model (VSOM) for binomial data

Once a binomial GLM is fitted to the data, a residual analysis is often used to identify outliers and/or influential observations in a dataset. Often outliers can be corrected or removed from the data and the analysis redone. However, in most cases they are anomalous and we may want to include them in the analysis. If the data are overdispersed we could use the beta-binomial or quasi-binomial models. In this section I introduce a VSOM for both independent and longitudinal binomial data. This model can be viewed as a model for overdispersion associated with the i^{th} observation, with observations having large overdispersion considered as potential outliers. I will fit these models using the HGLM approach of Lee and Nelder (1996).

6.2.1 A VSOM for independent binomial data

The models outlined in sections 6.1.2 and 6.1.3 were intended to handle overdispersion in the data due to variation between the response probabilities. In this section a model is introduced which handles overdispersion due to specific observations, that is outliers.

6.2.1.1 A VSOM with no covariates

Considering a model with no covariates, a VSOM for the i^{th} proportion is given by

$$\log\left(\frac{\mathrm{E}(p_i|\delta_i)}{1-\mathrm{E}(p_i|\delta_i)}\right) = \nu_{\delta_i}, \tag{6.5}$$

where δ_i is a random effect for the i^{th} proportion. Following Lee and Nelder (1996), it is assumed that $\nu_{\delta_i} = \log\left(\frac{\delta_i}{1-\delta_i}\right)$, where δ_i is a random effect which follows a beta distribution with parameters $(1/\alpha_i, 1/\alpha_i)$. This makes model (6.5) a beta-binomial HGLM (Lee and Nelder, 1996). This model can also be given by

$$\mathrm{E}(p_i|\delta_i) = \frac{\exp(\nu_{\delta_i})}{1+\exp(\nu_{\delta_i})}. \tag{6.6}$$

Since $\nu_{\delta_i} = \log\left(\frac{\delta_i}{1-\delta_i}\right)$ (6.6) can be simplified as follows

$$\begin{aligned}
\mathrm{E}(p_i|\delta_i) &= \left(\frac{\delta_i}{1-\delta_i}\right) / \{1 + \left(\frac{\delta_i}{1-\delta_i}\right)\} \\
&= \left(\frac{\delta_i}{1-\delta_i}\right) / \left(\frac{1}{1-\delta_i}\right) \\
&= \delta_i.
\end{aligned}$$

Since $Y_i = n_i p_i$,

$$\mathrm{E}(Y_i|\delta_i) = n_i \mathrm{E}(p_i|\delta_i) = n_i\delta_i. \tag{6.7}$$

The conditional variance of $Y_i|\delta_i$ is given by

$$\mathrm{var}(Y_i|\delta_i) = n_i\delta_i(1-\delta_i). \tag{6.8}$$

The properties of δ_i are

$$\mathrm{E}(\delta_i) = \pi_i = 1/2,$$
$$\mathrm{var}(\delta_i) = \rho_i\pi_i(1-\pi_i) = \rho_i/4,$$

where $\rho_i = \alpha_i/(\alpha_i + 2)$; ρ_i can be reparameterized as $\rho_i = \lambda_i/(1+\lambda_i)$, where $\lambda_i = \alpha_i/2$, λ_i will be considered as the dispersion parameter for δ_i. The marginal properties of Y_i are derived using iterated expectations. The marginal mean Y_i is given by

$$\begin{aligned}
\mathrm{E}(Y_i) &= \mathrm{E}[\mathrm{E}(Y_i|\delta_i)] \\
&= n_i\mathrm{E}(\delta_i) \\
&= n_i\pi_i,
\end{aligned}$$

the constraints imposed by Lee and Nelder (1996) simplify the marginal mean to $n_i/2$. The marginal variance of Y_i is given by

$$\begin{aligned}
\mathrm{var}(Y_i) &= \mathrm{E}[\mathrm{var}(Y_i|\delta_i)] + \mathrm{var}[\mathrm{E}(Y_i|\delta_i)] \\
&= \mathrm{E}[n_i\delta_i(1-\delta_i)] + \mathrm{var}(n_i\delta_i) \\
&= n_i[\mathrm{E}(\delta_i) - \{\mathrm{var}(\delta_i) + [\mathrm{E}(\delta_i)]^2\}] + n_i^2\mathrm{var}(\delta_i) \\
&= n_i[\pi_i - \frac{\lambda_i}{(1+\lambda_i)}\pi_i(1-\pi_i) - \pi_i^2] + n_i^2\frac{\lambda_i}{(1+\lambda_i)}\pi_i(1-\pi_i) \\
&= n_i\pi_i(1-\pi_i) + \frac{\lambda_i}{(1+\lambda_i)}n_i(n_i-1)\pi_i(1-\pi_i)
\end{aligned} \tag{6.9}$$

It can be seen from equation (6.9) that the overdispersion associated with the i^{th} observation is accommodated by the term $\frac{\lambda_i}{2(1+\lambda_i)} n_i(n_i-1)\pi_i(1-\pi_i)$. The difference between (6.4) and (6.9) is that the dispersion parameter in (6.4), γ, accommodates the overdispersion for all the observations in the dataset while λ_i accounts for the overdispersion due the i^{th} observation only. As the size of λ_i approaches zero the marginal variance of Y_i approaches the variance associated with a binomial response. Large values of λ_i are indicative of outlying observations.

6.2.1.2 A VSOM with covariates

Considering model (6.1) as the null model, a model in which the data are not overdispersed, the linear predictor was shown to be

$$\log\left(\frac{\mathrm{E}(p_i)}{1-\mathrm{E}(p_i)}\right) = \mathbf{x}_i\boldsymbol{\beta}.$$

The linear predictor can be written as

$$\mathrm{E}(p_i) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i\boldsymbol{\beta})} = \pi_i.$$

Since $Y_i = n_i p_i$

$$\mathrm{E}(Y_i) = \frac{n_i\exp(\mathbf{x}_i\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i\boldsymbol{\beta})} = n_i\pi_i.$$

The variance is given by

$$\mathrm{var}(Y_i) = n_i\pi_i(1-\pi_i).$$

A VSOM for the i^{th} proportion is given by

$$\log\left(\frac{\mathrm{E}(p_i|\delta_i)}{1-\mathrm{E}(p_i|\delta_i)}\right) = \mathbf{x}_i\boldsymbol{\beta} + \nu_{\delta_i}, \tag{6.10}$$

where δ_i is a random effect for the i^{th} proportion. Following Lee and Nelder (1996), it is assumed that $\nu_{\delta_i} = \log\left(\frac{\delta_i}{1-\delta_i}\right)$, where δ_i is a random effect which follows a beta distribution with parameters $(1/\alpha_i, 1/\alpha_i)$. The properties of δ_i are

$$\mathrm{E}(\delta_i) = \pi_i = 1/2,$$

$$\mathrm{var}(\delta_i) = \rho_i\pi_i(1-\pi_i) = \rho_i/4,$$

where $\rho_i = \alpha_i/(\alpha_i+2)$; ρ_i can be reparameterized as $\rho_i = \lambda_i/(1+\lambda_i)$, where $\lambda_i = \alpha_i/2$, λ_i will be considered as the dispersion parameter of δ_i.

Model (6.10) can be written as

$$\mathrm{E}(p_i|\delta_i) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta} + \nu_{\delta_i})}{1+\exp(\mathbf{x}_i\boldsymbol{\beta} + \nu_{\delta_i})},$$

thus

$$\mathrm{E}(Y_i|\delta_i) = \frac{n_i\exp(\mathbf{x}_i\boldsymbol{\beta} + \nu_{\delta_i})}{1+\exp(\mathbf{x}_i\boldsymbol{\beta} + \nu_{\delta_i})}.$$

Since $\nu_{\delta_i} = \log\left(\frac{\delta_i}{1-\delta_i}\right)$

$$\mathrm{E}(Y_i|\delta_i) = \frac{n_i\delta_i\exp(\mathbf{x}_i\boldsymbol{\beta})}{1+\delta_i[1-\exp(\mathbf{x}_i\boldsymbol{\beta})]}.$$

It is not possible to get a closed form expression for the marginal mean and the marginal variance for this model. The estimation of the unknown parameters for this model can be done by joint maximization of the h-likelihood for the fixed and random effects as described by Lee and Nelder (1996). It will be shown in section 6.3 that the VSOM is able to identify and downweight outlying observations. These outlying observations will be shown to have large values of λ_i. The outlying observations identified by the VSOM will be compared to the observations identified by existing model diagnostic techniques.

6.2.2 A VSOM for longitudinal binomial data

The null model for longitudinal binomial data is given by the HGLM described in chapter 2.3. It takes the form

$$\eta_{ij} = \log\left(\frac{\mathrm{E}(p_{ij}|b_i)}{1-\mathrm{E}(p_{ij}|b_i)}\right) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i}, \tag{6.11}$$

for $i = 1,\ldots,q$ and for $j = 1,\ldots,n_i$ with q being the number of groups in the study, where $\mathbf{x}_{ij}^{'}$ is the j^{th} rows of the design matrix $\mathbf{X}_i$. Where $\mathbf{X}_i$ is a $n_i \times p$ design matrix of covariates, with $\boldsymbol{\beta}$ being a p-dimensional vector of fixed effect coefficients, b_i is the group random effect with $\nu_{b_i} = \log\left(\frac{b_i}{1-b_i}\right)$. The random effects b_i are assumed to follow a beta distribution with dispersion parameter γ.

A VSOM for the j^{th} observation of the i^{th} group is given by

$$\eta_{ij} = \log\left(\frac{\mathrm{E}(p_{ij}|b_i, \delta_{ij})}{1 - \mathrm{E}(p_{ij}|b_i, \delta_{ij})}\right) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i} + \nu_{\delta_{ij}}, \tag{6.12}$$

where $\nu_{\delta_{ij}} = \log\left(\frac{\delta_{ij}}{1-\delta_{ij}}\right)$, with δ_{ij} being a random effect which follows a beta distribution with dispersion parameter λ_{ij}. It is also assumed that all random effects are independent. The constraints on the random effects are $b_i \sim \mathrm{Beta}(1/\alpha_{\nu_{b_i}}, 1/\alpha_{\nu_{b_i}})$ and $\delta_{ij} \sim \mathrm{Beta}(1/\alpha_{\delta_{ij}}, 1/\alpha_{\delta_{ij}})$, where λ_{ij} is the dispersion parameter for the δ_{ij} random effect. Once again the marginal variance of this model cannot be found in a closed form. The unknown parameter estimates can be found using joint maximization of the h-likelihood described by Lee and Nelder (1996). It will be shown in section 6.3 that observations with relatively large values of λ_{ij} are indicative of outliers. When presenting the results of the VSOM for individual observations in clustered data, I will consider the j^{th} observation of the i^{th} group as being the k^{th} unit of the n-dimensional vector of responses $\mathbf{Y}$. Thus I will present my results in terms of the k^{th} unit. This implies that units with relatively large values of λ_k are indicative of outliers.

6.2.2.1 A VSOM for outlying subjects in longitudinal binomial data

A VSOM for all observations of the i^{th} group is given by

$$\eta_{ij} = \log\left(\frac{\mathrm{E}(p_{ij}|b_i, \delta_{ij})}{1 - \mathrm{E}(p_{ij}|b_i, \delta_{ij})}\right) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \nu_{b_i} + \nu_{\zeta_i}, \tag{6.13}$$

for $i = 1, \ldots, q$ and for $j = 1, \ldots, n_i$ with q being the number of groups in the study, where $\mathbf{x}_{ij}^{'}$ is the j^{th} row of the design matrix $\mathbf{X}_i$ with $\mathbf{X}_i$ being a $n_i \times p$ design matrix of covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effect coefficients, b_i is the group random effect with $\nu_{b_i} = \log\left(\frac{b_i}{1-b_i}\right)$ with b_i being a random effect which follows a beta distribution with dispersion parameter γ, $\nu_{\zeta_i} = \log\left(\frac{\zeta_i}{1-\zeta_i}\right)$, with ζ_i being a random effect which follows a beta distribution with dispersion parameter ψ_i. It is also assumed that all random effects are independent. The constraints on the random effects are $b_i \sim \mathrm{Beta}(1/\alpha_{b_i}, 1/\alpha_{b_i})$ and $\zeta_i \sim \mathrm{Beta}(1/\alpha_{\zeta_i}, 1/\alpha_{\zeta_i})$, where ψ_i is the dispersion parameter for the ζ_i random effect, for $i = 1, ..., q$. Groups with

relatively large values of ψ_i are indicative of outliers. The difference between the VSOM for outlying individual observations and the VSOM for outlying groups in longitudinal binomial data is that the marginal variance for individual observations depends on λ_{ij} which accounts for the overdispersion due to a specific j^{th} observation belonging to the i^{th} group; while the marginal variance for groups depends on ψ_i which accounts for the overdispersion of all the observations being to the i^{th} group.

6.3 Examples

6.3.1 Seeds germination dataset

The germinating seeds dataset (Crowder, 1978) was used as an example of the application of the VSOM to independent binomial data. In order to select an appropriate null model for the data a binomial GLM (M_0), quasi-binomial GLM (M_1) and beta-binomial HGLM (M_2) were fitted to the data. The results of the model fitting process are shown in Table 6.1. From this table it can be seen that the beta-binomial provided the best fit to the data, with a deviance of 184.40 relative to the deviances of 188.96 and 185.31 for the binomial GLM and quasi-binomial GLM respectively. The beta-binomial model was thus used to fit the null model for this example. This model was fitted in GENSTAT by fitting a binomial model with saturated random effects which follow a beta distribution, this is similar to the formulation of the negative binomial model in chapter 5.1.2.

The null model fitted had the linear predictor

$$\eta_i = \log\left(\frac{\mathrm{E}(p_i|s)}{1-\mathrm{E}(p_i|s)}\right) = \beta_0 + S_i\beta_1 + E_i\beta_2 + (S_i \times E_i)\beta_3 + \nu_s, \qquad (6.14)$$

where S_i is the species of the i^{th} observation with values of 0 for species Orobanche aegyptiaca 75 (o75) and 1 for species Orobanche aegyptiaca 73 (o73), E_i is the root extract for the i^{th} observation with values of 0 for the bean extract and 1 for the cucumber extract, s is the observation random effect with dispersion parameter γ.

The VSOM was applied to all the observations in turn and the results

Table 6.1: **Parameter estimates of models fitted to the seeds germination dataset.**

Parameter	M_0	M_1	M_2
constant	-0.558 (0.126)	-0.558 (0.176)	-0.543 (0.187)
species o73	0.146 (0.223)	0.146 (0.312)	0.080 (0.303)
extract cucumber	1.318 (0.177)	1.318 (0.248)	1.337 (0.265)
species o73 $\times$ extract cucumber	-0.778 (0.306)	-0.778 (0.428)	-0.822 (0.423)
ϕ		1.958 (0.672)	
γ			0.023 (0.012)
deviance	188.96	185.31	184.40

where γ is the dispersion parameter due to the observations random effect.

are shown in Figure 6.1. The VSOM for the i^{th} observation takes the form

$$\eta_i = \log\left(\frac{\mathrm{E}(p_i|s,\delta_i)}{1-\mathrm{E}(p_i|s,\delta_i)}\right) = \beta_0 + S_i\beta_1 + E_i\beta_2 + (S_i \times E_i)\beta_3 + \nu_s + \nu_{\delta_i}, \quad (6.15)$$

where δ_i is the random effect for the i^{th} observation, which follows a beta distribution. It can be seen from Figure 6.1 that there were no outlying observations. This is supported by a plot of absolute residuals against fitted values shown in Figure 6.2. This plot shows that all absolute residuals were less than 2 thus none of the observations were clearly outlying, though observation 16 did have a high absolute residual.

In order to illustrate how the VSOM works for independent binomial data two outliers were inserted at observations 3 and 6. This was done by making the proportions of successes, for observation 3 and 6, 1 and 0 respectively.

The null model M_{S0} was then fitted to this updated dataset and the results are shown in Table 6.2. The VSOM was then applied and these observations were identified as outliers as shown in Figure 6.3. Model M_{S1} was fitted with observations 3 and 6 as random effects. The results are shown in Table 6.2.

There was a relatively large change in the fixed effect estimates in model M_{S1} as compared to model M_{S0}, thus the outlying observations were poten-

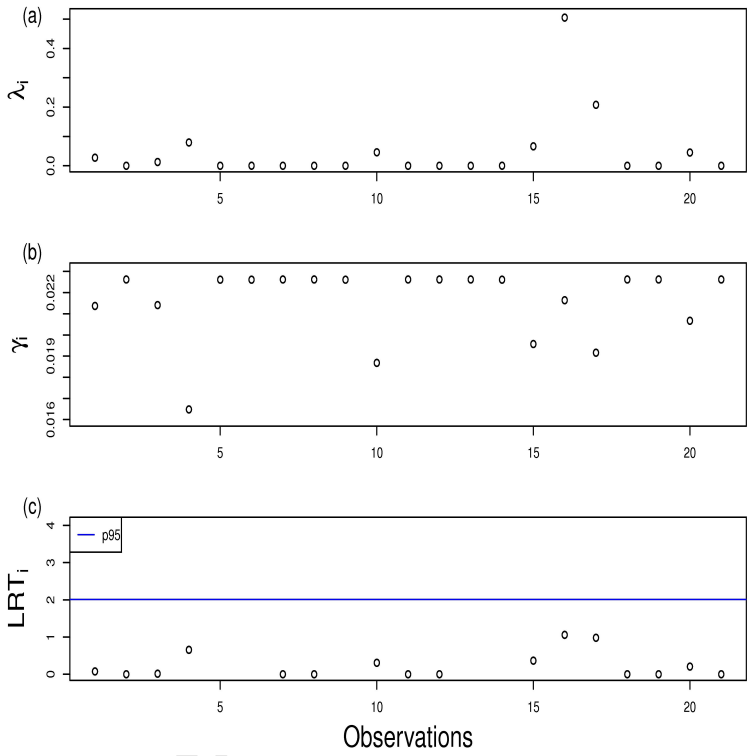


Figure 6.1: **VSOM statistics plotted against observation number for the seeds dataset.** (a) Variance shift estimates, λ_i. (b) Dispersion parameter γ_i. (c) Likelihood ratio statistics, LRT_i, with a 95^{th} percentile cut-off value.

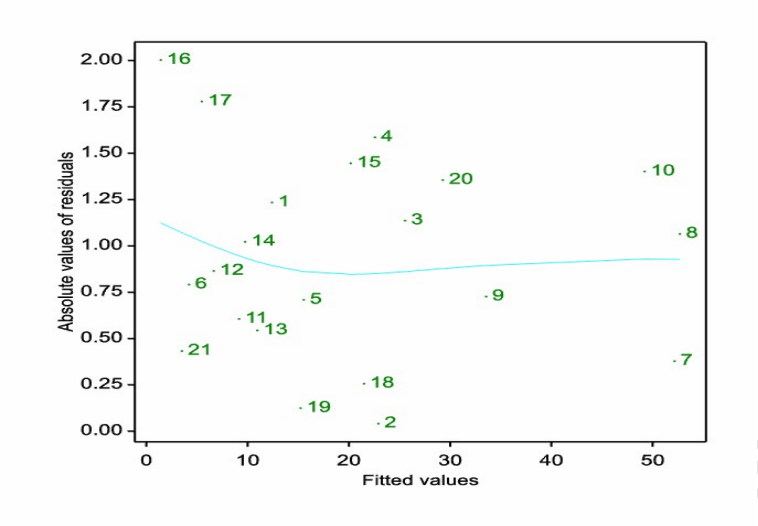


Figure 6.2: **Plot of the absolute residuals against fitted values for model M_{S0} from the seeds dataset.**

Table 6.2: **Parameter estimates of models fitted to the adjusted seeds germination dataset.**

Parameter	M_{S0}	M_{S1}	M_{S2}
constant	0.098 (0.523)	-0.419 (0.209)	-0.426 (0.210)
species o73	-0.715 (0.734)	-0.042 (0.315)	-0.035 (0.316)
extract cucumber	0.398 (0.695)	1.173 (0.281)	1.191 (0.282)
species o73 $\times$ extract cucumber	0.138 (0.995)	-0.658 (0.423)	-0.675 (0.431)
γ	0.240 (0.083)	0.021(0.012)	0.021 (0.012)
λ_3		3.128(3.410)	
λ_6		1.610 (1.884)	
deviance	208.80	181.89	168.53

where γ is the dispersion parameter due to the observations random effect and λ_i is the dispersion parameter for the i^{th} observation.

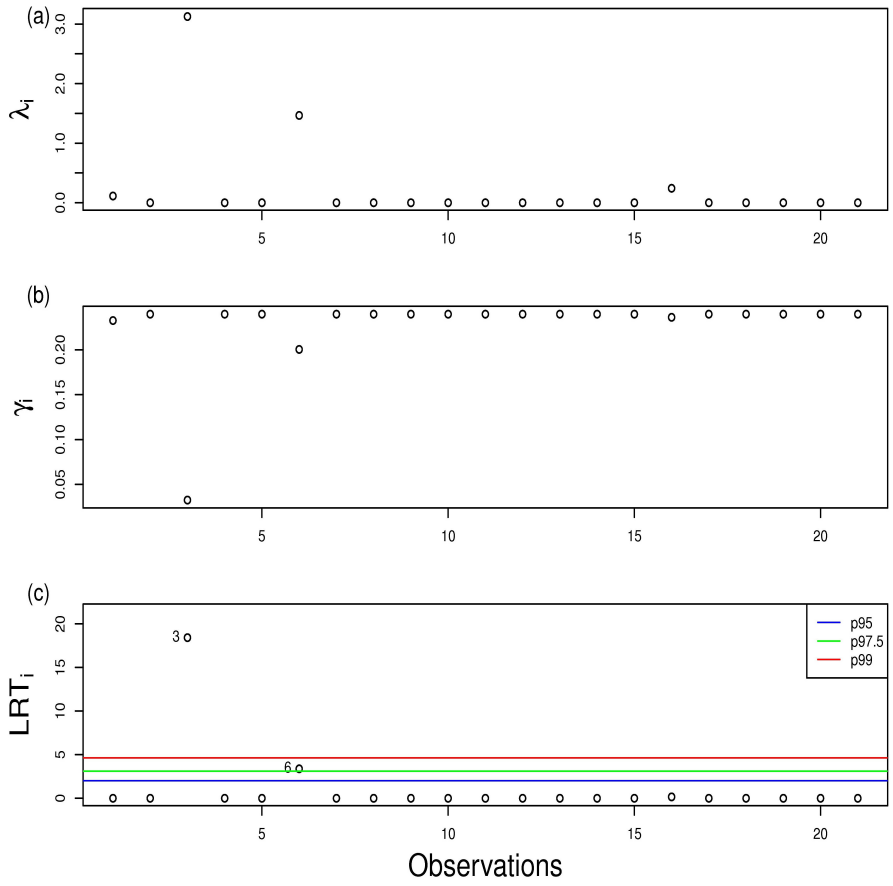


Figure 6.3: **VSOM statistics plotted against observation number for the adjusted seeds dataset.** (a) Variance shift estimates, λ_i. (b) Dispersion parameter γ_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values.

tially influential. This was expected due to the extremities of the outliers input into the dataset. Model M_{S2} was fitted after observations 3 and 6 were deleted from the dataset. The fixed effects for models M_{S1} and M_{S2} were similar thus showing that large outliers can be down-weighted to such an extent that they are effectively deleted from the study.

6.3.2 Contagious bovine pleuropneumonia dataset

The contagious bovine pleuropneumonia dataset (Lesnoff et al., 2004) was used as an example of the application of the VSOM to longitudinal binomial

data. The null model fit to the data was M_{H0} with linear predictor

$$\eta_{ij} = \log\left(\frac{\mathrm{E}(p_{ij}|b)}{1-\mathrm{E}(p_{ij}|b)}\right) = \beta_0 + \beta_1 Period2_{ij} + \beta_2 Period3_{ij} + \beta_3 Period4_{ij} + \nu_{b_i}, \tag{6.16}$$

where p_{ij} is the proportion of the i^{th} herd at the j^{th} time which is infected with CBPP for $i = 1, \ldots, 15$ with each herd being observed for at most $j = 4$ times, $Period2_{ij}$ is the effect of time period 2 relative to time period 1 for the i^{th} herd at the j^{th} time with $Period2_{ij}$ given a value of 1 if the i^{th} herd is observed at time period 2 and it is coded as 0 otherwise, $Period3_{ij}$ is the effect of time period 3 relative to time period 1 for the i^{th} herd at the j^{th} time with $Period2_{ij}$ given a value of 1 if the i^{th} herd is observed at time period 3 and it is coded as 0 otherwise, $Period4_{ij}$ is the effect of time period 4 relative to time period 1 for the i^{th} herd at the j^{th} time with $Period4_{ij}$ given a value of 1 if the i^{th} herd is observed at time period 4 and it is coded as 0 otherwise, b_i is the subject random effect with $\nu_{b_i} = \log\left(\frac{b_i}{1-b_i}\right)$ assuming b_i follows a beta distribution with a dispersion parameter of γ for $i = 1, \ldots, 15$.

The VSOM was applied to all the observations in turn and the results are shown in Figure 6.4. The VSOM for the j^{th} observation of the i^{th} herd takes the form

$$\begin{aligned}\eta_{ij} = \log\left(\frac{\mathrm{E}(p_{ij}|b,\delta_{ij})}{1-\mathrm{E}(p_{ij}|b)}\right) =& \beta_0 + \beta_1 Period2_{ij} + \beta_2 Period3_{ij} + \beta_3 Period4_{ij} \\ &+ \nu_{b_i} + \nu_{\delta_{ij}},\end{aligned} \tag{6.17}$$

δ_{ij} is the random effect specific to the j^{th} observation of the i^{th} herd with $\nu_{\delta_{ij}} = \log\left(\frac{\delta_{ij}}{1-\delta_{ij}}\right)$ assuming δ_{ij} follows a beta distribution with a dispersion parameter of λ_{ij}. As previously stated, I will consider the j^{th} observation of the i^{th} subject as being the k^{th} unit of the n-dimensional vector of counts $\mathbf{Y}$. Thus I will present my results in terms of the k^{th} observation. Using that approach observations 3, 37, 38 and 49 were found to be potential outliers. Model M_{H1} was fitted with these identified observations treated as random effects. The results are shown in Table 6.3. It can be seen from this model that there is a small change in the fixed effect estimates, compared to model M_{H0}, thus the outlying observations were not influential.

When the VSOM was applied to each herd in turn it was found that there were no outlying subjects as shown in Figure 6.5. Model M_{H2} was fit after deleting observations 3, 37, 38 and 49 from the dataset. The fixed effect estimates in model M_{H2} are similar to those obtained from model M_{H1}, thus the VSOM had the same effect as deleting the outlying observations.

Table 6.3: **Parameter estimates of models fitted to the CBPP dataset.**

Parameter	M_{H0}	M_{H1}	M_{H2}
constant	-1.364 (0.227)	-1.324 (0.220)	-1.357 (0.222)
period 2	-0.978 (0.303)	-1.153 (0.322)	-1.246 (0.347)
period 3	-1.113 (0.323)	-1.476 (0.367)	-1.496 (0.376)
period 4	-1.562 (0.424)	-1.706 (0.434)	-1.666 (0.435)
γ	0.095 (0.043)	0.068 (0.034)	0.067 (0.034)
λ_3		0.791 (1.004)	
λ_{37}		1.49 (1.758)	
λ_{38}		0.053 (0.135)	
λ_{49}		0.447 (0.617)	
deviance	328.80	309.99	275.81

where γ is the dispersion parameter for the herd random effect and λ_k is the dispersion parameter for the k^{th} observation.

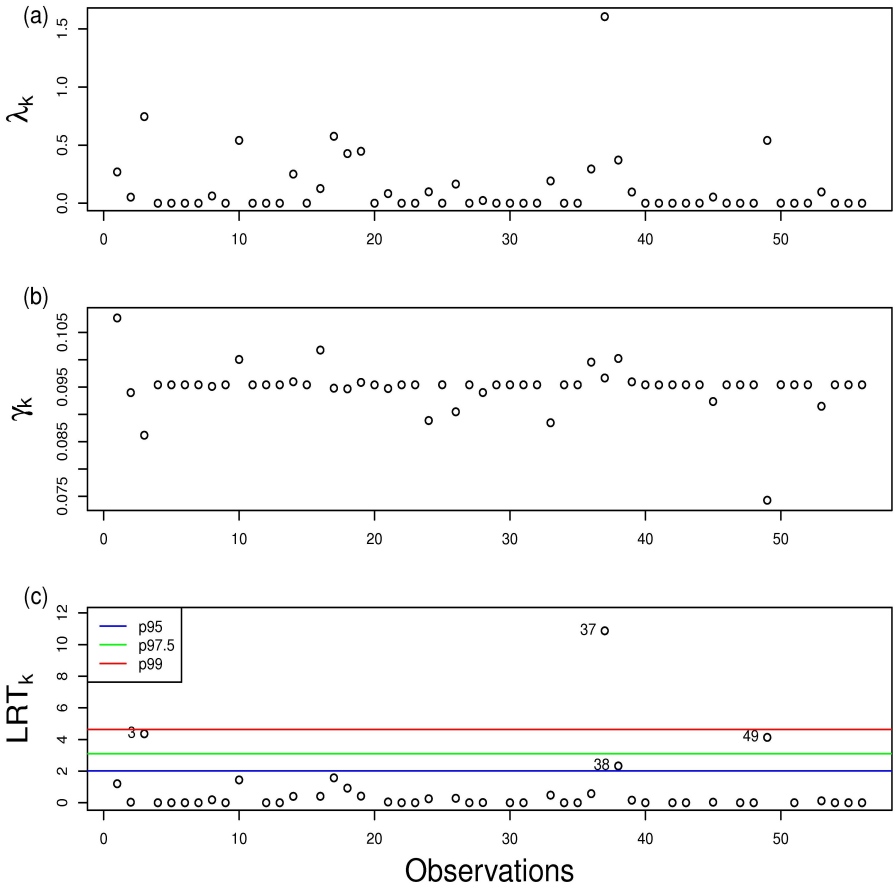


Figure 6.4: **VSOM statistics plotted against observation number for the CBPP dataset.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter for the herd random effect γ_k. (c) Likelihood ratio statistics, LRT_k, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values.

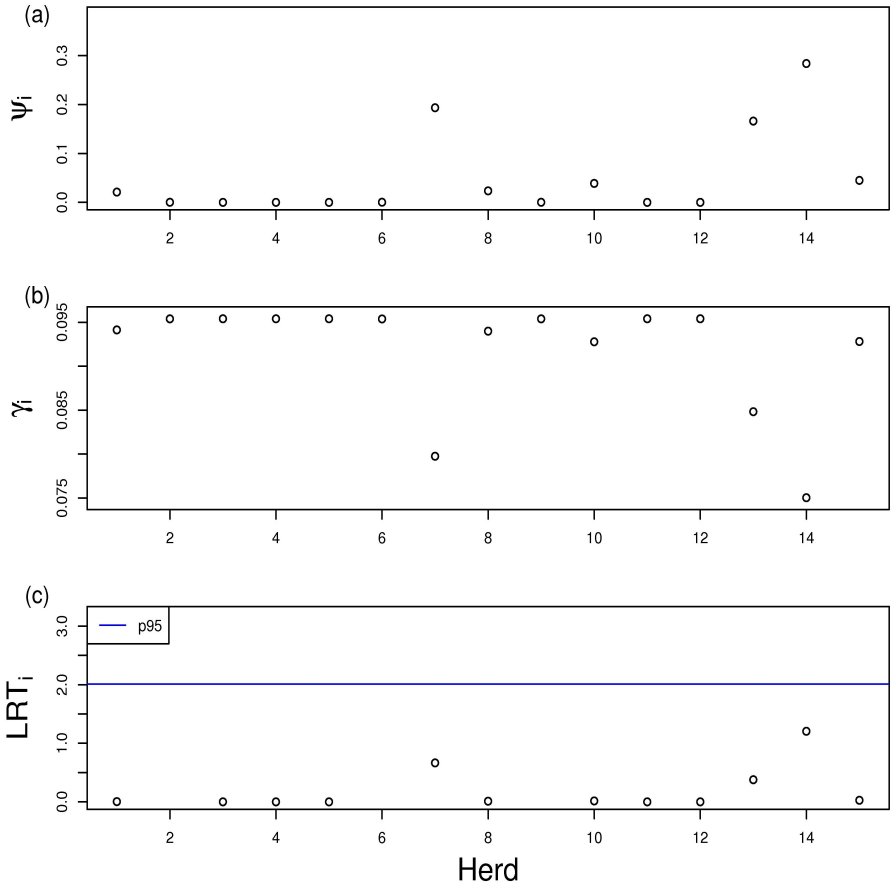


Figure 6.5: **VSOM statistics plotted against herd number for the CBPP dataset.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter for the herd random effect γ_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values.

Additional model checking was done using a plot of absolute residuals against fitted values (Figure 6.6). This plot shows that observations 3, 10, 17, 18, 37, 38 and 49 are potentially outlying. The VSOM was able to identify that all these observations had higher variance inflation parameters than the rest of the data, with observations 3, 37, 38 and 49 having variance inflation parameters which are significantly higher than the rest of the data. As a result it can be concluded that the VSOM gives consistent results with standard residual analysis.

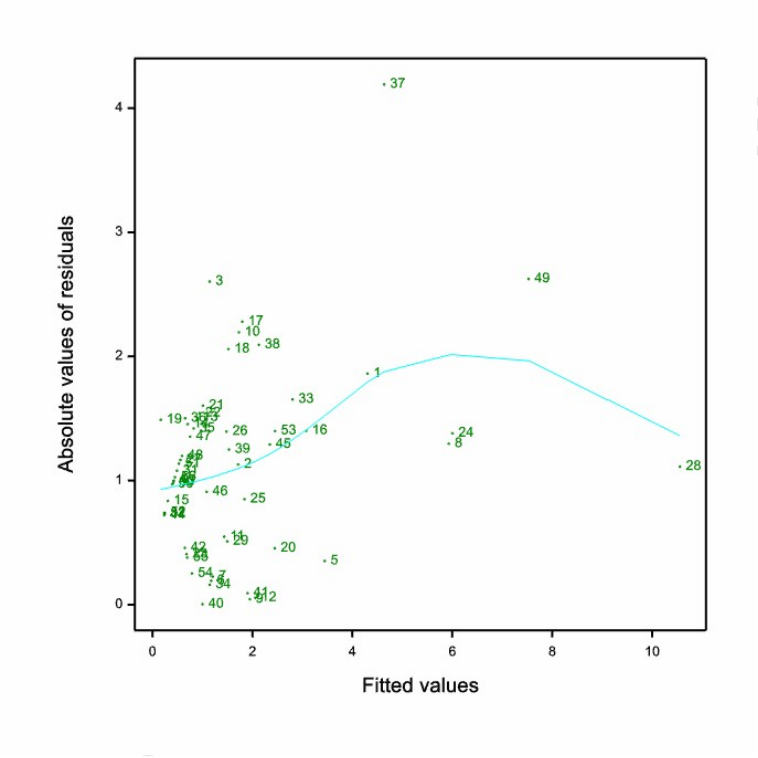


Figure 6.6: **Plot of absolute residuals against fitted values for individual observations using model M_{H0} in the CBPP dataset.**

Chapter 7

Cytokine dataset

In this section the VSOM was applied to a single dataset where the responses were analyzed as being normally distributed responses and counts. It has been shown in the previous sections that the VSOM can be applied to the these aforementioned types of responses in order to detect and down-weight outlying observations. The dataset used is from a study by Mansoor et al. (2009) which aimed to perform a comprehensive risk benefit analysis on the use of BCG vaccine. This dataset has a key feature of being skew with a long tail to the right. Given data with such a feature, researchers can choose to transform the data and thus analyze it as being normally distributed using a linear mixed model. Another alternative is to analyze the data in its original state using generalized linear mixed models or hierarchical generalized linear models assuming underlying distributions for skew data. The VSOM will be compared with existing model diagnostic techniques when the data are treated as being normally distributed or counts.

In this dataset the benefit of the BCG vaccine was assessed by whether or not it produced an immune response that is widely believed to indicate the quality of the immunity of an individual to TB. A secondary aim outlined in the paper was the investigation into whether the immune strength caused by BCG vaccination was the same in HIV-uninfected infants from HIV-infected and HIV-uninfected parents respectively.

It is widely believed that the presence of the T helper type 1 (Th1) response cytokine, which is characterized by, interferon (IFN)-γ, tumor necro-

sis factor (TNF)-α and interleukin (IL)-2 production, is vital to having a strong immunity to TB. It has been found that BCG induces CD4 and CD8 T cell populations that consist of combinations of IFN-γ, TNF-α and IL-2 (Soares et al., 2008). CD4 T cells that consist of more than one cytokine marker are referred to as polyfunctional, and this is considered as a very good indicator of the quality of immunity to TB. The levels of these polyfunctional CD4 T cell markers were thus treated as the response variable in the study.

The infants who were recruited in the study were infants born to HIV infected and HIV uninfected parents, in the Worcester region of the Western Cape in South Africa between 2003 and 2006. All infants recruited had to have received a BCG vaccination at the date of birth and they also needed to have had an HIV test at 6 weeks of age. There were 63 infants who were initially recruited into the study. However, there were 9 observations which were clearly outlying as they gave cytokines counts of 0 which is not possible. As a result these observations were deleted from the dataset. The resultant dataset which was analyzed had a sample size of 182 observations from 61 infants. The infants were assigned to three groups which were the HIV infected infants (Group 1), the HIV exposed and uninfected infants (Group 2) and the HIV unexposed and uninfected infants (Group 3). The term HIV exposed means that the parents of the infant are HIV positive. Throughout the duration of the study antiretroviral therapy was not available to HIV infected infants, thus any immune strength can only be considered to be due to the effect of BCG.

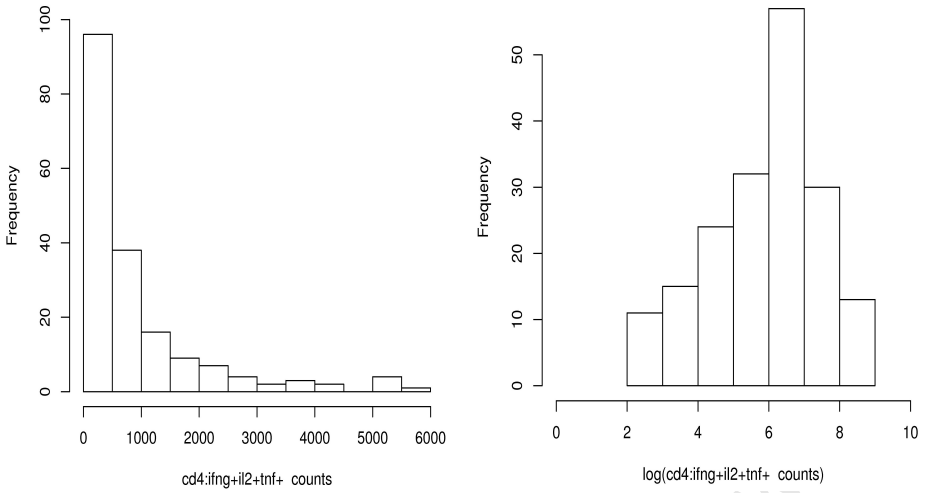
Infants were seen and blood samples were taken at 3, 6, 9 and 12 months. The information on how the blood samples were analyzed to get the T-cell marker measurements is outlined in Mansoor et al. (2009) . The measurements acquired were in the form of absolute count (unadjusted number of cells which have a particular combination of IFN-γ, TNF-α and IL-2 cytokines) and frequency data. The frequency data is derived from absolute count data and it is given as the proportion of absolute counts to the total number of cells in a patient, where the frequency value and absolute count values are for a particular combination of cytokines.

The data can be combined into the cells with 1, 2 or 3 cytokines present to

indicate a measure of polyfunctionality. In this thesis I will only consider the number of cells with all three cytokines as the response (cd4:inf+tnf+il2+). In this chapter it will be shown that the data can be analyzed as normally distributed or count data depending on the transformations applied to it. The VSOM will then be applied to each of these forms of data.

7.1 Data exploration

From Figure 7.1a it can be seen that the distribution of the response is skew to the right with most cd4:inf+tnf+il2+ counts being less than 1000 and some as large as 6000. As a result of the skewness the mean will be affected by the few large values. This is shown by the mean being considerably larger than the median, as indicated by the sample mean of 892.34 compared to the sample median of 485.5. The logarithmic transformation of the response reduces the size of the large values. This makes them lie nearer to the mean of the data values, thus resulting in a symmetrical distribution that appears to follow the normal distribution as shown in Figure 7.1b. This transformation results in a sample mean of 5.89 compared to a sample median of 6.19 for the transformed response, the relative similarity of these statistics indicates that the distribution is symmetrical.



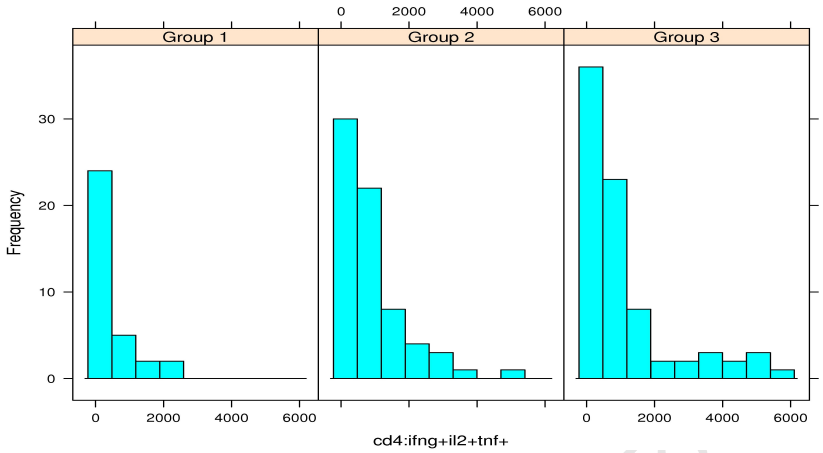
(a) Histogram of cd4:inf+tnf+il2+ (b) Histogram of log(cd4:inf+tnf+il2+)

Figure 7.1: **Histograms of the cd4:inf+tnf+il2+ counts and the logarithm of the cd4:inf+tnf+il2+ counts.**

Figure 7.2 shows that the cd4:inf+tnf+il2+ counts are skew across all the groups, with Group 3 having the longest tail. This implies that Group 3 has the largest extreme counts. It appears, from the graphical representation, that the levels of counts are not the same across all the groups. This is supported by the boxplots shown in Figure 7.3 which show that the levels of cd4:inf+tnf+il2+ are different for the three groups with the medians at different levels, though the interquartile ranges overlap. Group 3 also has the largest interquartile range, thus implying that there is more variability in this group as compared to the other groups, with group 1 having the smallest interquartile range. The boxplots also show that there are several outlying observations in each group which will need to be investigated further.

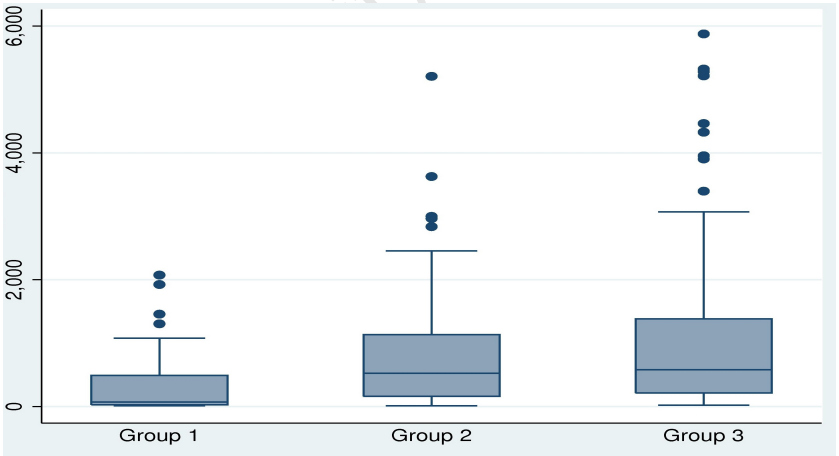
A non-parametric method of testing whether the median cd4:inf+tnf+il2+ counts were the same for all treatment groups is the Kruskal-Wallis test (Kruskal and Wallis, 1952). The the p-value calculated by applying this test was 0.0003, implying that there is a significant difference in the levels of cd4:inf+tnf+il2+ counts by groups.

Figure 7.4 shows that the logarithmic transformation of the cd4:inf+tnf+il2+ counts is fairly symmetrical across groups 2 and 3, while group 1 appears



Group 1 is the HIV infected infants, Group 2 the HIV exposed and uninfected infants and Group 3 the HIV unexposed and uninfected infants.

Figure 7.2: **Histograms of the cd4:inf+tnf+il2+ counts by treatment group**.

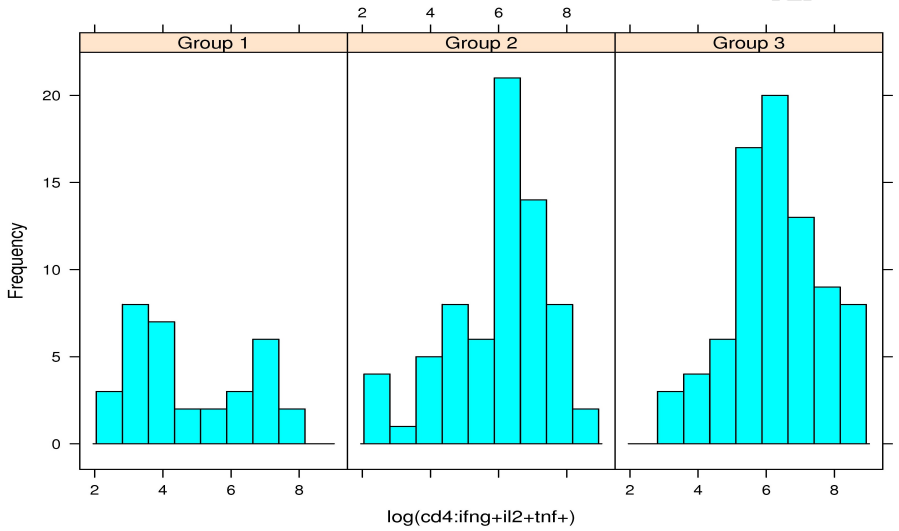


Group 1 is the HIV infected infants, Group 2 the HIV exposed and uninfected infants and Group 3 the HIV unexposed and uninfected infants.

Figure 7.3: **Boxplots of the cd4:inf+tnf+il2+ counts by groups**.

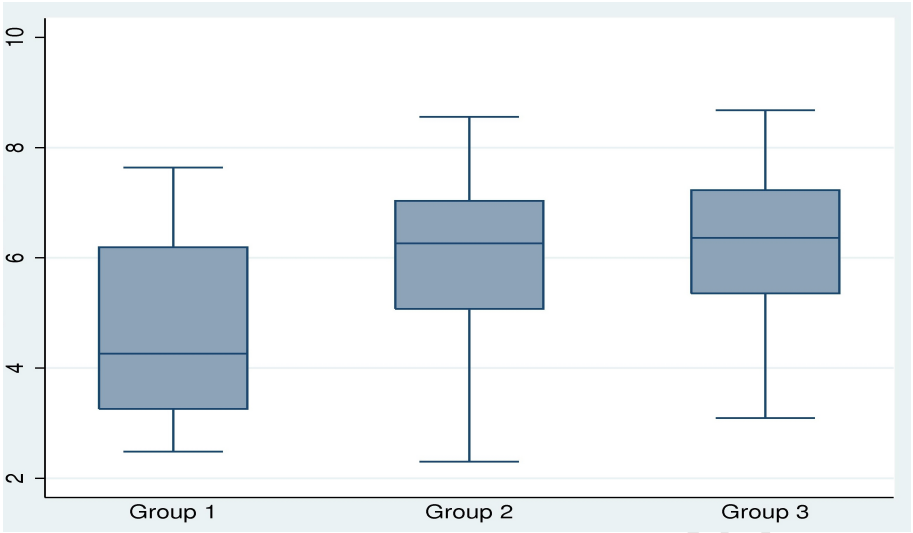
to have two modes. The boxplots of log(cd4:inf+tnf+il2+) (Figure 7.5) show that the levels of log(cd4:inf+tnf+il2+) are not the same with differing medians and interquartile ranges. It can be seen from Figure 7.5 that the transformation has removed the potentially outlying observations which were identified in Figure 7.3.

In order to test whether the means of the logarithm of the cd4:inf+tnf+il2+ counts were the same for all treatment groups an analysis of variance (ANOVA) test was conducted (Miller, 1997). The p-value calculated by applying this test was less than 0.0001, thus implying that there is a highly significant difference in the means across these groups.



Group 1 is the HIV infected infants, Group 2 the HIV exposed and uninfected infants and Group 3 the HIV unexposed and uninfected infants.

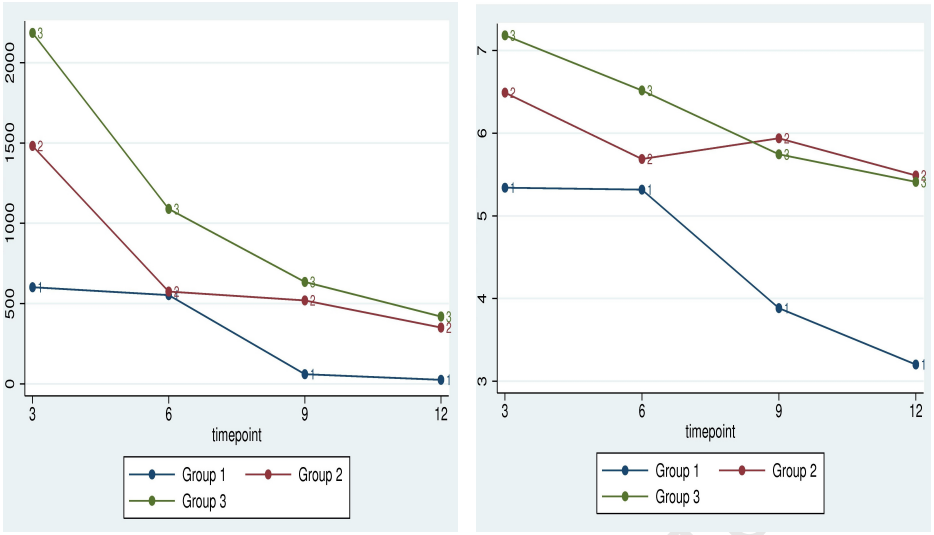
Figure 7.4: **Histograms of the logarithm of the cd4:inf+tnf+il2+ counts against time by treatment group**.



Group 1 is the HIV infected infants, Group 2 the HIV exposed and uninfected infants and Group 3 the HIV unexposed and uninfected infants.

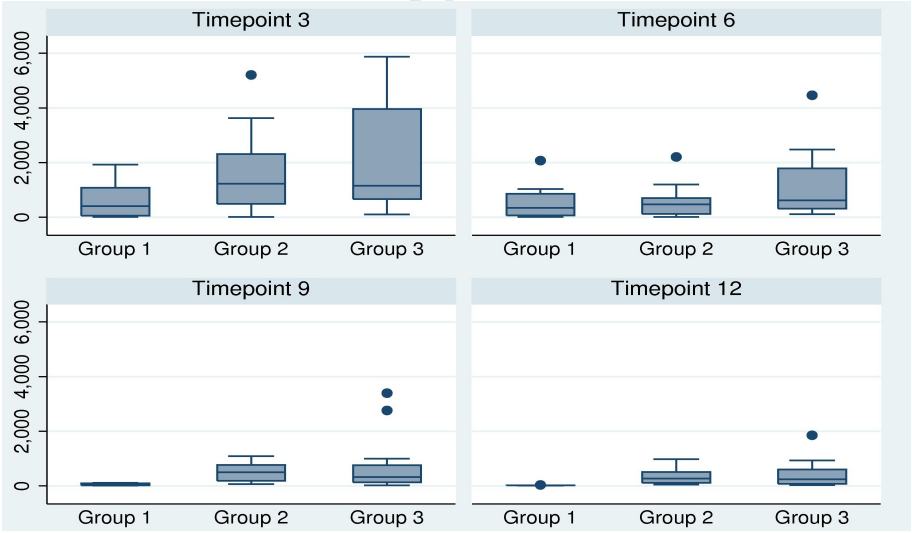
Figure 7.5: **Boxplots of the logarithm of the cd4:inf+tnf+il2+ counts by groups**.

The mean profiles for the levels cd4:inf+tnf+il2+ over time are shown in Figure 7.6a. This figure shows that in general the levels of cd4:inf+tnf+il2+ decrease over time with Group 2 and Group 3 being consistently higher than Group 1. The same can be said about Figure 7.6b which is the mean profile of the logarithmic transformation of the levels of cd4:inf+tnf+il2+. It can thus be concluded that time might also be a factor which affects the levels of cd4:inf+tnf+il2+ in the subjects. This is further supported by the boxplots of cd4:inf+tnf+il2+ and log(cd4:inf+tnf+il2+) over time, which are Figure 7.7 and Figure 7.8 respectively. These boxplots also emphasize that the difference in the levels of the cd4:inf+tnf+il2+ counts and its logarithmic transformation across groups is still apparent at all time points.



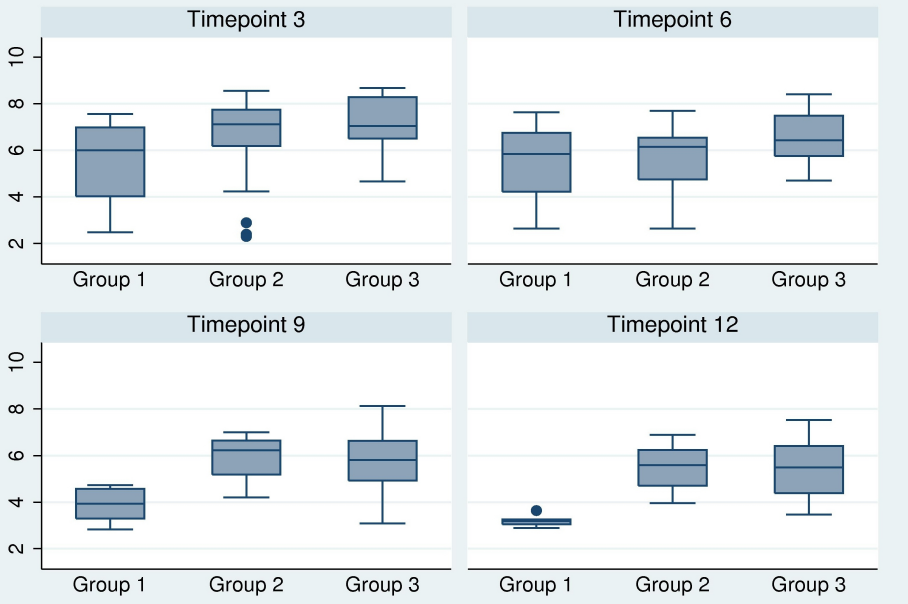
(a) Mean profile of cd4:inf+tnf+il2+ (b) Mean profile of log(cd4:inf+tnf+il2+)
Group 1 is the HIV infected infants, Group 2 the HIV exposed and uninfected infants and Group 3 the HIV unexposed and uninfected infants.

Figure 7.6: **Mean profiles of the cd4:inf+tnf+il2+ counts and the logarithm of the cd4:inf+tnf+il2+ counts.**



Group 1 is the HIV infected infants, Group 2 the HIV exposed and uninfected infants and Group 3 the HIV unexposed and uninfected infants.

Figure 7.7: **Boxplots of the cd4:inf+tnf+il2+ counts against treatment group by time.**



Group 1 is the HIV infected infants, Group 2 the HIV exposed and uninfected infants and Group 3 the HIV unexposed and uninfected infants.

Figure 7.8: **Boxplots of the logarithm of the cd4:inf+tnf+il2+ counts against treatment group by time.**

A scatterplot of the cd4:inf+tnf+il2+ counts grouped by treatment group is presented in Figure 7.9 and it reveals that there may be a few potentially outlying observations present in the data. For example the observation at time 6 for subject 5, the observations at times 3 and 6 for subject 41, the observation at time 3 for subject 49 and the observations belonging to subjects 28 and 44. These potentially outlying observations are also present in the transformed data as shown in Figure 7.10. The logarithmic transformation also highlights outlying observations at the lower end of the scale such as time 3 of subject 53, time 6 of subject 17 and all the observations of subjects 32 and 38.

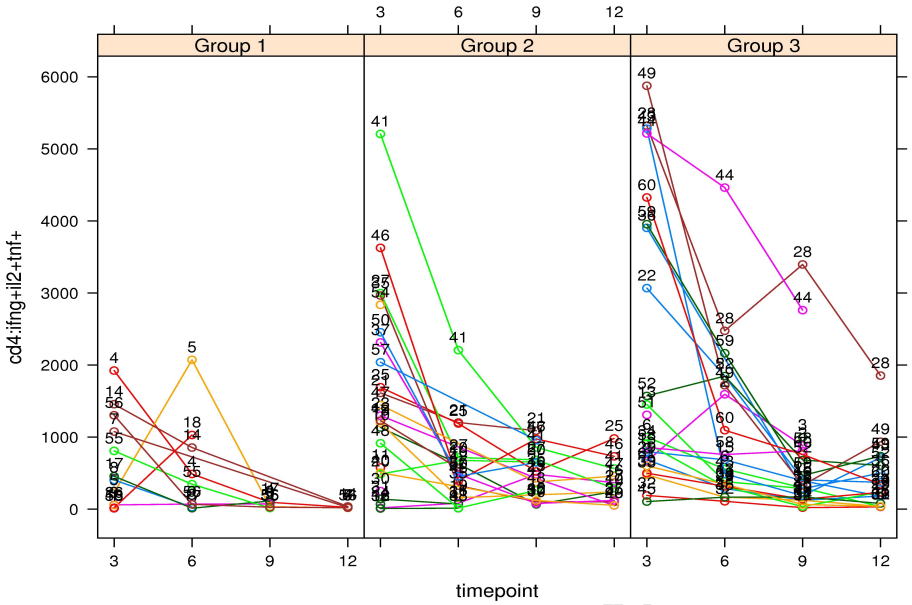


Figure 7.9: **Scatterplot of the cd4:inf+tnf+il2+ counts against by treatment group.**

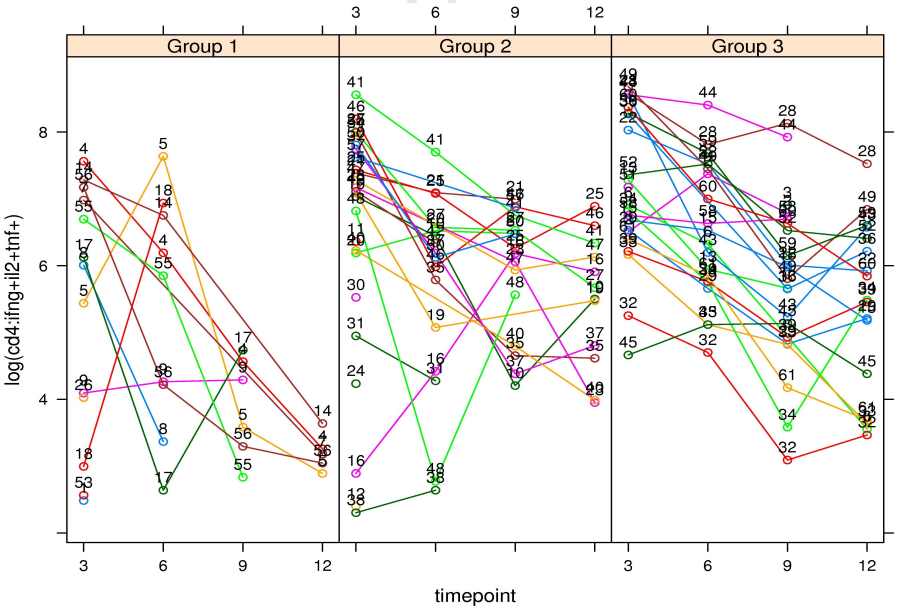


Figure 7.10: **Scatterplot of the logarithm of the cd4:inf+tnf+il2+ counts against time by treatment group.**

7.2 Cytokine data analyzed as normally distributed responses

In this section the log-transformed cd4:inf+tnf+il2+ count response was analyzed using the **lme4** package in R (Bates et al., 2012) and the code for this is found in the appendix. The analysis was also conducted using GENSTAT (Payne et al., 2011) and similar results were obtained.

The null model fitted to the data (M_{G0}) is given as:

$$y_{ij} = \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i3} + \beta_3 t_{ij} + b_i + e_{ij},$$

where y_{ij} is the logarithm of the cd4:inf+tnf+il2+ count response for the i^{th} patient during their visit at time j, x_{i2} and x_{i3} are covariates recorded as a value of one when the patient is from group 2 and group 3 respectively, and zero otherwise. The fixed terms in the model are the constant term (β_0), the effect of group 2 relative to group 1 (β_1), the effect of group 3 relative to group 1 (β_2) and the effect of time (β_3).

The addition of the group-time interaction terms did not improve the model significantly, thus this was not included in the model. The random effect for the ith subject is given as b_i for subjects i, $i = 1, ..., 61$ and e_{ij} is the random error. The random effects and random error are assumed to follow a normal distribution with properties $b_i \sim \mathrm{N}(0, \sigma_b^2)$ and $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$. These variance components are also assumed to be independent such that $\mathrm{cov}(b_i, e_{ij}) = 0$.

Model M_{G0} can also be written in the general form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e},$$

where $\mathbf{Z}$ is a $n \times q$ design matrix for a q-dimensional vector of unknown random effects $\mathbf{b}$, $\mathbf{X}$ is a $n \times p$ design matrix of covariates for a p-dimensional vector of coefficients $\boldsymbol{\beta}$ (with $n = 182$, $p = 4$ and $q = 61$), and $\mathbf{e}$ represents the vector of residual errors.

The VSOM was applied to each of the observations in turn. The VSOM for the k^{th} observation takes the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \delta_k \mathbf{d}_k + \mathbf{e},$$

where δ_k is the random effect for the k^{th} observation, with $\delta_k \sim \mathrm{N}(0, \sigma_k^2)$, and $\mathbf{d}_k$ is a vector with 1 in the k^{th} position and zeros in the other positions. The results are shown in Figure 7.11. The size of the variance inflation for the k^{th} observation is given as $\omega_k = \sigma_k^2/\sigma_e^2$. In order to generate an empirical distribution for the deviance statistic, used to identify outlying observations, a parametric bootstrap procedure was used. The bootstrap procedure used is described in chapter 4. At the 95^{th} percentile cut-off observations 11 (subject 5 at time 6), 29 (subject 12 at time 3), 40 (subject 16 at time 3), 47 (subject 18 at time 3), 48 (subject 18 at time 6) and 143 (subject 48 at time 6) were identified as potential outliers, this is shown in Figure 7.11.

Model M_{G1} was then fitted to the data with the outlying observations treated as random effects and the results are shown in Table 7.1. There was a substantial decrease in the deviance statistic, from 593.9 to 551, showing that this was a better model than M_{G0}. The residual variance also decreased quite substantially from 0.916 to 0.597. There was also down-weighting of the observations when it came to estimation of the fixed effects as the estimates changed in value.

In order to identify outlying subjects the VSOM was applied to each of the subjects in turn. The VSOM for the i^{th} subject takes the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \zeta_i \mathbf{d}_i + \mathbf{e},$$

where ζ_i is the random effect for the i^{th} subject, with $\zeta_i \sim \mathrm{N}(0, \sigma_i^2)$, and $\mathbf{d}_i$ is a vector with values of 1 in the positions corresponding to the i^{th} subject and zeros in the other positions. The results are illustrated in Figure 7.12. A parametric bootstrap was used to get the empirical distribution of the deviance statistic in order to find cut-off values which would be used to determine whether a subject is a potential outlier. After applying the parametric bootstrap procedure subjects 12 and 38 were identified as potential outliers at the 95^{th} percentile cut-off level. Model M_{G2} was then fitted with these subjects treated as random effects and the results are shown in Table 7.1. The results show that there was substantial down-weighting of the effect of these subjects on the estimation of the fixed effects. However, it can be seen that this model is not an improvement to model M_{G1} as it has a higher deviance (579 compared to 551) as well as a higher residual variance

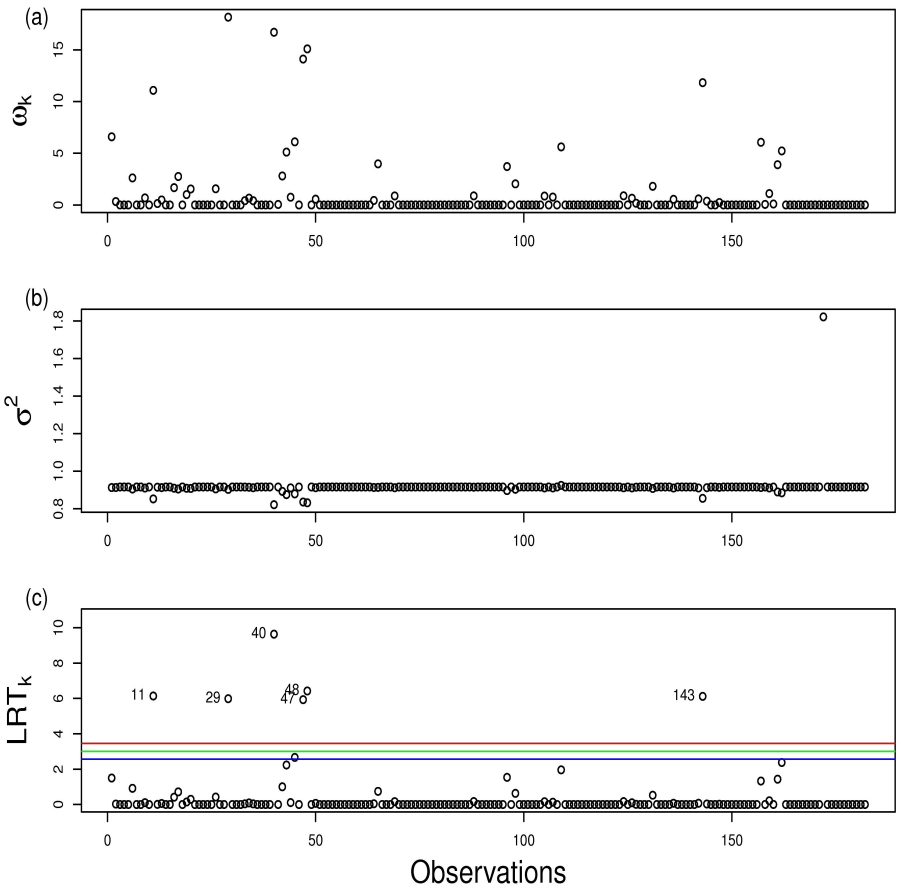


Figure 7.11: **VSOM statistics plotted against observation number for the log-cytokine dataset.** (a) Variance shift estimates, ω_k. (b) Residual variance estimates, σ^2. (c) Likelihood ratio statistics, LRT_k, with 95th percentile of the empirical distribution under the null hypothesis shown for the first r order statistics for each test: $r = 4$ (red line), $r = 5$ (green line) and $r = 6$ (blue line).

(0.903 compared to 0.597).

Model M_{G3} which treats observations 11 (subject 5 at time 6), 29 (subject 12 at time 3), 40 (subject 16 at time 3), 47 (subject 18 at time 3), 48 (subject 18 at time 6) and 143 (subject 48 at time 6) as random effects as well as subjects 12 and 38 was then fitted to data. This model also showed substantial down-weighting of the effect of these values in the estimation of the fixed effects. Also the deviance associated with the model is considerably lower than the other fitted models (541.2). The form of model M_{G3} is

$$\begin{aligned}\mathbf{y} =&\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \delta_{11}\mathbf{d}_{11} + \delta_{29}\mathbf{d}_{29} + \delta_{40}\mathbf{d}_{40} + \delta_{47}\mathbf{d}_{47} + \delta_{48}\mathbf{d}_{48} + \delta_{143}\mathbf{d}_{143} \\ &+ \zeta_{12}\mathbf{d}_{sub(12)} + \zeta_{38}\mathbf{d}_{sub(38)} + \mathbf{e},\end{aligned}$$

where δ_k is the random effect for the k^{th} observation and $\mathbf{d}_k$ a vector with 1 in the k^{th} position and zeros in the other positions; ζ_i is the random effect for the i^{th} subject and $\mathbf{d}_{sub(i)}$ is vector with values of 1 in the positions corresponding to the i^{th} subject and zeros in the other positions.

In order to compare the parameter estimates of the VSOM and case-deletion model, whereby the potentially outlying observations where deleted from the dataset, model M_{G4} was fitted. In this model all potentially outlying observations and subjects identified by the VSOM were deleted and then the null model was fitted to this data, in total 8 observations out of 182 were deleted. From Table 7.1 it can be seen that the estimated parameters of the fixed parameters were fairly similar. It is apparent that the VSOM provides similar estimates as compared to case-deletion, with the added advantage that data is not deleted but down-weighted. The level of down-weighting depends on the extent to which the observation identified is outlying from the rest of the data. This feature of the VSOM is very valuable when a researcher is unsure about whether an observation is truly an outlier.

Model M_{G3} was used to interpret the fixed effect coefficients. It can be seen that the coefficients of the baseline effects of group 2 and group 3 relative to group 1, imply that being in group 2 increases the level of cd4:inf+tnf+il2+ cells by 6.30 fold ($6.30 = e^{1.841}$) relative to group 1. While being in group 3 increases the level of the cells by 8.25 fold ($8.25 = e^{2.110}$). The time trend implies that a one unit increase in the time points will result

in a decrease of cd4:inf+tnf+il2+ cells by 19.5% (19.5% = (1 - $e^{-0.217}$) * 100).

Additional model checking was performed on the null model (M_{G0}) using residual analysis and the results are shown in Figure 7.13 and Figure 7.14 for individual observations. Figure 7.15 and Figure 7.16 are used to analyze subjects. These figures were plotted using STATA (StataCorp, 2009).

The histogram of standardized error residuals (Figure 7.13) can be seen to be fairly symmetrical and normally distributed with no extreme tails. The scatterplot of residuals does reveal that there are a few outlying observations in the data. These observations are the same ones that are identified by the VSOM.

Figure 7.15 has a long tail to the left thus showing the presence of an outlying subject in the data. The scatterplot of the standardized subject residuals shows that subjects 12 and 38 are the most outlying subjects. These subjects were identified by the VSOM.

Table 7.1: **Parameter estimates of models fitted to the log-cytokine dataset assuming an underlying normal distribution.**

Parameter	M_{G0}	M_{G1}	M_{G2}	M_{G3}	M_{G4}
constant	5.756 (0.366)	5.730 (0.351)	5.809 (0.321)	5.758 (0.318)	5.769 (0.322)
group 2	1.382 (0.419)	1.699 (0.409)	1.622 (0.364)	1.841 (0.368)	1.870 (0.372)
group 3	1.991 (0.422)	2.142 (0.409)	1.949 (0.360)	2.110 (0.365)	2.113 (0.369)
time	-0.200 (0.022)	-0.217 (0.018)	-0.202 (0.022)	-0.217 (0.018)	-0.219 (0.018)
σ^2	0.916	0.597	0.903	0.600	0.600
σ^2_{sub}	1.092	1.085	0.707	0.811	0.817
ω_{11}		16.17		16.05	
ω_{29}		29.42			
ω_{40}		24.32		24.63	
ω_{47}		7.21		7.25	
ω_{48}		11.57		10.79	
ω_{143}		18.03		18.46	
ψ_{12}			30.05	32.26	
ψ_{38}			25.55	26.98	
deviance	593.9	551	579	541.2	501.5

Subject 12 has only one observation which is observation 29 hence in model M_{G3}, ω_{29} is omitted.

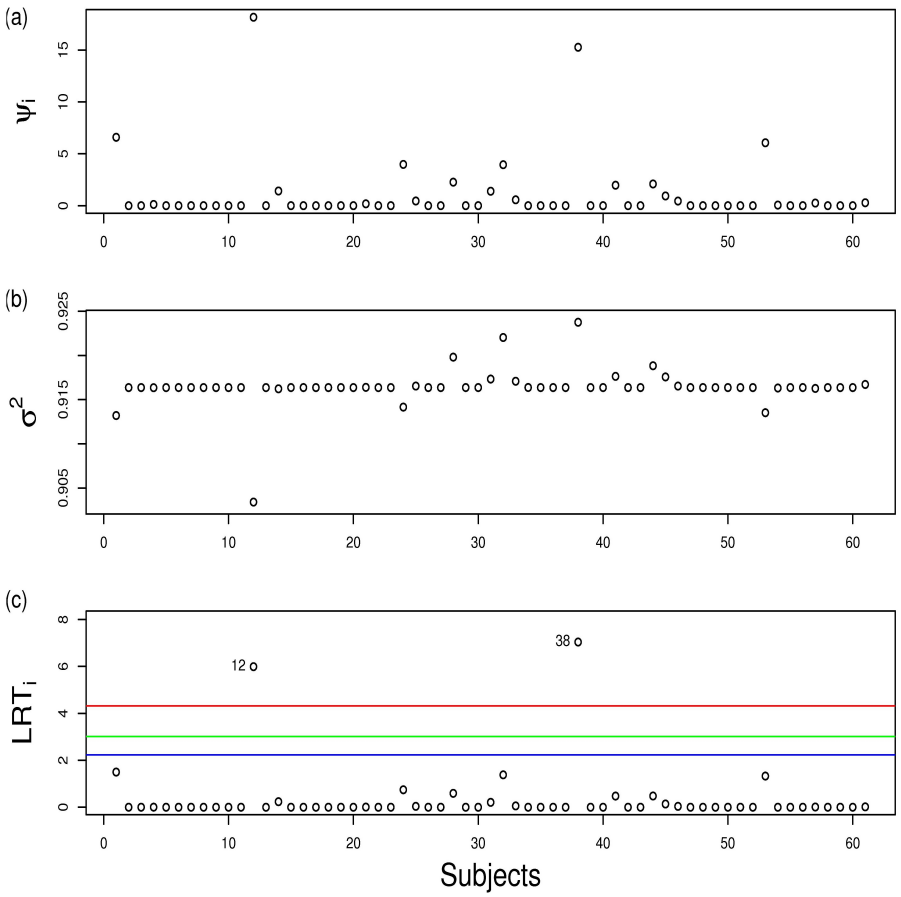


Figure 7.12: **VSOM statistics plotted against subject number for the log-cytokine dataset.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, σ^2. (c) Likelihood ratio statistics, LRT_i, with 95th percentile of the empirical distribution under the null hypothesis shown for the first r order statistics for each test: $r = 2$ (red line), $r = 3$ (green line) and $r = 4$ (blue line).

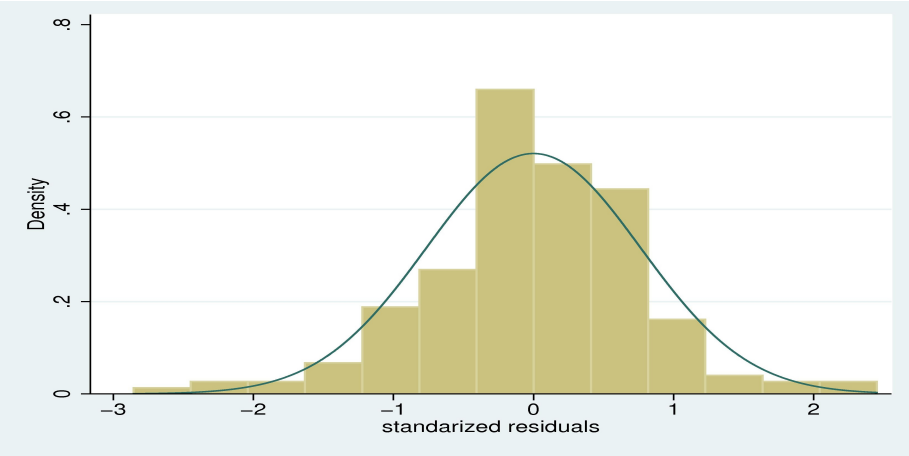


Figure 7.13: **Histogram of standardized error residuals for the log-cytokine dataset.**

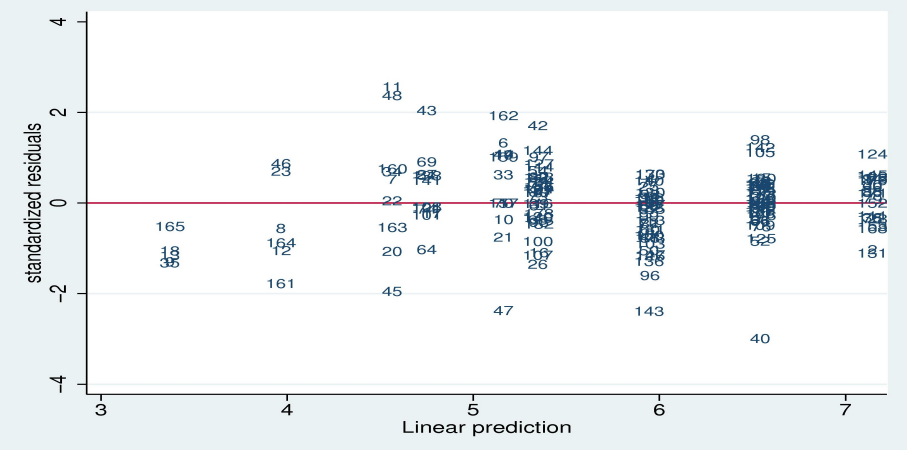


Figure 7.14: **Scatterplot of standardized residuals for the log-cytokine dataset.**

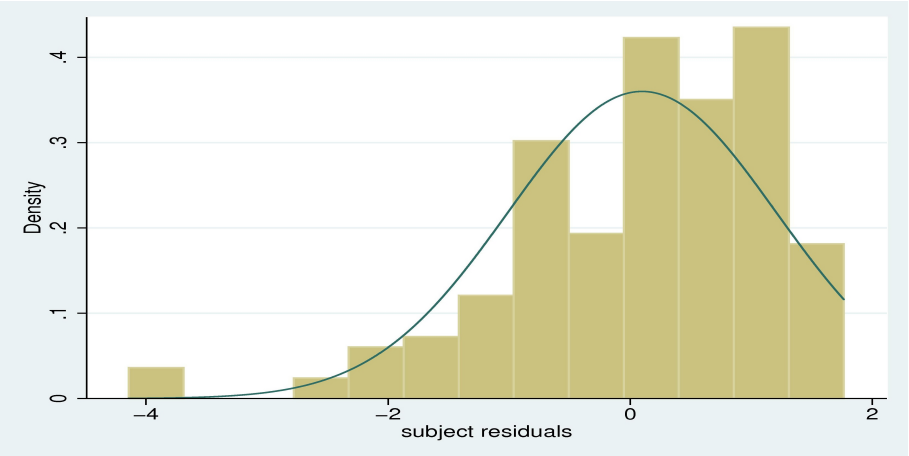


Figure 7.15: **Histogram of standardized subject residuals for the log-cytokine dataset.**

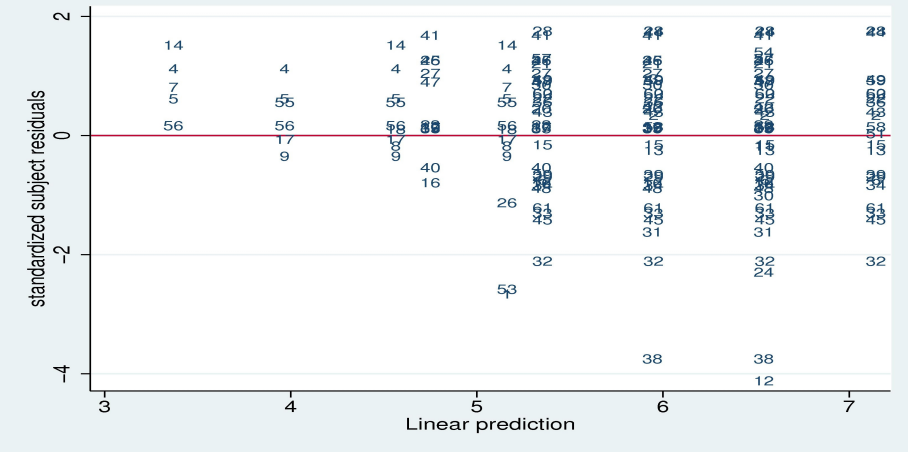


Figure 7.16: **Scatterplot of standardized subject residuals for the log-cytokine dataset.**

In conclusion it has been shown that the VSOM is able to identify the same outliers as standard residual analysis. The VSOM is also able to down-weight the influence of potentially outlying observations on the estimation of parameters. The down-weighting is very important when it is not clear if an observation is truly an outlier.

7.3 Cytokine dataset analyzed as counts

In this section I will analyze the cytokine dataset (Mansoor et al., 2009) as counts using a quasi-Poisson-gamma and a negative binomial HGLM approach.

7.3.1 Quasi-Poisson-gamma HGLM

The null model (M_{C0}) fitted to the data used the group effect and time effects as covariates based on model selection which will not be shown in this thesis. The linear predictor of model M_{C0} is given as

$$\log(\mathrm{E}(\mathbf{Y}|\mathbf{b}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu_b},$$

where b_i is the random effect for the i^{th} subject for $i = 1, \ldots, 61$, $\nu_{b_i} = \log(b_i)$ with b_i following a gamma distribution with a mean of one and dispersion parameter of γ. Based on this model the marginal variance is given by

$$\mathrm{var}(Y_{ij}) = \phi_{ij}\mu_{ij} + \gamma\mu_{ij}^2.$$

The VSOM was then applied to all the observations in turn. The VSOM for the k^{th} observation, of the n-dimensional vector of responses $\mathbf{Y}$, takes the form

$$\log(\mathrm{E}(\mathbf{Y}|\mathbf{b}, \delta_k)) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu} + \nu_{\delta_k}\mathbf{d}_k,$$

where $\nu_{\delta_k} = \log(\delta_k)$ with δ_k following a gamma distribution with dispersion parameter λ_k and $\mathbf{d}_k$ is a vector of length n with a value of one in the k^{th} position and zeros everywhere else. The results are shown in Figure 7.17. I will use the potential outliers identified at the 99^{th} percentile cut-off level to illustrate the VSOM (model M_{C1}). The observations are observation 11 (subject 5 at time 6), 47 (subject 18 at time 3), 124 (subject 43 at time 3)

and 125 (subject 43 at time 6). It can be seen from Table 7.2 that the fixed effect estimates are fairly similar for models M_{C0} and M_{C1}, thus indicating that the outlying observations are not influential observations. However, the VSOM does reduce the size of the dispersion parameter thus reducing the overdispersion in the model with a reduction in ϕ from 323.5 to 259.6. It can also be seen that model M_{C1} fits the data better than model M_{C0} with a deviance of 2769.20 compared to 2795.08.

In order to identify any potentially outlying subjects, the VSOM was applied to all the subjects in turn. The VSOM for the i^{th} subject takes the form

$$\log(\mathrm{E}(\mathbf{Y}|\mathbf{b},\zeta_i)) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu} + \nu_{\zeta_i}\mathbf{d}_{sub(i)},$$

where $\nu_{\zeta_i} = \log(\zeta_i)$ with ζ_i following a gamma distribution with a dispersion parameter of ψ_i and $\mathbf{d}_{sub(i)}$ is a vector of length n with values of 1 in positions corresponding to the i^{th} subject and zeros everywhere else. The results are shown in Figure 7.18. From this figure it can be seen that subjects 12, 32 and 38 are potentially outlying subjects. Model M_{C2} was fitted using the outlying subjects as random effects. This model gave fixed estimates which were similar to the null model thus the subjects were not influential subjects.

Model M_{C3} was fitted using the outlying observations and subjects as random effects. The linear predictor for model M_{C3} is given by

$$\begin{aligned}\log(\mathrm{E}(\mathbf{Y}|\mathbf{b},\delta_k,\zeta_i)) =& \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu} + \nu_{\delta_{11}}\mathbf{d}_{11} + \nu_{\delta_{47}}\mathbf{d}_{47} + \nu_{\delta_{124}}\mathbf{d}_{124} + \nu_{\delta_{125}}\mathbf{d}_{125} \\ &+ \nu_{\zeta_{12}}\mathbf{d}_{sub(12)} + \nu_{\zeta_{32}}\mathbf{d}_{sub(32)} + \nu_{\zeta_{38}}\mathbf{d}_{sub(38)},\end{aligned}$$

where δ_k is the random effect for the k^{th} observation and $\mathbf{d}_k$ a vector with 1 in the k^{th} position and zeros in the other positions; ζ_i is the random effect for the i^{th} subject and $\mathbf{d}_{sub(i)}$ is vector with values of 1 in the positions corresponding to the i^{th} subject and zeros in the other positions.

Model M_{C3} fit the data better than models M_{C0}, M_{C1} and M_{C2} with a deviance of 2760.38. The size of the overall overdispersion is considerably reduced by the use of the VSOM in this model as ϕ dropped from 323.5 in the null model to 254.9 in model M_{C3}. When the potentially outlying observations and subjects were deleted (model M_{C4}) the fixed effect estimates and deviance statistics were found to be similar to those of model M_{C3}.

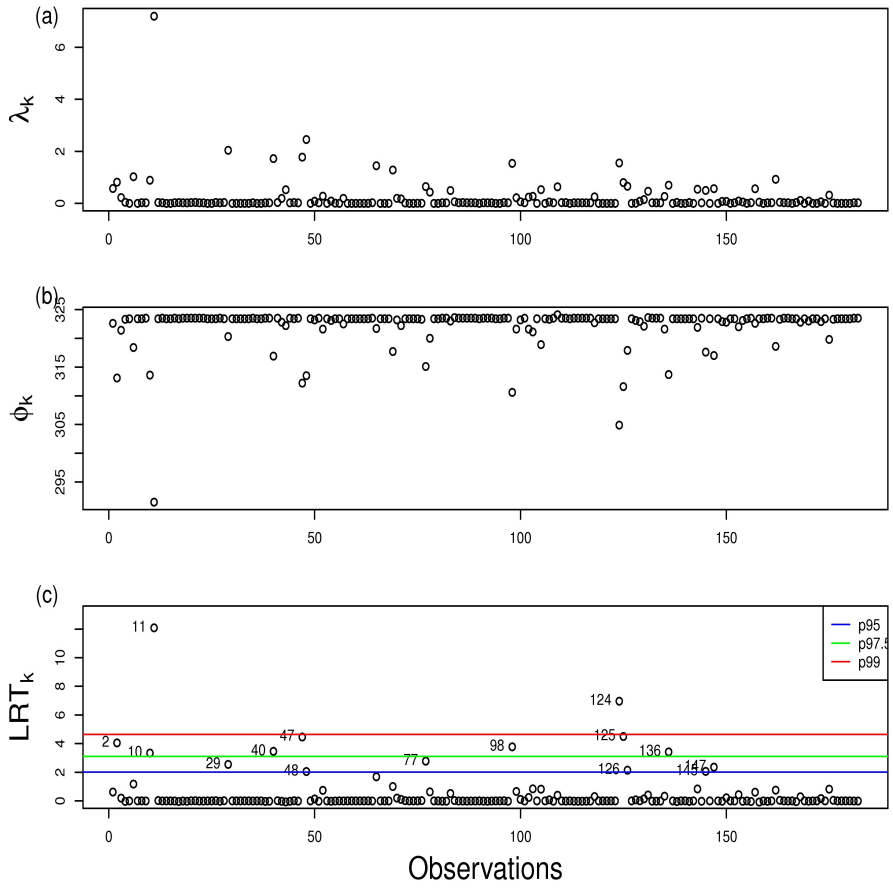


Figure 7.17: **VSOM statistics plotted against observations for the cytokine count dataset analyzed using a quasi-Poisson-gamma model.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, ϕ_k. (c) Likelihood ratio statistics, LRT_k, with $95^{th}, 97.5^{th}$ and 99^{th} percentile cut-off values.

Table 7.2: **Parameter estimates of quasi-Poisson-gamma models fitted to the count cytokine dataset.**

Parameter	M_{C0}	M_{C1}	M_{C2}	M_{C3}	M_{C4}
constant	7.036 (0.267)	6.977 (0.273)	7.041 (0.255)	6.978 (0.260)	6.972 (0.262)
group 2	0.810 (0.303)	0.849 (0.311)	0.867 (0.289)	0.912 (0.295)	0.934 (0.298)
group 3	1.268 (0.303)	1.282 (0.313)	1.296 (0.288)	1.313 (0.295)	1.331 (0.298)
time	-0.204 (0.017)	-0.201 (0.015)	-0.205 (0.017)	-0.202 (0.015)	-0.201 (0.015)
ϕ	323.5 (39.47)	259.6 (32.19)	317.0 (0.122)	254.9 (31.61)	259.2 (32.92)
γ	0.497 (0.100)	0.547 (0.108)	0.425 (0.211)	0.464 (0.095)	0.470 (0.097)
λ_{11}		7.377 (9.295)		7.199 (9.071)	
λ_{47}		2.042 (2.675)		2.028 (2.677)	
λ_{124}		1.356 (1.993)		1.289 (1.908)	
λ_{125}		0.154 (0.383)		0.176 (0.411)	
ψ_{12}			2.222 (1.31)	2.526 (3.233)	
ψ_{32}			1.899 (1.30)	1.907 (2.479)	
ψ_{38}			2.777 (1.25)	3.024 (3.750)	
deviance	2795.08	2769.20	2786.57	2760.38	2598.82

where γ is the dispersion parameter of the subject random effect, λ_k is the dispersion parameter of the k^{th} observation and ψ_i is the variance of the i^{th} subject.

Model M_{C3} however, has the added advantage that there is no loss of data (11 out of 182 observations deleted) especially when a researcher is unsure about whether an observation or subject is truly outlying.

Model M_{C3} was used to interpret the fixed effect coefficients. It can be seen that the coefficients of the baseline effects of group 2 and group 3 relative to group 1 imply that being in group 2 increases the level of cd4:inf+tnf+il2+ cells by 2.49 fold ($2.49 = e^{0.912}$) relative to group 1. While being in group 3 increases the level of the cells by 3.72 fold ($3.72 = e^{1.313}$). The time trend implies that a one unit increase in the time points will result in a decrease of cd4:inf+tnf+il2+ cells by 18.3% (18.3% = (1 - $e^{-0.202}$) * 100).

Residual analysis using plots of the absolute residuals for the individual observations (Figure 7.19) was also conducted. It can be seen from this graphic that the VSOM was able to identify similar outlying observations as standard residual analysis.

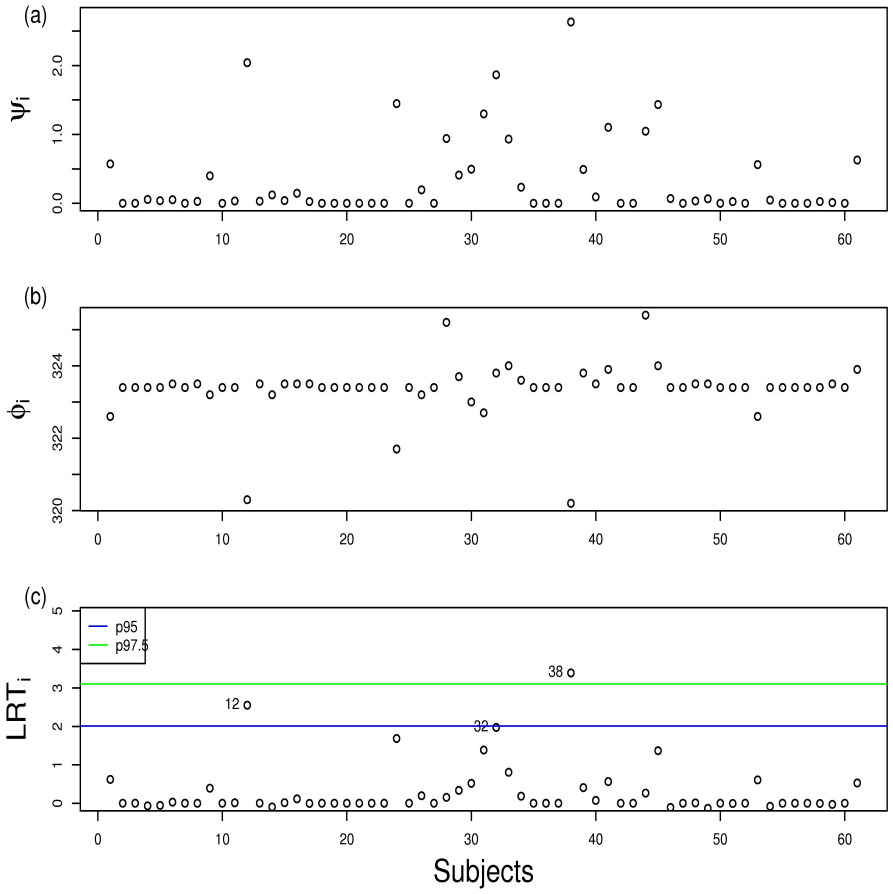


Figure 7.18: **VSOM statistics plotted against patient number for the cytokine count dataset analyzed using a quasi-Poisson-gamma model.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, ϕ_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values.

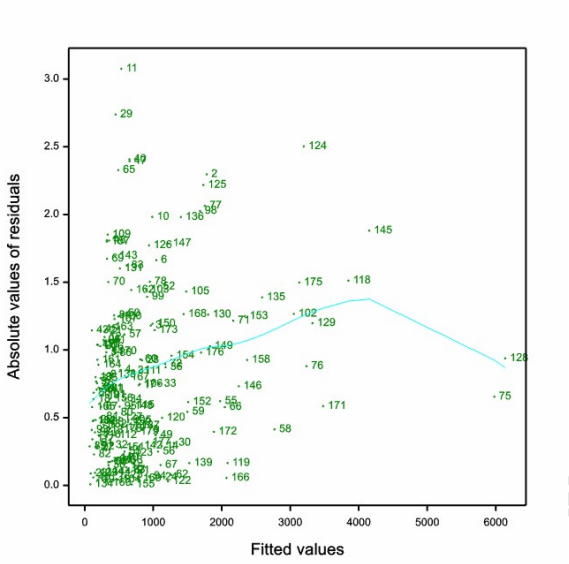


Figure 7.19: **Absolute residuals against fitted values plot for individual observations from the cytokine count dataset using the quasi-Poisson-gamma HGLM as the null model**

7.3.2 Negative binomial HGLM

In this section I will use the negative binomial HGLM when applying the VSOM (Lee and Nelder, 2000). This model depends on using saturated gamma distributed random effects, $\boldsymbol{\varsigma}$. The null model, M_{N0}, which is given by

$$\log(\mathrm{E}(\mathbf{Y}|\mathbf{b},\mathbf{s})) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_s\boldsymbol{\nu_s} + \mathbf{Z}_b\boldsymbol{\nu_b},$$

where $\mathbf{Z}_s$ is a $n \times n$ design matrix with values of 1 along the diagonal and zeros everywhere else with s being the observation random effect which follows a gamma distribution with a mean of one and dispersion parameter of α, $\mathbf{Z}_b$ is a $n \times q$ design matrix with values of 1 corresponding to positions where the i^{th} subject is observed and zeros everywhere else with b being the subject random effect which follows a gamma distribution with a mean of one and dispersion parameter of γ. The marginal variance of Y_{ij} is given by

$$\mu_{ij} + \alpha_{ij}\mu_{ij}^2 + \gamma\mu_{ij}^2.$$

The results of the model fit are shown in Table 7.3. It can be seen from this table that the deviance for the negative binomial null model (M_{N0}) is lower than the deviance from the quasi-Poisson-gamma HGLM (M_{C0}) which was fit in the previous section with a deviance of 2720.16 relative to 2795.08. The VSOM was applied to each observation in turn. The VSOM for the k^{th} observation, of the n-dimensional vector of responses $\mathbf{Y}$, takes the form

$$\log(\mathrm{E}(\mathbf{Y}|\mathbf{b},\mathbf{s},\delta_k)) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_s\boldsymbol{\nu_s} + \mathbf{Z}_b\boldsymbol{\nu_b} + \nu_{\delta_k}\mathbf{d}_k,$$

where δ_k follows a gamma distribution with dispersion parameter λ_k and $\mathbf{d}_k$ is a vector of length n with a value of one in the k^{th} position and zeros everywhere else. The results are shown in Figure 7.20. In order to illustrate the VSOM I will use observations identified as outliers at the 99^{th} percentile level as random effects in model M_{N1}. The observations are observation 11 (subject 5 at time 6), 40 (subject 16 at time 3) and 47 (subject 18 at time 3). The results of this model fitting are shown in Table 7.3. Model M_{N1} fits the data better than M_{N0} as shown by the lower deviance of 2697.39 compared to 2720.16. The fixed effect estimates for model M_{N1} are similar to those of model M_{N0} thus the outlying observations are not influential observations.

In order to identify any potentially outlying subjects, the VSOM was applied to all the subjects in turn. The VSOM for the i^{th} subject takes the form

$$\log(\mathrm{E}(\mathbf{Y}|\mathbf{b},\mathbf{s},\zeta_i)) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_s\boldsymbol{\nu_s} + \mathbf{Z}_b\boldsymbol{\nu_b} + \nu_{\zeta_i}\mathbf{d}_{sub(i)},$$

where ζ_i follows a gamma distribution with a dispersion parameter of ψ_i and $\mathbf{d}_{sub(i)}$ is a vector of length n with values of 1 in positions corresponding to the i^{th} subject and zeros everywhere else. The results are shown in Figure 7.21. To illustrate the VSOM I used subjects 12 and 38 as random effects in model M_{N2}. This model did not fit the data better than model M_{N1} thus a model with the potentially outlying observations and subjects treated as random effects was fit to the data (model M_{N3}). The linear predictor of model M_{N3} is given by

$$\begin{aligned}\log(\mathrm{E}(\mathbf{Y}|\mathbf{b},\mathbf{s},\delta_k,\zeta_i)) =& \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_s\boldsymbol{\nu}_s + \mathbf{Z}_b\boldsymbol{\nu}_b + \nu_{\delta_{11}}\mathbf{d}_{11} + \nu_{\delta_{40}}\mathbf{d}_{40} + \nu_{\delta_{47}}\mathbf{d}_{47} \\ &+ \nu_{\zeta_{12}}\mathbf{d}_{sub(12)} + \nu_{\zeta_{38}}\mathbf{d}_{sub(38)}.\end{aligned}$$

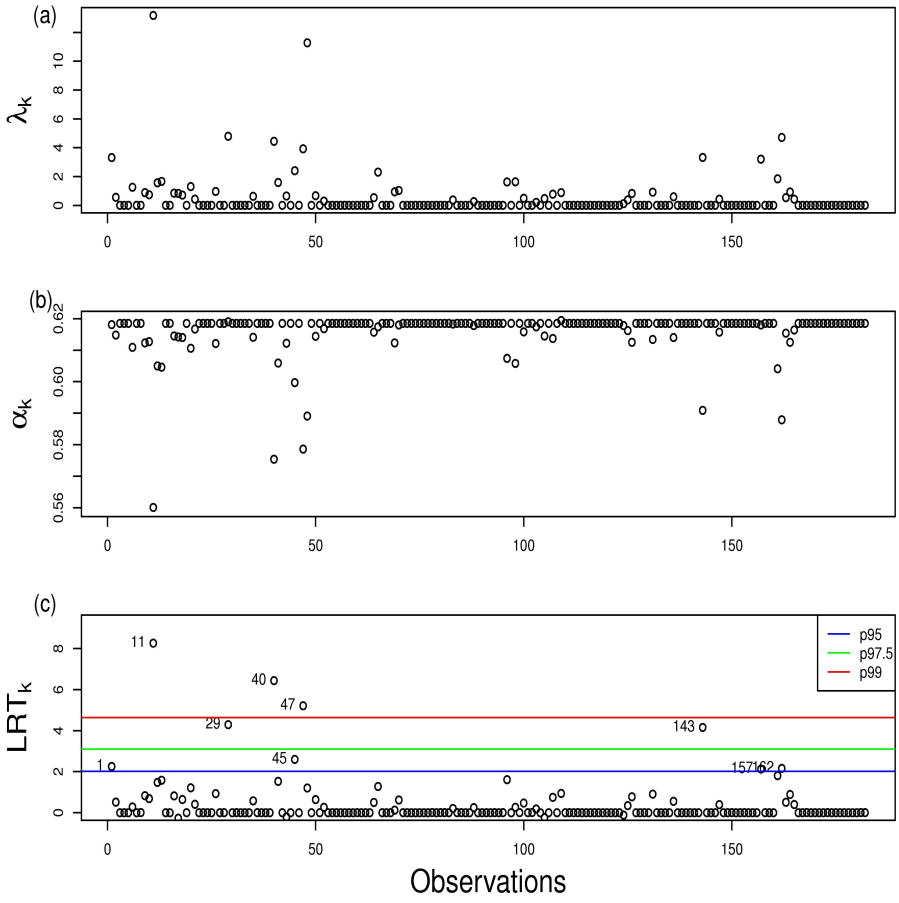


Figure 7.20: **VSOM statistics plotted against observations for the cytokine count dataset analyzed using a negative binomial HGLM.** (a) Variance shift estimates, λ_k. (b) Dispersion parameter estimates, α_k. (c) Likelihood ratio statistics, LRT_k, with $95^{th}, 97.5^{th}$ and 99^{th} percentile cut-off values.

Table 7.3: **Parameter estimates of models fitted to the count cytokine data using a negative binomial HGLM null model.**

Parameter	M_{N0}	M_{N1}	M_{N2}	M_{N3}	M_{N4}
constant	6.709 (0.304)	6.647 (0.308)	6.713 (0.273)	6.625 (0.278)	6.633 (0.280)
group 2	1.066 (0.357)	1.205 (0.364)	1.147 (0.319)	1.302 (0.330)	1.322 (0.333)
group 3	1.537 (0.362)	1.677 (0.370)	1.502 (0.319)	1.677 (0.331)	1.681 (0.333)
time	-0.207 (0.019)	-0.214 (0.016)	-0.207 (0.019)	-0.214 (0.017)	-0.216 (0.017)
α	0.619 (0.072)	0.476 (0.057)	0.635 (0.072)	0.484 (0.058)	0.484 (0.058)
γ	0.810 (0.156)	0.886 (0.165)	0.565 (0.119)	0.669 (0.132)	0.679 (0.134)
λ_{11}		14.16 (17.13)		13.59 (16.58)	
λ_{40}		4.51 (5.32)		4.52 (5.33)	
λ_{47}		3.98 (4.73)		3.87 (4.60)	
ψ_{12}			4.89 (5.72)	4.95 (5.80)	
ψ_{38}			4.39 (5.19)	4.42 (5.22)	
deviance	2720.16	2697.39	2709.92	2687.10	2612.63

where γ is the dispersion parameter of the subject random effect, α is the dispersion parameter for all observations, λ_k is the dispersion parameter of the k^{th} observation and ψ_i is the dispersion parameter of the i^{th} subject.

Model M_{N3} gave similar fixed effect estimates to those found from deleting the potentially outlying observations and subjects.

Model M_{N3} was used to interpret the fixed effect coefficients. It can be seen that the coefficients of the baseline effects of group 2 and group 3 relative to group 1 imply that being in group 2 increases the level of cd4:inf+tnf+il2+ cells by 3.68 fold ($3.68 = e^{1.302}$) relative to group 1. While being in group 3 increases the level of the cells by 5.35 fold ($5.35 = e^{1.677}$). The time trend implies that a one unit increase in the time points will result in a decrease of cd4:inf+tnf+il2+ cells by 19.3% (19.3% = (1 - $e^{-0.214}$) * 100).

Residual analysis using plots of the absolute residuals for the individual observations (Figure 7.22) was also conducted. It can be seen from this graphic that the VSOM was able to identify similar outlying observations as standard residual analysis.

In conclusion it has been shown that the VSOM can be applied to various types of models which can handle overdispersion in count data including the quasi-Poisson-gamma HGLM and negative binomial HGLM. The advantage of the VSOM is the ability to down-weight outlying observations whilst preserving them in the dataset. This is beneficial when it is not clear whether to delete an observation from the study. It has also been shown, specifically for the cytokine dataset, that similar outliers are identified when the underlying models fit to the data are the quasi-Poisson gamma and negative binomial HGLM.

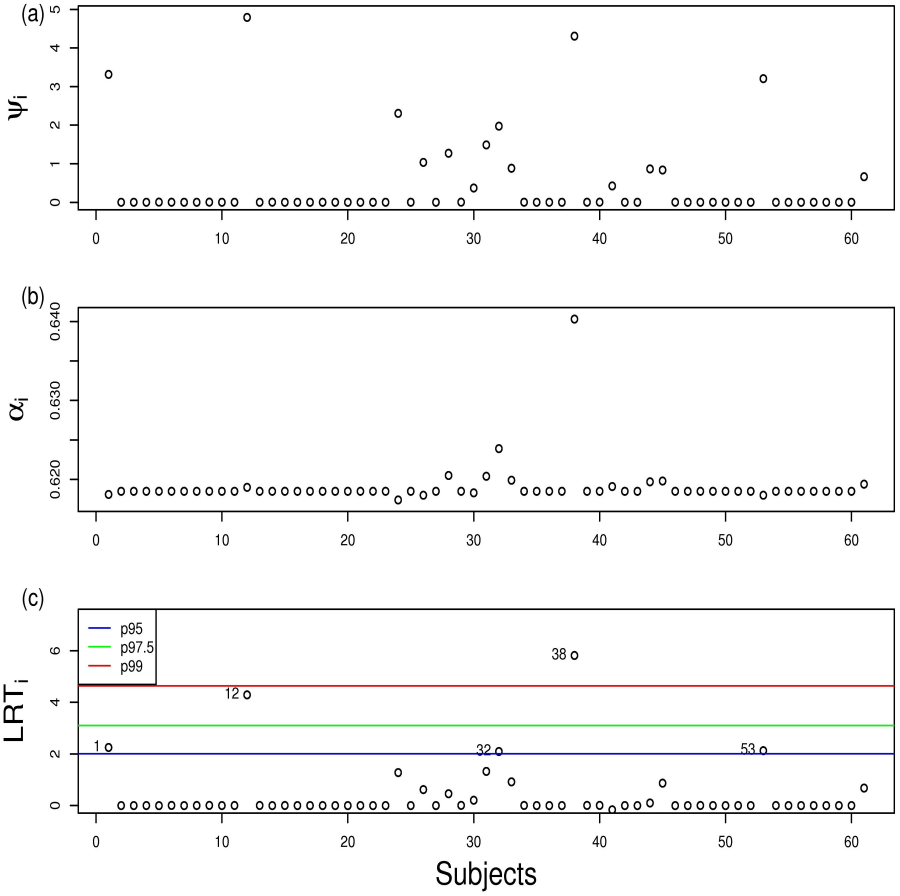


Figure 7.21: **VSOM statistics plotted against patient number for the cytokine count dataset analyzed using a negative binomial HGLM.** (a) Variance shift estimates, ψ_i. (b) Dispersion parameter estimates, α_i. (c) Likelihood ratio statistics, LRT_i, with 95^{th}, 97.5^{th} and 99^{th} percentile cut-off values.

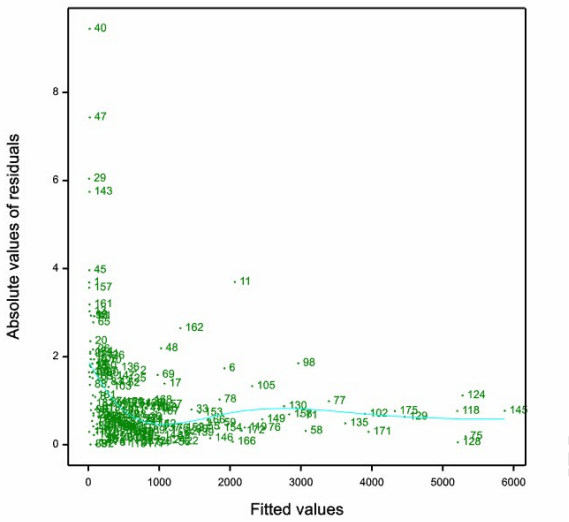


Figure 7.22: **Absolute residuals against fitted values plot for individual observations from the cytokine count dataset using the negative binomial HGLM as the null model**

7.4 Summary

In this section the cytokine data has been analyzed as normally distributed and count data. In all the models fitted to the various forms of this data the conclusion has been the same. The subjects who were in Group 2 (HIV negative and exposed) and Group 3 (HIV negative and unexposed) had substantially higher levels of CD4 T cells as compared to subjects in Group 1 (HIV positive). It was also revealed that there was a time trend which decreased the levels of CD4 T cells across all the groups. The general conclusion was that subjects in Group 1 were not assisted to a great degree by the BCG vaccination.

Another interesting result which arose from the study was the difference in the magnitude of the estimates when log-transformed responses and the count responses (using both the quasi-Poisson-gamma and negative binomial HGLM approaches) were used. A possible reason is that the log-transformation assumes that the mean is proportional to the standard error

while the assumption used in the Poisson models assumes that the mean is proportion to the variance hence the log-transformation has a stronger effect down-weighting effect on outliers as compared to the Poisson model approach. Another possible reason for the differences is that linear predictors for the linear mixed model, quasi-Poisson-gamma and negative binomial HGLM models were not the same. The difference between the linear mixed model and quasi-Poisson-gamma was the distribution of the random effects and the constraints applied to the random effects by their distributional properties. The linear mixed model used normally distributed random effects while the quasi-Poisson-gamma used gamma distributed random effects. Another difference is the way that extreme values are handled by these models. The presence of the dispersion parameter ϕ when analyzing count data down-weights the effect of extreme values in the estimation process. When using the linear mixed model the initial log-transformation to the data scales down extreme values before the estimation process even begins, thus the extent of the down-weighting is not going to be same for both models. The negative-binomial and quasi-Poisson-gamma HGLMs both assume that their respective random effects follow a gamma distribution. However, they have different covariance structures and linear predictors as shown in Chapter 5. These differences have been shown, in this example, to give rise to different parameter estimates.

It has been shown that the VSOM was able to identify similar potentially outlying observations and subjects regardless of the choice of the null model or the distribution of the response variable. Subjects 12 and 38 were identified as potentially outlying in the variance shift outlier models fitted when the data was assumed to be normally distributed and when the data were treated as counts. This is an interesting finding as it shows that the VSOM is able to give consistent results for normally distributed and count data.

Chapter 8

Conclusion

Several methods of analyzing overdispersed count and binary data have been presented in this thesis. These include the quasi-Poisson, quasi-binomial, negative binomial and the beta-binomial models for independent data. In biological areas longitudinal data are very common. Such data are associated with within subject correlation along with overdispersion. The two types of mixed effects models used in this thesis are GLMMs and HGLMs. The primary interest of this thesis was the detection and down-weighting of outliers in the presence of overdispersion and correlation.

In practice most researchers prefer to either delete outlying observations or use robust estimation techniques when handling outliers. The VSOM can be used as an alternative method. The VSOM is able to identify and down-weight the effect of outlying observations on model estimates, with the added advantage that data is preserved and not deleted. This study has shown that the methodology outlined by Gumedze et al. (2010) can be extended to count data as well as binomial data with overdispersion present. In the examples used in this study it was also seen that the VSOM is able to pick up the majority of outliers identified by standard residual analysis.

At present the VSOM in count data can only be applied using the HGLM procedure in GENSTAT. HGLMs have the added advantage of allowing random effects in mixed models to follow non-normal distributions. The HGLMs used in this study are the Poisson-normal, Poisson-gamma, quasi-Poisson-gamma, Poisson-gamma with saturated random effects, binomial-

normal and beta-binomial HGLMs. These HGLMs for count data provide simple interpretations of the variance shift, due to outlying observations, by using the linear nature of the marginal variance. The marginal variances for binomial data using the VSOM cannot be found in a closed form however, it has been shown that the VSOM is able to identify the same outlying observations as standard residual analysis.

It was shown in this thesis that the VSOM is able to identify and down-weight outlying observations with the added advantage that data is preserved and not deleted. Some fitted VSOM models may not converge because of poor starting values for estimated variance shift estimates. A possible solution to this problem would be to estimate only the variance parameters associated with the additional random effect under the VSOM when the VSOM is fitted, and keep the other variance components at their null model estimates. This is to help the variance component estimation process. Gumedze et al. (2010) adopted a similar fitting strategy and called the resulting variance estimates partial variances to reflect that they are not based on full iteration. The approach I have used in this thesis has not taken account of the problem of multiple testing which occurs as the VSOM is applied to all the observations in a dataset, further research will go into applying a parametric bootstrap to account for this.

There are other models which could possibly be used to model data in the presence of overdispersion and correlation. Kassahun et al. (2012) proposed a combined model which could be used to accommodate for overdispersion and correlation in data. This model used a combination of beta and normal random effects. The normally distributed random effects were used primarily to account for the correlation among repeated measures, they were also used to accommodate some of the overdispersion in the data. The beta distributed random effects would then account for the majority of the overdispersion in the model, by formulating a beta-binomial GLM for the fixed effects. The advantage of using the beta random effects is that they are a conjugate distribution of the binomial distribution, as a result the sensitivity of the model due to assumptions about the random effects is reduced (Molenberghs et al., 2010).

An alternative to the Poisson-gamma model is the Poisson model with

random effects which follow a generalized log-gamma distribution. This model was proposed by Fabio et al. (2012) and it was found that these random effects were able to accommodate the overdispersion in count data and as well as the within-cluster correlation. Numerical methods were used to derive the marginal models in the paper and the authors were also able to obtain a multivariate negative binomial model after setting parameter restrictions in the hierarchical model. The score function and Fisher information matrix for the multivariate negative binomial model were derived in this paper. An iterative process for obtaining the maximum likelihood estimates for the parameters in the multivariate negative binomial model along with goodness of fit procedures and residual analysis were also derived in the paper.

An area of further research would be the application of the VSOM to the models developed by Kassahun et al. (2012) and Fabio et al. (2012) for overdispersed data.

Count data with excess zeros (zero inflated data) is very common in medical data. Further research will also go into adapting the VSOM to count data with excess zeros (zero inflated data) (Yau1 et al., 2003). The zeros in this type of data can be modeled using a Bernoulli distribution while the counts that are greater than zero can be modeled using a Poisson distribution, as a result the VSOM can be applied theoretically to identify and down-weight outlying subjects in both the zero and non-zero parts of the data.

Chapter 9

References

Banerjee, M. and Frees, E. (1997). Influence diagnostics for longitudinal models, *Journal of the American Statistical Association* **92**: 999–1005.

Bates, D., Maechler, M. and Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-0.
URL: *http://CRAN.R-project.org/package=lme4*

Besag, J., P.J., H. and Mengersen, K. (1995). Bayesian computation and stochastic systems, *Statistical Science* **10**: 3–66.

Bissell, A. F. (1972). A negative binomial model with varying element sizes, *Biometrika* **59**: 435–441.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421): pp. 9–25.

Collett, D. (1996). *Modelling Binary Data*, Chapman and Hall.

Cook, R. (1977). Detection of influential variables in linear regression, *Technometrics* **19**: 15–19.

Cook, R. (1986). Assessment of local influence (with discussion), *Journal of the Royal Statistical Society* **48**: 133–169.

Cook, R., Holschuh, N. and Weisberg, S. (1982). A note on an alternative outlier model, *Journal of the Royal Statistical Society* **44**: 370–376.

Cook, R. and Weisberg, S. (1982). *Residuals and Influence in Regression.*, Chapman and Hall.

Crowder, M. (1978). Beta-binomial anova for proportions, *Applied Statistics* **27**: 34–37.

Davison, A. C. and Tsai, C.-L. (1992). Regression model diagnostics, *International Statistical Review* **60**: 337–353.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*, John Wiley and Sons.

Dobson, A. and Barnett, A. (2008). *An introduction to generalized linear models*, Texts in statistical science, CRC Press.

Fabio, L. C., Paula, G. A. and de Castro, M. (2012). A poisson mixed model with nonnormal random effect distribution, *Computational Statistics and Data Analysis* **56**: 1499–1510.

Fahemeir, L. and Tutz, G. (2001). *Multivariate statistical modelling based on Generalized Linear Models (2nd ed.)*, Springer Series in Statistics, Springer-Verlag, New York.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995). *Bayesian Data Analysis*, Texts in Statistical Science. London: Chapman and Hall.

Gu, M. and Kong, F. (1998). A stochastic approximation algorithm with markov chain monte-carlo method for incomplete data estimation problems, *Proc. Natl. Acad. Sci. USA* **95**: 7270–7274.

Guimarães, P. (2005). A simple approach to fit the beta-binomial model, *The Stata Journal* **5**(3): 385–394.

Gumedze, F. N., Welham, S. J., Gogel, B. J. and Thompson, R. (2010). A variance shift model for detection of outliers in the linear mixed model, *Computational Statistics and Data Analysis* **54**: 2128–2144.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model, *Biometrics* **31**: 423–447.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on gibbs sampling in hierarchical linear mixed models, *Journal of the American Statistical Association* **91**(436): pp. 1461–1473.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects; salamander mating revisited, *Biometrics* **48**(2): pp. 631–644.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C. and Verbeke, G. (2012). Modeling overdispersed longitudinal binary data using a combined beta and normal random-effects model, *preprint* .

Kruskal, W. and Wallis, A. (1952). Use of ranks in one-criterion variance, *Journal of the American Statistical Association* **47**(260): 583–621.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data, *Biometrics* **38**: 963–974.

Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society* **58**: 619–678.

Lee, Y. and Nelder, J. (2001). Hierarchical generalised linear models : A synthesis of generalised linear models, random-effect models and structured dispersions, *Biometrika* **88**: 987–1006.

Lee, Y. and Nelder, J. A. (2000). Two ways of modelling overdispersion in non-normal data, *Journal of the Royal Statistical Society* **49**(4): 591–598.

Lee, Y., Nelder, J. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*, Chapman and Hall/CRC.

Lesnoff, M. and Lancelot, R. (2012). *aod: Analysis of Overdispersed Data*. R package version 1.3.
URL: *http://cran.r-project.org/package=aod*

Lesnoff, M., Laval, G., Bonnet, P., Abdicho, S., Workalemahu, A., Kifle, D., Peyraud, A., Lancelot, R. and Thiaucourt, F. (2004). Within-herd spread of contagious bovine pleuropneumonia in ethiopian highlands, *Preventive Veterinary Medicine* **64**: 2740.

Liang, K. Y. and Zeger, S. (1986). Longitudinal data analysis using generalised linear models, *Biometrika* **73**: 13–22.

Lindsey, J. (1996). *Parametric statistical inference*, Clarendon Press.

Mansoor, N., Scriba, T. J., de Kock, M., Tameris, M., Abel, B., Keyser, A., Little, F., Soares, A., Gelderbloem, S., Mlenjeni, S., Denation, L., Hawkridge, A., Boom, W. H., Kaplan, G., Hussey, G. D. and Hanekom, W. A. (2009). Hiv-1 infection in infants severely impairs the immune response induced by bacille calmette-gurin vaccine, *The Journal of infectious diseases* **199**(7): 982–990.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*, Chapman and Hall, London.

McCulloch, C. (1994). Maximum likelihood variance components estimation in binary data, *Journal of the American Statistical Association* **89**: 330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association* **92**(437): pp. 162–170.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear, and Mixed Models*, New York, John Wiley.

Miller, R. G. (1997). *Beyond ANOVA: Basics of Applied Statistics*, Chapman and Hall, Boca Raton, FL.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer-Verlag.

Molenberghs, G., Verbeke, G., Demétrio, C. and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects, *Statistical Science* **25**: 325–347.

Myers, P., Montgomery, D. and Vining, G. (2002). *Generalized linear models with applications in engineering and the sciences*, New York: John Wiley and Sons.

Nelder, J. and Pregibon, D. (1987). An extended quasi-likelihood function, *Biometrika* **74**(2): 221–232.

Ouwens, M. J. N., Tan, F. E. S. and Berger, M. P. F. (2001). Local influence to detect influential data structures for generalized linear mixed model, *Biometrics* **57**: 1166–1172.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**: 545–554.

Payne, R., Harding, S., Murray, D., Soutar, D., Baird, D., Glaser, A., Welham, S., Gilmour, A., Thompson, R. and Webster, R. (2011). *GenStat Release 14 Reference Manual, Part 3 Procedure Library PL22*, 5 The Waterhouse, Waterhouse Street, Hemel Hempstead, Hertfordshire HP1 1ES, UK.

Pierce, D. and Schafer, D. (1986). Residuals in generalized linear models, *Journal of the American Statistical Association* **81**: 977–986.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2012). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-104.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*, Springer.

Pregibon, D. (1981). Logistic regression diagnostics, *Annals of Statistics* **9**: 705–724.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org*

Schall, R. (1991). Estimation in generalized linear models with random effects, *Biometrika* **78**(4): pp. 719–727.

Self, S. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions, *Journal of the American Statistical Association* **82**: 605–610.

Soares, A. P., Scriba, T. J., Joseph, S., Harbacheuski, R., Murray, R. A., Gelderbloem, S. J., Hawkridge, A., Hussey, G. D., Maecker, H., Kaplan, G. and Hanekom, W. A. (2008). Bacillus calmette-gurin vaccination of human newborns induces t cells with complex cytokine and phenotypic profiles, *The Journal of Immunology* **180**(5): 3569–3577.

StataCorp (2009). Stata statistical software, *Release 11* .

Stram, D. and Lee, J. (1995). Correction to variance components testing in longitudinal mixed effects model, *Biometrics* **51**: 1196.

Thall, P. and Vail, S. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics* **46**: 657–671.

Thompson, R. (1985). A note on restricted maximum likelihood estimation with an alternative outlier model, *Journal of the Royal Statistical Society* **47**: 53–55.

Tuerlinckx, F., Rijmen, F., Verbreke, G. and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review, *British Journal of Mathematical and Statistical Psychology* **59**: 225–255.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method, *Biometrika* **61**(3): pp. 439–447.

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

Williams, D. (1987). Generalized linear model diagonstics using the deviance and single case deletions, *Applied Statistics* **36**(2): 181–191.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models, *Applied Statistics* **31**: 144–148.

Xiang, L., Tse, S. and Lee, A. H. (2002). Influence diagnostics for generalized linear mixed models: applications to clustered data, *Computational Statistics and Data Analysis* **40**: 759–774.

Xu, L., Lee, S. Y. and Poon, W. Y. (2006). Deletion measures for generalized linear mixed effects models, *Computational Statistics and Data Analysis* **51**: 1131–1146.

Yau1, K., Wang, K. and Lee, A. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros, *Biometrical Journal* **45**: 437–452.

Zewotir, T. (2007). Infinitesimal model perturbation influence in the linear mixed model, *South African Statistical Journal* **41**: 105–126.

Zhu, H. and Lee, S. (2003). Local influence for generalized linear mixed models, *The Canadian Journal of Statistics* **31**(3): 293–309.

Chapter 10

Appendix

10.1 VSOM code for GENSTAT and R

10.1.1 Cytokine linear mixed model R code

```
library(MASS)
library(lme4)
a<-read.table("C:/Users/Officeworks/Desktop/thesis_drop
/abs_dropz.csv",header=T,sep=",")
#reads in the data
a
n<-length(a[,5]) #number of observations
q<-61 #number of subjects
cd4<-log(a[,5])
cd4

hist(a[,5], main = " ", xlab = "cd4_ifngpil2ptnfp counts",ylim=c(0,100))
hist(cd4, main = " ", xlab = "log(cd4_ifngpil2ptnfp) counts",xlim=c(0,10), ylim=c(0,60))
subject<-a[,1]
subject<-factor(subject) # creates a factor
group<-factor(a[,3])
a[,3]<-group
a[,1]<-subject # replaces the numeric with a factor

b<-lmer(cd4~group + timepoint+ (1|subject),data= a)# fits model
summary(b)
b
attributes(b)
# allows you to see the X and Z matrices as well as residuals and fitted values

m<-VarCorr(b) # gives the variances
m
varu<-as.numeric(m)# subject random effect variance
sigma<-attributes(m)$sc #gives standard deviation of the error term

X<-attributes(b)$X
X<-as.numeric(X) # the data has to be converted to numerical form
X<-matrix(X,n) # then the vector produced must be made into a matrix
X # gives the design matrix
ll<-attributes(b)$Z
Z<-t(ll)
Z
```

```
Z<-as.numeric(Z)# the data has to be converted to numerical form
Z<-matrix(Z,n) # makes a numeric Z matrix
Z # gives the designx for random effects


# Estimating the LRT values for each observation
b<-lmer(cd4~group + timepoint+ (1|subject),data= a)# fits model
b
dev<-attributes(b)$deviance
dev<-as.numeric(dev)
reml<-dev[2]
reml # gives the reml of the null model
m<-VarCorr(b) # gives the variances
sigma<-attributes(m)$sc #gives standard deviation of residuals

lrtalt<-di<-omegaV<-newvariance<-rep(0,times=n)

for( i in 1:n)
{

di[i]<-1 # ith position of the vector becomes 1
vsom1<-lmer(cd4~group + timepoint+ (1|subject)+ (0+di|di),data =a) #VSOM model
devI<-attributes(vsom1)$deviance
devI<-as.numeric(devI)
remlI<-devI[2]
 # gives the restricted maximum likelihood value of the model with ith random effect
mI<-VarCorr(vsom1) # gives the variances
sigmanew<-attributes(mI)$sc #gives standard deviation of residuals
newvariance[i]<-sigmanew^2
vc <-VarCorr( vsom1 ) #gives variances of random effects
vc1<-as.numeric(vc)
vardi<-vc1[2] # gives variance of di
omegaV[i]<-vardi/(newvariance[i]) #omega ratio
lrtalt[i]<-(reml-remlI) # LRT statistic
di[i]<-0# resetting the d vector to zeroes
}
plot(lrtalt) # LRT statistic
plot(omegaV) # size of omega
plot(newvariance) #residual variance

# Bootstrapping

sim<-2000
u<-rep(0,times=q)
e<-rep(0,times=n)
di<-rep(0,times=n)
lrtalt1<-rep(0,times= sim*n)
lrtalt1<-matrix(lrtalt1,n)
lrtalt2<-rep(0,times=n)

tau<-attributes(b)$fixef
tau<-as.numeric(tau)
tau # beta values of the null model

for(j in 1:sim)
{
for(i in 1:n)
{e[i]<-rnorm(1,0,sigma)} # generates error terms

for(i in 1:q)
{u[i]<-rnorm(1,0,varu^0.5)} # generates subject effects

cd41<-(X%*%tau)+ ( Z%*%u)+ e # generating dummy data set

nb<-lmer(cd41~group + timepoint +(1|subject),data= a)# fits model
```

```
dev1<-attributes(nb)$deviance
dev1<-as.numeric(dev1)
reml<-dev1[2]

for( i in 1:n)
{
  di[i]<-1 # ith position of the vector becomes 1
  vsomnew<-lmer(cd41~group + timepoint+ (1|subject) + (0+di|di),data =a)
  devI<-attributes(vsomnew)$deviance
  devI<-as.numeric(devI)
  remlI<-devI[2]
# gives the residual maximum likelihood value of the model with ith random effect
  lrtalt2[i]<-(reml-remlI)
  di[i]<-0# resetting the d vector to zeroes
}

lrtalt1[,j]<-lrtalt2
}

bbb3<-rep(0,times=sim*n)
bbb3<-matrix(bbb3,n,sim) # creates a dummy matrix


hh <- transform(data.frame(lrtalt1)) # create a dataframe, this is a method to sort the data

for( i in 1:sim)
{
bbb3[,i]<-sort(hh[,i],decreasing=T) # fills the dummy matrix with sorted columns for lrtalt
 }

cutterlrtalt<-rep(0,times =n)
for(i in 1:n)
{
cutterlrtalt[i]<-quantile(bbb3[i,],0.95) # creates a quantile for all values of lrtalt
}
cutterlrtalt

write.csv(cutterlrtalt, "C:/Users/Officeworks/Desktop/thesis_drop/cutterlrtalt_indy.csv")
write.csv(lrtalt, "C:/Users/Officeworks/Desktop/thesis_drop/lrtalt_indy.csv")



lrtalt
plot(lrtalt)
abline(cutterlrtalt[4],0,col="blue") # 95% CI of 4th highest lrt value
abline(cutterlrtalt[5],0,col="red") # 95% CI of 5th highest lrt value
abline(cutterlrtalt[6],0,col="black") # 95% CI of 6th highest lrt value

plot(lrtalt)#actual likelihood

##Plots#############

boot_out<-c(40,48,11,143,29,47)
boot_out_lik<-c(9.637,6.423,6.1387,6.117788553,5.98657211,5.9312946)


par(mar=c(5, 5, 0, 2) + 0.1)
par(mfrow=c(3,1))
#plot of omega values
plot(omegaV, main = "",ylim=c(min(omegaV)-0.01,(max(omegaV)+.01)),xlab="",
ylab = (expression(paste(omega[i]))),font.axis=6,cex=1.05,cex.lab=2)
mtext(paste("(a)"), side=3,at= -17 ,line = -1,font.lab=6, cex.lab=1 )

#plot of residual variances
plot(newvariance, main ="",xlab="", ylab = (expression(paste(sigma^2))),
```

```
ylim=c((min(newvariance)-0.0005),(max(newvariance)
+0.0005)),font.axis=6,cex=1.05,cex.lab=2)
mtext(paste("(b)"), side=3,at=-17,font.lab=6,cex.lab=1)

#plot of likelihood statistics
plot(lrtalt, main = "", xlab="Observations",ylab=  (expression(paste(  LRT[i]))),ylim=c(0,(max(lrtalt)+1)),font.axis=6,cex=1.05,cex.lab=2)
mtext(paste("(c)"), side=3,at=-17,font.lab=6,cex.lab=1)
text(boot_out,boot_out_lik,boot_out, pos =  2)
abline(cutterlrtalt[4],0,col="red") # 95% CI of the highest lrt value
abline(cutterlrtalt[5],0,col="green") # 95% CI of 4th highest lrt value
abline(cutterlrtalt[6],0,col="blue") # 95% CI of 10th highest lrt value



# VSOM for outlying observations

d40<-rep(0,times=n)
d48<-rep(0,times=n)
d11<-rep(0,times=n)
d143<-rep(0,times=n)
d29<-rep(0,times=n)
d47<-rep(0,times=n)

d40[40]<-1
d48[48]<-1
d11[11]<-1
d143[143]<-1
d29[29]<-1
d47[47]<-1
bbb2<-lmer(cd4~group + timepoint+ (1|subject)+ (0+d11|d11)+ (0+d29|d29)+(0+d40|d40) +(0+d47|d47)+(0+d48|d48)+(0+d143|d143),data= a)# fits model
summary(bbb2)


#case deletion

aaa<-read.table("C:/Users/Officeworks/Desktop/thesis_drop/normal/abs_dropz_drop.csv",
header=T,sep=",")#reads in the data
aaa
cd4new<-log(aaa[,5])
cd4new

subject<-aaa[,1]
subject<-factor(subject) # creates a factor
group<-factor(aaa[,3])
aaa[,3]<-group
aaa[,1]<-subject # replaces the numeric with a factor
aaa

bbb<-lmer(cd4new~group + timepoint+ (1|subject),data= aaa)# fits model
summary(bbb)
```

10.1.1.1 R code for subjects

```
library(MASS)
library(lme4)
a<-read.table("C:/Users/Officeworks/Desktop/thesis_drop
/abs_dropz.csv",header=T,sep=",")
#reads in the data
a
n<-length(a[,5]) #number of observations
q<-61 #number of subjects
cd4<-log(a[,5])
cd4
```

```
hist(a[,5], main = " ", xlab = "cd4_ifngpil2ptnfp counts",ylim=c(0,100))
hist(cd4, main = " ", xlab = "log(cd4_ifngpil2ptnfp) counts",xlim=c(0,10), ylim=c(0,60))
subject<-a[,1]
subject<-factor(subject) # creates a factor
group<-factor(a[,3])
a[,3]<-group
a[,1]<-subject # replaces the numeric with a factor

b<-lmer(cd4~group + timepoint+ (1|subject),data= a)# fits model
summary(b)
b
attributes(b)
# allows you to see the X and Z matrices as well as residuals and fitted values

m<-VarCorr(b) # gives the variances
m
varu<-as.numeric(m)# subject random effect variance
sigma<-attributes(m)$sc #gives standard deviation of the error term

X<-attributes(b)$X
X<-as.numeric(X) # the data has to be converted to numerical form
X<-matrix(X,n) # then the vector produced must be made into a matrix
X # gives the design matrix
ll<-attributes(b)$Z
Z<-t(ll)
Z
Z<-as.numeric(Z)# the data has to be converted to numerical form
Z<-matrix(Z,n) # makes a numeric Z matrix
Z # gives the designx for random effects


# Estimating the LRT values for each subject
b<-lmer(cd4~group + timepoint+ (1|subject),data= a)# fits model
b
dev<-attributes(b)$deviance
dev<-as.numeric(dev)
reml<-dev[2]
reml # gives the reml of the null model
m<-VarCorr(b) # gives the variances
sigma<-attributes(m)$sc #gives standard deviation of residuals

lrtalt<-di<-omegaV<-newvariance<-rep(0,times=q)

for( i in 1:q)
{

di<-Z[,i] # ith position of the vector becomes 1
vsom1<-lmer(cd4~group + timepoint+ (1|subject)+ (0+di|di),data =a) #VSOM model
devI<-attributes(vsom1)$deviance
devI<-as.numeric(devI)
remlI<-devI[2]
# gives the restricted maximum likelihood value of the model with ith random effect
mI<-VarCorr(vsom1) # gives the variances
sigmanew<-attributes(mI)$sc #gives standard deviation of residuals
newvariance[i]<-sigmanew^2
vc <-VarCorr( vsom1 ) #gives variances of random effects
vc1<-as.numeric(vc)
vardi<-vc1[2] # gives variance of di
omegaV[i]<-vardi/(newvariance[i]) #omega ratio
lrtalt[i]<-(reml-remlI) # LRT statistic

}
lrtalt
plot(lrtalt) # LRT statistic
```

```
plot(omegaV) # size of omega
plot(newvariance) #residual variance

# Bootstrapping

sim<-2000
u<-rep(0,times=q)
e<-rep(0,times=n)
di<-rep(0,times=n)
lrtalt11<-rep(0,times= sim*q)
lrtalt11<-matrix(lrtalt11,q)
lrtalt21<-rep(0,times=q)

tau<-attributes(b)$fixef
tau<-as.numeric(tau)
tau # beta values of the null model

for(j in 1:sim)
{
for(i in 1:n)
{e[i]<-rnorm(1,0,sigma)} # generates error terms

for(i in 1:q)
{u[i]<-rnorm(1,0,varu^0.5)} # generates subject effects

cd41<-(X%*%tau)+ ( Z%*%u)+ e # generating dummy data set

nb<-lmer(cd41~group + timepoint +(1|subject),data= a)# fits model
dev1<-attributes(nb)$deviance
dev1<-as.numeric(dev1)
reml<-dev1[2]

for( i in 1:q)
{
  di<-Z[,i] # ith position of the vector becomes 1
  vsomnew<-lmer(cd41~group + timepoint+ (1|subject) + (0+di|di),data =a)
  devI<-attributes(vsomnew)$deviance
  devI<-as.numeric(devI)
  remlI<-devI[2]
# gives the residual maximum likelihood value of the model with ith random effect
  lrtalt21[i]<-(reml-remlI)
}

lrtalt11[,j]<-lrtalt21
}

bbb3<-rep(0,times=sim*q)
bbb3<-matrix(bbb3,q,sim) # creates a dummy matrix


hh <- transform(data.frame(lrtalt11)) # create a dataframe, this is a method to sort the data

for( i in 1:sim)
{
bbb3[,i]<-sort(hh[,i],decreasing=T) # fills the dummy matrix with sorted columns for lrtalt
 }

cutterlrtalt<-rep(0,times =q)
for(i in 1:q)
{
cutterlrtalt[i]<-quantile(bbb3[i,],0.95) # creates a quantile for all values of lrtalt
}
cutterlrtalt

write.csv(cutterlrtalt, "C:/Users/Officeworks/Desktop/thesis_drop/cutterlrtalt_sub.csv")
```

```
lrtalt
plot(lrtalt)
abline(cutterlrtalt[2],0,col="blue") # 95% CI of the highest lrt value
abline(cutterlrtalt[3],0,col="red") # 95% CI of 4th highest lrt value
abline(cutterlrtalt[4],0,col="black") # 95% CI of 10th highest lrt value

plot(lrtalt)#actual likelihood

##Plots#############

boot_out<-c(12,38)
boot_out_lik<-c(5.986572, 7.035780)
par(mar=c(5, 5, 0, 2) + 0.1)
par(mfrow=c(3,1))
#plot of omega values
plot(omegaV, main = "",ylim=c(min(omegaV)-0.01,(max(omegaV)+.01)),xlab="",
 ylab = (expression(paste(omega[j]))),font.axis=6,cex=1.05,cex.lab=2)
mtext(paste("(a)"), side=3,at= -5 ,line = -1,font.lab=6, cex.lab=1 )

#plot of residual variances
plot(newvariance, main ="",xlab="", ylab = (expression(paste(sigma^2))),ylim=c((min(newvariance)-0.0005),(max(newvariance)+0.0005))
,font.axis=6,cex=1.05,cex.lab=2)
mtext(paste("(b)"), side=3,at=-5,font.lab=6,cex.lab=1)

#plot of likelihood statistics
plot(lrtalt, main = "", xlab="Observations",ylab=  (expression(paste(  LRT[j]))),ylim=c(0,(max(lrtalt)+1)),font.axis=6,cex=1.05,cex.lab=2)
mtext(paste("(c)"), side=3,at=-5,font.lab=6,cex.lab=1)
text(boot_out,boot_out_lik,boot_out, pos =  2)
abline(cutterlrtalt[2],0,col="red") # 95% CI of the highest lrt value
abline(cutterlrtalt[3],0,col="green") # 95% CI of 4th highest lrt value
abline(cutterlrtalt[4],0,col="blue") # 95% CI of 10th highest lrt value

##########VSOM for individuals#############

d40<-rep(0,times=n)
d48<-rep(0,times=n)
d11<-rep(0,times=n)
d143<-rep(0,times=n)
d29<-rep(0,times=n)
d47<-rep(0,times=n)

d40[40]<-1
d48[48]<-1
d11[11]<-1
d143[143]<-1
d29[29]<-1
d47[47]<-1
bbb2<-lmer(cd4~group + timepoint+ (1|subject)+ (0+d11|d11)+ (0+d29|d29)+(0+d40|d40) +(0+d47|d47)+(0+d48|d48)+(0+d143|d143),data= a)# fits model
summary(bbb2)




######VSOM subject 12 and 40#####
ds12<-ds38<-rep(0,times=n)
ds12<-Z[,12]
ds38<-Z[,38]

vsom2<-lmer(cd4~group + timepoint+ (1|subject) + (0+ds12|ds12)
+(0+ds38|ds38),data =a)
vsom2

#########VSOM combined##################
vsom3<-lmer(cd4~group + timepoint+ (1|subject) + (0+ds12|ds12)
```

```
+(0+ds38|ds38)+(0+d11|d11)+(0+d40|d40)
+(0+d47|d47)+(0+d48|d48)+(0+d143|d143),data= a)# fits model,data =a)
vsom3

#######case deletion###################

a2<-read.table("C:/Users/Officeworks/Desktop/thesis_drop/normal/abs_dropz_drop.csv"
,header=T,sep=",")#reads in the data
a2
cd4n<-log(a2[,5])
cd4n

subject<-a2[,1]
subject<-factor(subject) # creates a factor
group<-factor(a2[,3])
a2[,3]<-group
a2[,1]<-subject # replaces the numeric with a factor

group
b2<-lmer(cd4n~group + timepoint+ (1|subject),data= a2)# fits model
summary(b2)
b
```

10.1.2 Cytokine Count VSOM GENSTAT code

```
SPLOAD 'C:/Users/Officeworks/Desktop/thesis_drop/count/count.gsh'
scalar n; value=182
scalar a; value=61
variate[;1...#n]obs
variate[;1...#a]obs1
print obs,subject,group,timepoint,cd4

" Poisson gamma HGLM  "
HGRANDOMMODEL [DIST=gamma; LINK=log] subject
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=1;] group+timepoint
HGANALYSE    cd4

" Quassi Poisson gamma HGLM  "
HGRANDOMMODEL [DIST=gamma; LINK=log] subject
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
HGANALYSE    cd4

\Quassi- Poisson null model
HGRANDOMMODEL [DIST=gamma; LINK=log] subject
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
HGANALYSE    cd4
HGKEEP [modeltype=mean;rmethod=simple] residuals=resm;fittedvalues=fitm;estimates=meanest;\
likelihoodstat=dev11
 HGKEEP [modeltype=dispersion] estimates=disp
print dev11
print dev11$[4]
print disp$[1]

HGPLOT fitted,normal,histogram,absresidual
HGPLOT [random=subject]fitted,normal,histogram,absresidual
"Fit a glm + random effect for each observation, keep var. components and obstain
likelihood ratio tests using HGLM"
VARIATE [NVALUES=#n]lik,d,pval
VARIATE [NVALUES=#n]sig2,omega
scal mis;val=(*)
```

```
FOR [NTIMES=#n;INDEX=k]
Calc d=mis
calc d$[k]=1
 GROUPS [PRINT=summary; LMETHOD=give] d; FACTOR=d1
HGRANDOMMODEL [DIST=gamma; LINK=log] subject+d1
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
HGANALYSE    [MAXCYCLE=50; EXIT= check] cd4

 IF check  "retry with adjusted Aitkin extrapolation"
   HGANALYSE   [MAXCYCLE=50; EMETHOD=adjusted; EXIT=check] cd4
 ENDIF
 IF check  "retry with no Aitkin extrapolation"
   HGANALYSE   [MAXCYCLE=50; EMETHOD=*; EXIT=check] cd4
 ENDIF
 IF check  "fit without di"
HGRANDOMMODEL [DIST=gamma; LINK=log] subject
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
HGANALYSE     [MAXCYCLE=50; EMETHOD=*; EXIT =check]cd4
   EXIT [CONTROL=for; REPEAT=yes] check
   HGKEEP [modeltype=dispersion] estimates=dispest1
   HGKEEP [modeltype=mean;rmethod=simple] residuals=resm1;
fittedvalues=fitm1;estimates=meanest1;\
   likelihoodstat=dev12

   calc lik$[k]= dev11$[4]-dev12$[4]
   Calc omega$[k]=0
   Calc sig2$[k]=exp(dispest1$[1])
 ELSE

 HGKEEP [modeltype=mean;rmethod=simple] residuals=resm1;
fittedvalues=fitm1;estimates=meanest1;\
  likelihoodstat=dev12
HGKEEP [modeltype=dispersion] estimates=dispest1
 calc lik$[k]= dev11$[4]-dev12$[4]
 Calc omega$[k]= exp(dispest1$[3])
Calc sig2$[k]=exp(dispest1$[1])
"Calc p-values for lik. "
calc pval$[k]=1-(0.68+0.32*(CLCHISQUARE(lik$[k];1;0)))
endif
Endfor
print obs,lik,omega,sig2,pval

"Individual VSOM"
VARIATE [NVALUES=#n]dl11,dl147,dl124,dl125
calc dl11$[11]=1
calc dl147$[47]=1
calc dl124$[124]=1
calc dl125$[125]=1
\print dl110,dl111,dl133,dl11
GROUPS [PRINT=summary; LMETHOD=give] dl11,dl147,dl124,dl125;
FACTOR=d11,d47,d124,d125
HGRANDOMMODEL [DIST=gamma; LINK=log] subject+d11+d47+d124+d125
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
HGANALYSE    cd4

"Fitting for just the subjects"
"Fit LMM using REML to get design matrices X and Z"
VCOMPONENTS [FIXED=group+timepoint] RANDOM= subject
REML [PRINT=model;\
 MVINCLUDE=*;parameteri=gamma;pterms=subject;Rmethod=final;METHOD=ai]cd4
VKEEP[vest=v1;vare=v2;dev=dev1]
print v1;dec=7
& dev11
VKEEP term=group+timepoint;DESIGNMATRIX=x;effe=t1
```

```
VKEEP term=subject;DESIGNMATRIX=z;effe=t2
print z
calc z1 =z
 scal mis;val=(*)
   print mis
MATRIX[ROWS=#n;COL=#a]z2
calc z2 = mis
print z2

 FOR [NTIMES=#n;INDEX=k]
      FOR [NTIMES=#a;INDEX=m]
          If z$[k;m] == 1
             calc z2$[k;m] = z$[k;m]
          endIf
      EndFor
EndFor
print z2

\ my attempt at fitting patientid
VARIATE [NVALUES=#a]omega1,sig21,lik1,pval1


FOR [NTIMES=#a;INDEX=k]

          VARIATE [NVALUES=#n]d2
 calc d2 = z2$[*;k] \for the groups and remove calc d

 GROUPS [PRINT=summary; LMETHOD=give] d2; FACTOR=d11
HGRANDOMMODEL [DIST=gamma; LINK=log] subject+d11
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
HGANALYSE     [MAXCYCLE=50; EXIT=check] cd4

 IF check  "retry with adjusted Aitkin extrapolation"
   HGANALYSE   [MAXCYCLE=50; EMETHOD=adjusted; EXIT=check] cd4
 ENDIF
 IF check  "retry with no Aitkin extrapolation"
   HGANALYSE   [MAXCYCLE=50; EMETHOD=*; EXIT=check] cd4
 ENDIF
 IF check  "fit without di"
HGRANDOMMODEL [DIST=gamma; LINK=log] subject
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
   HGANALYSE   [MAXCYCLE=50; EMETHOD=*; EXIT=check] cd4
   EXIT [CONTROL=for; REPEAT=yes] check
   HGKEEP [modeltype=dispersion] estimates=dispest1
   HGKEEP [modeltype=mean;rmethod=simple] residuals=resm1;
fittedvalues=fitm1;estimates=meanest1;\
   likelihoodstat=dev12

   calc lik1$[k]= dev11$[4]-dev12$[4]
   Calc omega1$[k]=0
   Calc sig21$[k]=exp(dispest1$[1])
 ELSE

 HGKEEP [modeltype=mean;rmethod=simple] residuals=resm1;
fittedvalues=fitm1;estimates=meanest1;\
  likelihoodstat=dev12

HGKEEP [modeltype=dispersion] estimates=dispest1

 calc lik1$[k]= dev11$[4]-dev12$[4]
 Calc omega1$[k]= exp(dispest1$[3])
Calc sig21$[k]=exp(dispest1$[1])
"Calc p-values for lik. "
calc pval1$[k]=1-(0.68+0.32*(CLCHISQUARE(lik1$[k];1;0)))
endif
```

```
Endfor
print obs1,lik1,omega1,sig21,pval1

"VSOM for outlying subjects"
variate [nvalues = #n] zs12,zs38, zs32

calc zs12 =z2$[*;12]
calc zs32 = z2$[*;32]
calc zs38 = z2$[*;38]

 GROUPS [PRINT=summary; LMETHOD=give] zs12,zs32,zs38; FACTOR=z12,z32,z38
HGRANDOMMODEL [DIST=gamma; LINK=log] subject+z12+z32+z38
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
HGANALYSE     [MAXCYCLE=50; EXIT=check] cd4

"VSOM combing subjects and observations"

HGRANDOMMODEL [DIST=gamma; LINK=log] subject+z12+z32+z38+d11+d47+d124+d125
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
HGANALYSE     [MAXCYCLE=50; EXIT=check] cd4

"Drop outliers"
SPLOAD 'C:/Users/Officeworks/Desktop/thesis_drop/count/abs_dropz_drop.gsh'
HGRANDOMMODEL [DIST=gamma; LINK=log] subject
HGFIXEDMODEL  [DIST=poisson; LINK=log;DISPERSION=*;] group+timepoint
HGANALYSE     [MAXCYCLE=50; EXIT=check] cd4
```

10.1.3 Updated seed germination dataset VSOM GENSTAT code

```
SPLOAD 'C:/Users/Officeworks/Desktop/binomial/beta_seeds/beta_seeds.gsh'
scalar n1; value=21
variate[;1...#n1]obs
print id,y,n,species2,extract2
 calc y$[3] =81
calc y$[6] =0
print id,y,n,species2,extract2
" beta-binomial HGLM  "
HGRANDOMMODEL [DIST=beta; LINK=logit] id
HGFIXEDMODEL  [DIST=binomial; LINK=logit;DISPERSION=1;] species2+extract2+ species2*extract2
HGANALYSE    y; nbinomial=n
HGKEEP [modeltype=mean;rmethod=simple] residuals=resm;fittedvalues=fitm
;estimates=meanest;\
likelihoodstat=dev11
 HGKEEP [modeltype=dispersion] estimates=disp
print dev11
print dev11$[4]
print disp$[1]

\HGPLOT fitted,normal,histogram,absresidual
\HGPLOT[random=herd] fitted,normal,histogram,absresidual
"Fit a glm + random effect for each observation, keep var. components and obstain
likelihood ratio tests using HGLM"
VARIATE [NVALUES=#n1]lik,d,pval
VARIATE [NVALUES=#n1]sig2,omega
scal mis;val=(*)
\calc init[1...3] = exp(-4,-8,-1)
FOR [NTIMES=#n1;INDEX=k]
Calc d=mis
calc d$[k]=1
 GROUPS [PRINT=summary; LMETHOD=give] d; FACTOR=d1
HGRANDOMMODEL [DIST=beta; LINK=logit] id +d1
```

```
HGFIXEDMODEL  [DIST=binomial; LINK=logit;DISPERSION=1;] species2+extract2+ species2*extract2
HGANALYSE   [MAXCYCLE=50; EXIT= check] y; nbinomial=n


\ IF check  "retry with adjusted Aitkin extrapolation"
\   HGANALYSE   [MAXCYCLE=50; EMETHOD=adjusted; EXIT=check] r; nbinomial=n
\ ENDIF
\ IF check  "retry with no Aitkin extrapolation"
 \  HGANALYSE   [MAXCYCLE=50; EMETHOD=*; EXIT=check] r;nbinomial=n
 \ENDIF
 IF check  "fit without di"
HGRANDOMMODEL [DIST=beta; LINK=logit] id
HGFIXEDMODEL  [DIST=binomial; LINK=logit;DISPERSION=1;] species2+extract2+ species2*extract2
HGANALYSE     [MAXCYCLE=50; EMETHOD=*; EXIT =check]y; nbinomial=n
   EXIT [CONTROL=for; REPEAT=yes] check
   HGKEEP [modeltype=dispersion] estimates=dispest1
   HGKEEP [modeltype=mean;rmethod=simple] residuals=resm1;fittedvalues=fitm1;estimates=meanest1;\
   likelihoodstat=dev12
\print dev12
   \print k, dev12
   calc lik$[k]= dev11$[4]-dev12$[4]
   Calc omega$[k]=0
   Calc sig2$[k]=exp(dispest1$[1]) \subject variance
 ELSE
 HGKEEP [modeltype=mean;rmethod=simple] residuals=resm1;fittedvalues=fitm1;estimates=meanest1;\
  likelihoodstat=dev12
\print dev12
HGKEEP [modeltype=dispersion] estimates=dispest1
\print dispest1
 calc lik$[k]= dev11$[4]-dev12$[4]
 Calc omega$[k]= exp(dispest1$[2])
Calc sig2$[k]=exp(dispest1$[1])
"Calc p-values for lik. "
calc pval$[k]=1-(0.68+0.32*(CLCHISQUARE(lik$[k];1;0)))
endif
Endfor
print obs,lik,omega,sig2,pval

"individual observations VSOM"
VARIATE [NVALUES=#n1]lik,dl3,dl6,pval
VARIATE [NVALUES=#n1]sig2,omega
scal mis;val=(*)
calc dl3$[3]=1
calc dl6$[6]=1
print dl3,dl6
 GROUPS [PRINT=summary; LMETHOD=give] dl3; FACTOR=d3
 GROUPS [PRINT=summary; LMETHOD=give] dl6; FACTOR=d6
HGRANDOMMODEL [DIST=beta; LINK=logit] id +d3+d6
HGFIXEDMODEL  [DIST=binomial; LINK=logit;DISPERSION=1;] species2+extract2+ species2*extract2
HGANALYSE   [MAXCYCLE=50; EXIT= check] y; nbinomial=n


"drop observations"
SPLOAD 'C:/Users/Officeworks/Desktop/binomial/beta_seeds/beta_seeds_drop2.gsh'
HGRANDOMMODEL [DIST=beta; LINK=logit] id
HGFIXEDMODEL  [DIST=binomial; LINK=logit;DISPERSION=1;] species2+extract2+ species2*extract2
HGANALYSE    y; nbinomial=n
```