Modeling Hierarchical Data, Allowing for Overdispersion and Zero Inflation, in Particular Excess Zeros

Wondwosen Kassahun Yimer

Promotor: Prof. dr. Geert Molenberghs

Co-Promotor: Prof. dr. Christel Faes

Acknowledgements

It is a pleasure to thank many people who made this thesis possible. This work would not have been possible without the support and help of them.

First and foremost I would like to express my heartfelt gratitude to my PhD promotors Prof. dr. Geert Molenberghs and Prof. dr. Christel Faes. I am very grateful to Geert for his enthusiasm, his inspiration, his guidance, his encouragement and understanding me. Geert, it has been a pleasure and a great opportunity to work with you. I would also like to express my sincere thanks to Christel, for her great ideas, suggestions and invaluable help. Christel, you have been always willing to help me, I thank you so much.

I would like to thank Prof. dr. Geert Verbeke for his invaluable contributions, comments and suggestions, as co-author. I am also very much thankful to my colleague Thomas Neyens for working together with me from the start, and for the useful discussions that we had together. Thomas, I thank you so much for your friendship and help in many ways.

I am indebted to the VLIR IUC-JU project for funding my PhD. I am very grateful to Prof. dr. Paul Janssen (Modeling Project Leader, North) for his help to get me first in contact with my promoters, and his continuous support in my visit at Hasselt University. I would like to thank Prof. dr. Luc Duchateau (Programme Coordinator, North) and Mr. Kora Tushune (Programme Coordinator, South) for their contribution and continuous support in the course of my PhD. My thanks go to Prof. dr. Ziv Shekedy for his support in various ways, specially during my visits at CENSTAT. I am grateful to all staff members of the program support and administrative units at Jimma, Hasselt and Ghent Universities, specially: Mr. Kasahun Eba, Mr. Boka Assefa, Mr. Jemal Abafita, Mr. Marc Tholen, Ms. Martine Machiels, Ms. Heleke

Bayens, Ms. Mira Jasheri and Ms. Annick Verheylezoon for assisting me in many different ways.

I am indebted to many colleagues at Jimma University for the permission to use several data sets resulting from our joint research or their own. In this line, I would like to thank Prof. Tefera Belachew and other research team members of the Jimma Longitudinal Family Survey of Youth for the kind permission to use the data, and colleagues in the malaria incidence study of under ten years old children around the Gilgel-Gibe hydroelectric dam: Dr. Delenasaw Yewhalaw, Prof. Kifle W/Michael, Mr. Kora Tushune, and Prof. dr. Luc Duchateau. I would also like to thank Mr. Fasil Tessema and Prof. Makonen Assefa for the permission to use the Jimma Infant Growth Data. Support from WHO Special Programme for Research and Training in Tropical Diseases (TDR) and IAP research Network P7/06 of the Belgian Government (Belgian Science Policy) is gratefully acknowledge.

I am indebted to my many colleagues at CENSTAT for providing me different support. I would like to thank Yehenew, Girma, Birhanu, Bedilu, Aklilu and Abel for their friendship and advice they provided.

I would also like to thank my entire extended family members for their unconditional love and supports and encouragement. My beloved mother Terengo H/Mariam, my lovely sisters and brother were particularly supportive.

Finally, and most importantly, I am forever indebted to my family, specially my wife, Meseret W/Cherkos, who have always encouraged me to pursue my studies and taken full responsibility of family issues in the course of my absence. To whom I dedicate this thesis.

Thank you everyone that I could not mention here, who have encouraged and helped me in different ways.

Wondwosen Kassahun Yimer

04 April 2014

Diepenbeek

List of Publications

- **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. *Archives of Public Health*, http://dx.doi.10.1186/0778-7367-70-7.

- **Kassahun, W.**, Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2013). A Joint Model for Hierarchical Continuous and Zero-inflated Overdispersed Count Data. *Journal of Statistical Computation and Simulation*, http://dx.doi.org/10.1080/00949655.2013.829058.

- **Kassahun, W.**, Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2014). A Zero-Inflated Overdispersed and Hirarchical Poisson Model. *Statistical Modelling, Accepted for Publication.*

- **Kassahun, W.**, Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2014). Marginalized Multilevel Hurdle and Zero-Inflated Models for Overdispersed and Correlated Count Data with Excess Zeros, *Submitted for Publication.*

Contents

List of Abbreviations

AED	Anti-Epileptic-Drug
AIC	Akaike Information Criterion
AIDS	Acquired Immuno Deficiency Syndrom
BMI	Body Mass Index
CDC	Center for Disease Control and Prevention
DIC	Deviance Information Criterion
GEE	Generalized Estimating Equations
GLM	Generalized Linear Models
GLMM	Generalized Linear Mixed Models
GPI	Gender Parity Index
HIV	Human Immunodeficiency Virus
H(PNG)$_\ell$	Hurdle Poisson-Normal-Gamma model with logit link
H(PNG)$_p$	Hurdle Poisson-Normal-Gamma model with probit link
IG	Inverse Gamma
IRC	Indoor Resting Collection
JLFSY	Jimma Longitudinal Family Survey of Youth
Kg	Kilogram
MH(PNG)$_\ell$	Marginalized Hurdle Poisson-Normal-Gamma model with logit link

MH(PNG)$_p$	Marginalized Hurdle Poisson-Normal-Gamma model with probit link
MMM	Marginalized Multilevel Model
M(P--)	Marginalized Poisson
M(PN-)	Marginalized Poisson-Normal
M(PNG)	Marginalized Poisson-Normal-Gamma
MZI(PNG)$_\ell$	Marginalized Zero Inflated Poisson-Normal-Gamma model with logit link
MZI(PNG)$_p$	Marginalized Zero Inflated Poisson-Normal-Gamma model with probit link
NLMIXED	Non-Linear Mixed Model
(P--)	Poisson
(P-G)	Poisson-Gamma
(PN-)	Poisson-Normal
(PNG)	Poisson-Normal-Gamma
ZI(P--)	Zero-Inflated Poisson
ZI(P-G)	Zero-Inflated Poisson-Gamma
ZI(PN-)	Zero-Inflated Poisson-Normal
ZI(PNG)	Zero-Inflated Poisson-Normal-Gamma
ZI(PNG)$_\ell$	Zero-Inflated Poisson-Normal-Gamma with logit link
ZI(PNG)$_p$	Zero-Inflated Poisson-Normal-Gamma with probit link

List of Tables

List of Figures

Chapter 1

Introduction

In medical and biomedical areas, binary and binomial, counts, and times to event outcomes are very common. The generalized linear model family (Agresti, 2002; Nelder and Wedderburn, 1972) offers, among others, a suitable modeling framework. When such data are collected longitudinally from a given subject repeatedly overtime, this results in clustering of the observations within subjects.

Suppose that r_{ij} is a longitudinal binary outcome for subject i at the j^{th} time point, such that each subject has n_i measurements. The sum $Y_i = \sum_{j=1}^{n_i} r_{ij}$ follow a binomial distribution. It is well known that, while i.i.d Bernoulli do not contradict the prescribed mean-variance relation, i.i.d. binomial data can exhibit extra variability beyond the binomial model, leading to so-called overdispersion in the latter, in addition to the correlation emanating from the repeated measures nature. In the past, overdispersion and correlation have been handled separately. To deal with overdispersion, the beta-binomial model is a popular and analytically tractable alternative to the binomial model, which accounts for the overdispersion not accommodated in the binomial model, thereby allowing for a better fit to the observed data (Hinde and Demétrio, 1998a; Hinde and Demétrio, 1998b). On the other hand, correlation is accommodated by making use of generalized linear mixed models (Engel and Keen, 1992; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993), which combine the general exponential family models with normally distributed random effects. These are attractive for repeated measurements. Molenberghs *et al.* (2010) formulated a model for correlated and overdispered repeated binary data using Gaussian and beta random effects simultaneously, which they termed the *combined model*.

In longitudinal studies, count data are encountered in a variety of fields, including biological, public health, medical, and social studies. For example, in entomological research, as we focus on in this thesis, mosquito counts are collected repeatedly in time to study the abundance and species composition of the vector over time so as to regulate and monitor the status of the ecosystem and design an appropriate intervention strategy, whenever necessary. However, statistical modeling of such data poses several challenges. This is because repeatedly measured insect counts often exhibit three features: first, correlated observations per subject, which results from the clustering of measurements within subjects; second, the variance exceeds the mean, leading to so-called overdispersion; and third, occurrence of an excessive number of zeros beyond what can be expected based on the commonly used count distributions.

For the researcher or statistician, who wants to model data with these forms of complexity, different options are not always straight-forward to choose from. Indeed, there already exists a number of possibilities, e.g., non-Gaussian clustered data, such as counts, are frequently modeled by making use of generalized linear mixed-effects models, which extend the broad class of generalized linear models by adding a subject-specific random effect, often of a Gaussian type, to capture the correlation between the repeated measurements per subject (Engel and Keen, 1992; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005).

On the other hand, overdispersion is dealt with by including a gamma random effect in the Poisson model, leading to the negative-binomial model (McCullagh and Nelder, 1989). In the *combined model* framework of Molenberghs *et al.* (2010) a Gaussian and a gamma random effect are employed at once to model both correlation and overdispersion.

Also, to account for the excessive proportion of zeros, either the hurdle or zero-inflated model are often used. The former is a two-part conditional model, using a zero mass and a truncated-at-zero count distribution, while the latter is a way of modeling excessive zeros by mixing a discrete point mass and a count distribution. Mullahy (1986) studied the hurdle model for univariate count data. An extension for longitudinal or clustered count data with excessive zeros was considered by Min and Agresti (2005). A separate strand of literature is devoted to zero-inflated model. Lambert (1992) and Greene (1994) studied zero-inflation for cross-sectional count data, and the multi-level extension was the focus of Lee *et al.* (2006). Min and Agresti (2005) and Lee *et al.* (2006) introduced two separate and possibly correlated subject-specific random effects, one in the count and the other in the zero-inflation part.

Many longitudinal studies involve collecting data on more than one outcome from a given subject repeatedly in time. These outcomes include, but are not limited to continuous, count, and binary data. For example, in HIV studies, seropositive patients are monitored until they develop AIDS or die, and their immune system is regularly measured using markers such as the CD4 lymphocyte count, the estimated viral load, or whether viral load is below detectable limits. In the case of the Jimma Infant Growth study, described in the next section, a continuous outcome, such as body weight is measured repeatedly from each infant. At the same time, the health condition of a child was also assessed to see if the child has experienced a specific disease, like diarrhea, whereby the number of days of illness, as a count outcome, was recorded so as to measure the magnitude of the disease burden.

Extensive literature is available on the analysis of each longitudinal outcome separately. For a Gaussian longitudinal response, the linear mixed model is very popular (Laird and Ware, 1982; Verbeke and Molenberghs, 2000). Subject-level random-effects, that are of a Gaussian type, are introduced in such a model to capture the within-subject correlation.

Many applications demand modeling of two or more longitudinal outcomes jointly to get better insight into their joint evolution so as to address certain real world problems. A lot of literature is available on joint modeling of a longitudinal outcome and time to an event (Tsiatis and Davidian, 2004). Horrocks and van den Heuvel (2009) consider the problem of predicting the achievement of successful pregnancy, in a population of women undergoing treatment for infertility, based on longitudinal measurements of adhesiveness. For this purpose, they used a joint model, consisting of a linear mixed-effects sub model for the longitudinal adhesion outcome and a generalized linear sub model for the primary binary endpoint. Molenberghs and Verbeke (2005) discuss a number of techniques that jointly model continuous and discrete outcomes.

Joint modeling of longitudinal continuous and count sequences, the latter possibly overdispersed and zero-inflated, requires to assemble aspects coming from each one of them into one single model. These include the correlation from the continuous, as well as the correlation, overdispersion, and zero-inflation features from the count sequence. The model is relatively complex because it combines various features; nevertheless, it can be implemented in standard software, such as the SAS procedure NLMIXED.

For non-Gaussian outcomes, such as counts, random-effect models provide parameter estimates having a subject-specific interpretation (Molenberghs and Verbeke, 2005). Heagerty (1999) and Heagerty and Zeger (2000) proposed a marginalized multilevel model (MMM), and simultaneously specified a marginal mean and a conditional mean by making use of the so called connector function, yielding marginally interpretable covariate effects. Molenberghs *et al.* (2010) derived marginal expressions of the combined model by integrating the hierarchically specified model over the random effects for a variety of settings, including count data. Such partial marginalization proceeds by integrating first with the overdispersion random effects, leaving the normal random effects untouched. The so resulting marginal means may not provide readily interpretable parameter estimates for covariate effects. Hence, Iddi and Molenberghs (2012) merged the concepts of the combined model of Molenberghs *et al.* (2010) and marginalized multilevel model (MMM) of Heagerty (1999) and proposed a corresponding marginal combined model, so that the resulting estimates have a direct marginal inferences. Lee *et al.* (2011) considered marginalized hurdle model as an extension of Heagerty (1999) for clustered count data with excessive zeros.

In **Chapter 2**, we briefly describe the datasets that have been used in this work. This will be followed by a review of basic terminology, concepts and the standard models for analysis of binary and count data, as outlined in **Chapter 3**.

Clustered binary data is subject to overdispersion in addition to the correlation due to the data hierarchy. The combined modeling approach of Molenberghs *et al.* (2010) for repeated binary data and its implementation in a Bayesian setting, as an alternative estimation technique, is studied in **Chapter 4**. Two longitudinal binary data sets, collected in south western Ethiopia: the Jimma infant growth study, where the child's early growth is studied, and the Jimma longitudinal family survey of youth where the adolescent's school attendance is studied over time, are considered. In addition to the combined model, the commonly used methods for binary and binomial data, such as the simple logistic, which accounts neither for the overdispersion nor the correlation, the beta-binomial model, and the logistic-normal model, which accommodate only for the overdispesion, and correlation, respectively, are also considered for comparison purposes.

In **Chapter 5**, an extension of the zero-inflation modeling framework to deal also with zero-inflation, in addition to overdispersion and correlation is presented and applied. Section 5.4 deals with a simulation study to investigate the importance of accounting for clustering, overdispersion, and a preponderance of zero counts.

In **Chapter 6**, we will employ the combined model idea of Molenberghs *et al.* (2010) and marginalized multilevel model Heagerty (1999) with concepts of hurdle or zero-inflated models, and present a unified marginalized hurdle combined model as well as a marginalized zero-inflated combined model, as two alternative modeling strategies for overdispersed and correlated count data with excessive zeros. The former was also studied by Lee *et al.* (2011), where the logit link function was used for the zero-inflation. We considered both logit and probit link functions, whereby the latter leads to closed-form expressions. In addition, instead of using only one of the logit or the probit, we make use of the connection between them (Griswold and Zeger, 2004), and specified a logit link for the marginal model, and a probit for the conditional model, so that the odds ratio interpretation is still retained, while taking computational advantage of the probit link.

Joint modeling of longitudinal continuous and count sequences, the latter possibly overdispersed and zero-inflated, requires to assemble aspects coming from each one of them in one single model. These include the correlation from the continuous, as well as the correlation, overdispersion, and zero-inflation features from the count sequence, which will be the topic of **Chapter 7**. In the case of the Jimma Infant Growth study, described in the next section, a continuous outcome, such as body weight is measured repeatedly from each infant is correlated. At the same time, the health condition of a child was also assessed to see if the child has experienced a specific disease, like diarrhea, whereby the number of days of illness, as a count outcome, were recorded so as to measure the maginitude of the disease burden. These two outcomes are jointly modeled and studied. Further, a simulation study is also reported in Section 7.5.

Finally, discussions and concluding remarks are given in **Chapter 8**.

Chapter 2

Motivating Examples

In this chapter, the data sets that will be used as key examples throughout this thesis are introduced. In Section 2.1, we introduce the Jimma Infant Growth study, conducted to investigate early growth characteristics of children, to establish risk factors affecting infant survival, and to study socio-economic, maternal, and infant-rearing factors that contribute most to the child's early survival. A longitudinal entomological study that aims to investigate the abundance and distribution of An. mosquito, around a hydroelectric dam, is introduced in Section 2.2. Section 2.3 introduces the Jimma Longitudinal Family Survey of Youth, where school attendance, involvement in work, sexual behaviour of adolescents, among others are studied. The epileptic data set, obtained from a randomized, double-blind, parallel group multi-center study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's is introduced in Section 2.4.

2.1 The Jimma Infant Growth Study

The Jimma Infant Survival Differential Longitudinal Growth Study is an Ethiopian study, set up to establish risk factors affecting infant survival and to investigate socio-economic, maternal, and infant-rearing factors that contribute most to the child's early survival. Children born in Jimma, Keffa and Illubabor, located in Southwestern Ethiopia were examined for their first year growth characteristics. At baseline, there were a total of 7969 infants enrolled in the study, whereby 4317, 1494, and 2158 were from rural, urban, and semi-urban areas, respectively. The children were followed-up every two months, until the age of one year. Of special interest in this thesis is the

Table 2.1: *Jimma Infant Growth Study. Percentage of overweight male and female infants by place of residence for each of the seven follow-up times.*

	rural		urban		semi-urban	
Time	female	male	female	male	female	male
0	11.5	12.2	16.5	14.5	20.3	21.5
2	12.1	12.7	13.4	13.5	20.6	22.4
4	12.1	12.4	12.7	16.4	22.5	20.2
6	13.4	12.3	13.8	14.9	18.3	21.0
8	12.7	11.8	14.9	19.5	20.2	23.1
10	13.4	11.4	14.9	14.9	19.5	22.6
12	13.8	14.1	16.9	16.0	17.6	18.2

risk factor for overweight in children. Overweight, among infants, is associated with various risk factors.

It is of particular interest to identify these risk factors in early life through weight and height measurements, which helps in prevention and treatment of overweight and obesity to reduce incidence of several adulthood diseases (Freedman *et al.*, 1999). This outcome is defined by dichotomization of the Body Mass Index (BMI), with a BMI over the 85th percentile for his or her age referring to overweight. The 85th percentile for age- and sex-specific BMI classification of overweight is used based on Center for Disease Control (CDC) recommendation (Mei *et al.*, 2002). The question of interest is whether the percentage of overweight infants changes over time, and whether the evolution differs for gender, place of residence (rural, urban and semi-urban), as well as breast feeding behavior. Table 2.1 gives a summary of the percentage of overweight infants as a function of gender, location and follow-up time (age). The second question of interest in the survey is to assess the diarrheal disease burden. It is investigated whether the number of days of diarrheal illness in the two months period prior to each visit, changes over time (i.e., age), whether the evolution differs for gender (male or female), place of residence (urban or rural), medical care (medical help given or not) and breast feeding behavior (breast or artificial feeding). Of the total 49,000 observations, only about 8,000 (i.e., roughly 85%), observations are non-zero, indicating that there is a non-negligible dominance of zero counts (Table 2.2).

Table 2.2: *Jimma Infant Growth Study. The mean number of days of illness and standard deviation at each of the seven follow-up times.*

Time	Mean	Std. Dev.
0	0.01	0.19
2	0.91	4.24
4	1.28	4.62
6	1.56	4.87
8	2.14	5.93
10	2.63	6.66
12	2.67	6.95

Figure 2.1 shows profiles of number of days of illness for 30 randomly selected subjects. We observe that the profiles touch the zero-axis many times. In addition, the observed values as well as the between subject variation seems to be higher at later ages. The average evolution, as displayed in Figure 2.2, indicates an increase in number of days of illness with increasing age.

Thirdly, two outcome variables, namely (1) body weight (kg), measured longitudinally from each infant and (2) number of days of diarrheal illness recorded at each visit to assess the diarrheal disease burden will be studied jointly (Table 2.3).

Figure 2.3 shows subject specific profiles of body weight for randomly selected infants, implying considerable between and within subjects variability. In addition, the average profiles, as shown in Figure 2.4, suggest that average weight increases with increasing age, as expected. It is then useful to assess the connection between body weight and days of illness, by studying their association. This can be addressed in the context of so-called Joint models.

2.2 Anopheles Mosquito Data

A longitudinal entomological study was conducted between September 2007 to 2009 (for three years) around the Gilgel-Gibe hydroelectric Power Dam, south-western Ethiopia, to investigate if the dam has influenced abundance and species composition of An. mosquito. For this purpose, all villages surrounding the dam (within a ten km radius) were classified into two (at risk and control) according to their distance from the dam i.e., villages within three kilometers from the dam identified as test (at

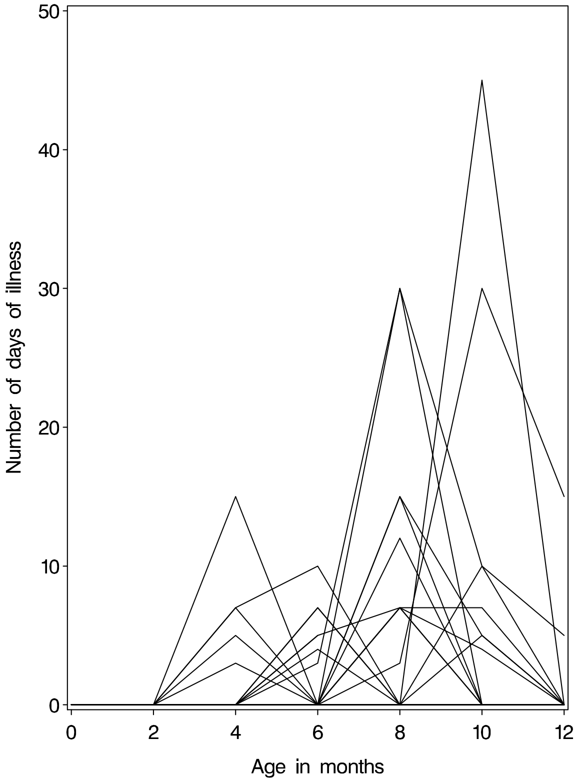

Figure 2.1: *Jimma Infant Growth Study. Selected profiles of number of days of illness per month per child.*

Table 2.3: *Jimma Infant Growth Study. Mean and standard deviation of weight and days of illness at each of the seven follow-up times.*

Age	Mean weight (s.d.)	Mean days of illness (s.d.)
0	3.11(0.52)	0.01(0.19)
2	4.88(0.78)	0.91(4.24)
4	5.97(0.99)	1.28(4.62)
6	6.67(1.12)	1.56(4.87)
8	7.13(1.21)	2.14(5.93)
10	7.50(1.26)	2.63(6.66)
12	7.84(1.28)	2.67(6.95)

risk) villages and the remaining villages, five to ten kilometers away from the dam

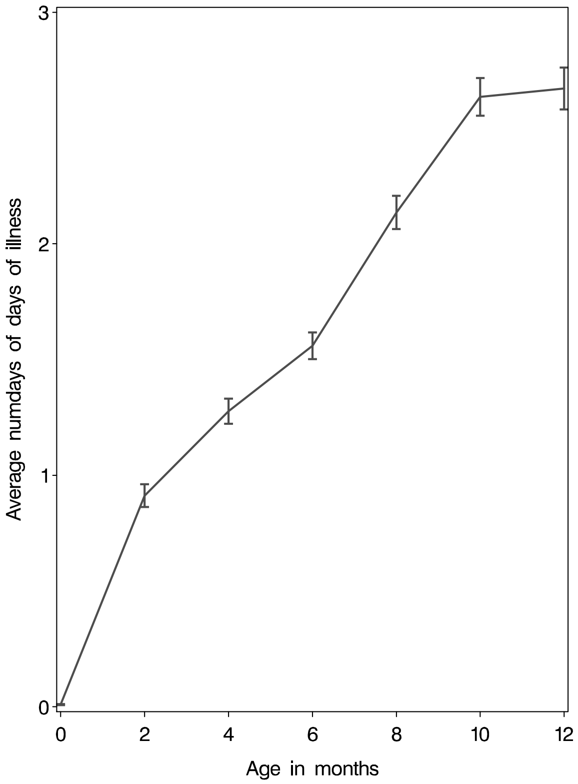

Figure 2.2: *Jimma Infant Growth Study. Average number of days of illness per month per child.*

were identified as controls and from each of these two groups of villages, 8 villages were selected based on various comparability factors for the study. Distance from the dam being the major attribute for considering a village as either at risk or control, the villages were intended to be similar in every other characteristics, such as similar eco-topography, access to health facilities, without major impounded water nearby and homogeneous with respect to socio-cultural and daily economic activities. The study area and setting is described in Yewhalaw *et al.* (2010).

2.2.1 Indoor Resting Collection (IRC)

One aspect of the research consists of the collection of mosquito at each month for the three study years using indoor resting collection (IRC). Ten houses were selected randomly from each selected at risk and control village, and then one room was selected in each house. All collected mosquitoes were counted and sorted based on

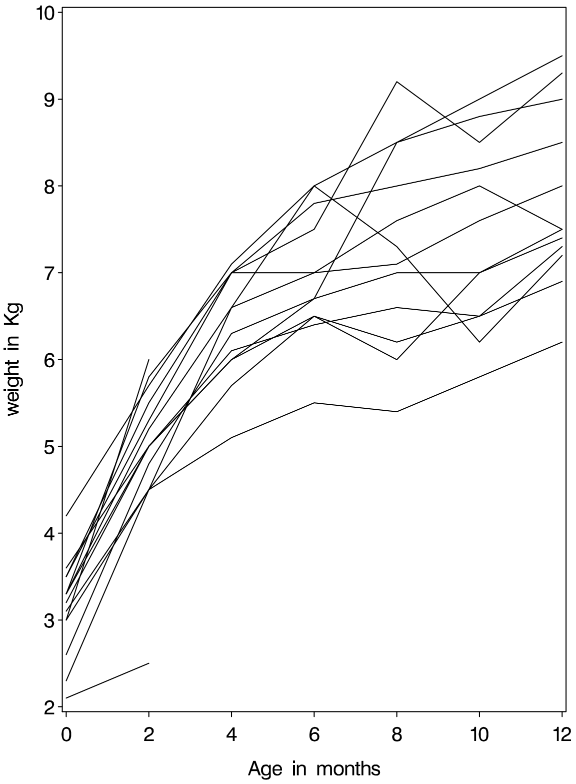

Figure 2.3: *Jimma Infant Growth Study. Individual profiles of weight versus age in months.*

their species type. Of the total nine mosquito species identified, An. gambiae is the most dominant one which constitutes above 95% of the total counts for every year in each village type and therefore is the species we focus on here. Table 2.4 shows the mean, standard deviation and proportion of zero counts among at risk and control villages over the three years. The larger proportion of zero observations and the relatively higher variances compared to the mean, imply that the data are likely to be subject to zero-inflation and overdispersion, in addition to the correlation due to the repeated measures. In this work, it is investigated whether the mean An. gambiae count changes over time (months), differ among village type (at risk vs control), and season (wet vs dry).

The individual profiles for randomly selected houses are displayed in Figure 2.5, and the mean evolution, per village type, are plotted in Figure 2.6. The average profiles indicate that at risk villages are consistently higher than the control villages.

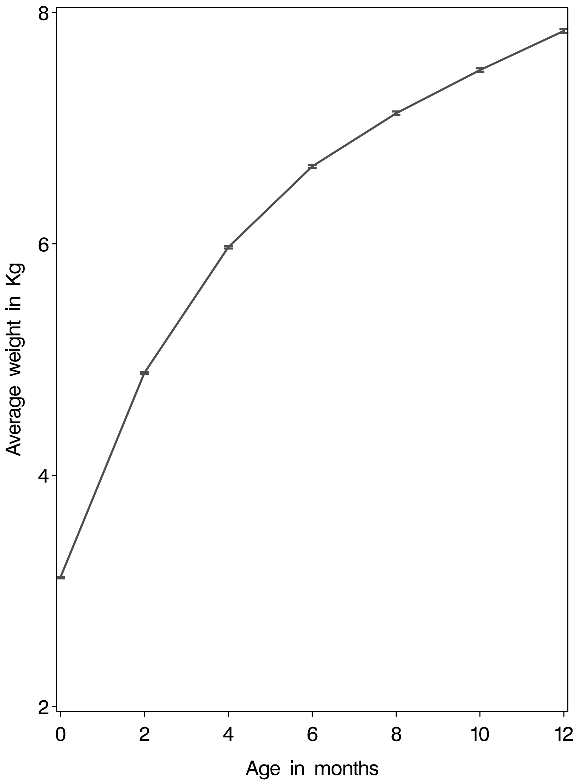

Figure 2.4: *Jimma Infant Growth Study. Average weight of infants versus age in months.*

Table 2.4: *IRC Data. Mean (s.d), and percentage of zeros of An. gambiae by village type and year of collection.*

	control		at risk	
Year	mean(s.d)	% of zeros	mean(s.d)	% of zeros
One	1.48(5.39)	80.2	7.00(16.24)	66.3
Two	2.29(8.50)	80.7	9.20(24.85)	65.5
Three	1.03(4.14)	88.4	4.57(11.28)	74.8
Overall	1.62(6.43)	83.6	6.92(18.76)	69.2

Of course, at this point it is not yet possible to decide on the significance of this difference. Both the individual profiles and average evolution suggest an oscillatory pattern with the observed values attaining higher values at wet seasons and lower

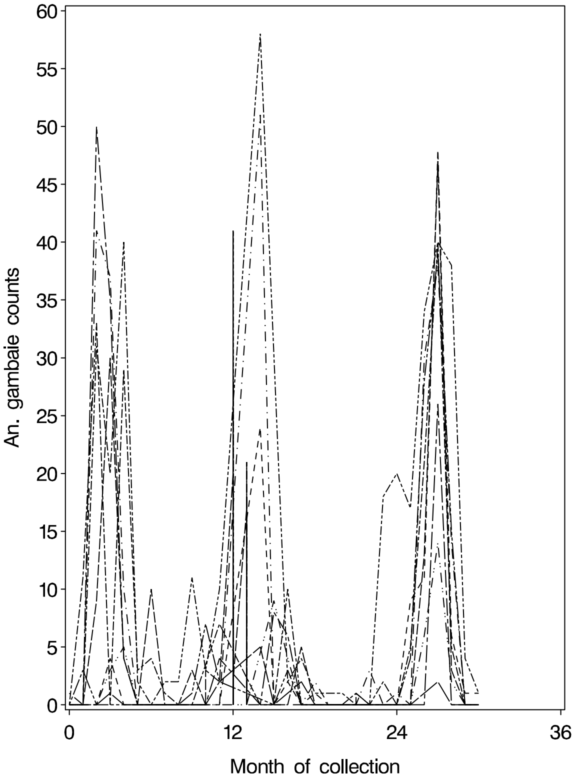

Figure 2.5: *IRC Data. Selected profiles of An. gambaie counts per house per month.*

values at dry seasons, which augment the result in Table 2.4.

2.2.2 CDC Light Trap Catches (CDC)

The alternative approach used to collect An. mosquito was the CDC light trap. Two houses were selected from each at risk and control villages included in the study based on their relative location in the village. One house located at the center of the village, whereas the second located at the periphery. The two houses served as sentinel stations for the study involving light trap catches (LTCs) using CDC light traps. Subject specific profiles for randomly selected houses are displayed in Figure 2.7, and the mean evolution, per village type, are plotted in Figure 2.8, conveying similar pattern with the IRC data except that the mean profiles in the CDC suggest that year one has a higher value, while it is year two in the case of the IRC, though it is not possible to generalize about the significance of time effect in both cases at this

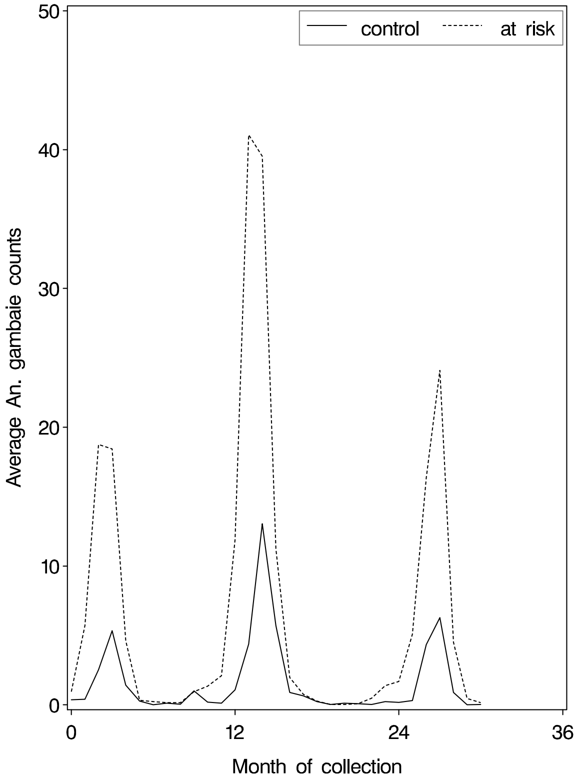

Figure 2.6: *IRC Data. Average number of An. gambaie counts per house per month.*

stage.

The mean, standard deviation and proportion of zero counts among at risk and control villages over the three years are shown in Table 2.5. These results also suggest presence of excessive zeros and a higher sample variance relative to the mean.

2.3 Jimma Longitudinal Family Survey of Youth

The Jimma Longitudinal Family Survey of Youth (JLFSY) is another Ethiopian study where data were collected from households. The study began in 2005, and was repeated in 2007. More than 90% of the study subjects present at baseline were visited and willing to respond in the second round. The study population is representative of the relatively large town of Jimma, the small towns of Yebu, Serbo, and Sheki, and nearby rural areas. The sample includes 3700 households as well as 700 adolescents. The outcome of interest is the adolescents' current school attendance coded as 0 (not

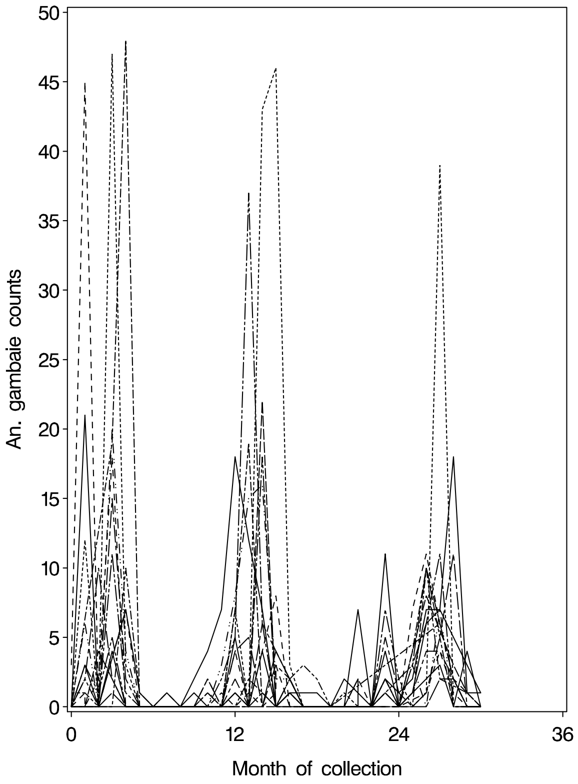

Figure 2.7: *CDC Data. Selected profiles of An. gambaie counts per house per month.*

Table 2.5: *CDC Data. Mean (s.d), and percentage of zeros of An. gambiae by village type and year of collection.*

	control		at risk	
Year	mean(s.d)	% of zeros	mean(s.d)	% of zeros
One	1.59(4.01)	68.8	5.05(12.69)	64.2
Two	0.56(1.88)	81.8	2.10(6.20)	66.5
Three	0.93(3.08)	72.7	1.87(3.99)	58.0
Overall	0.94(2.96)	75.3	2.67(7.65)	62.7

currently attending) or 1 (currently attending). Current school attendance was 90.2% and 91.1% in the first round survey and 93.5% and 92.8% in the second round for male and female adolescents, respectively. The research question is to examine whether or

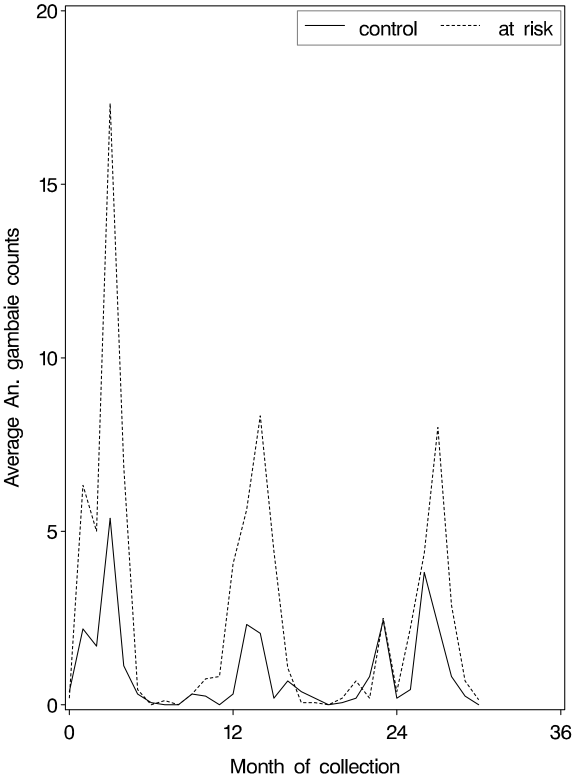

Figure 2.8: *CDC Data. Average number of An. gambaie counts per house per month.*

not the percentage of school attendance depends on adolescents involvement in work to support themselves or their families to earn money, whether they are living in urban towns or rural areas, as well as on gender and age (Belachew *et al.*, 2011).

Another outcome studied in the survey is the adolescents' average number of days of work per week measured repeatedly three times. Mean (s.d.) for year one, year two, and year three are 1.12(2.12), 0.95(2.01) and 1.15(2.23), respectively. The data exhibit higher proportion of zeros: year one (72.4%), year two (77.9%), and year three (76.2%). The research question is to examine whether or not the average number of days of work changes over time, and depends on adolescents age and sex.

2.4 A Clinical Trial in Epileptic Patients

The epileptic data set considered here is obtained from a randomized, double-blind, parallel group multi-center study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's. The study is described in full detail in Faught *et al.* (1996) and Molenberghs and Verbeke (2005). In the study, 45 patients were randomized to the placebo group and 44 to the active (new) treatment group. The number of epileptic seizures were measured on a weekly basis during a 16 weeks period. After this period, patients were entered into a long-term open-extension study, which contains follow-up measurements of patients up to 27 weeks. The key research question is whether or not the additional new treatment reduces the number of epileptic seizures.

The average and median evolutions are shown in Figure 2.9 and Figure 2.10, respectively, suggesting presence of extreme values. The unstable behaviour is also the result of the very little observations available at some of the time-points, especially past week 20 (Molenberghs and Verbeke, 2005). Zero observations account for 33% of the data, with sample average and standard deviation 3.18 and 6.14, respectively. Thus, there is a large proportion of zeros, as well as evidence of overdispersion and correlation stemming from the longitudinal aspect in this data set.

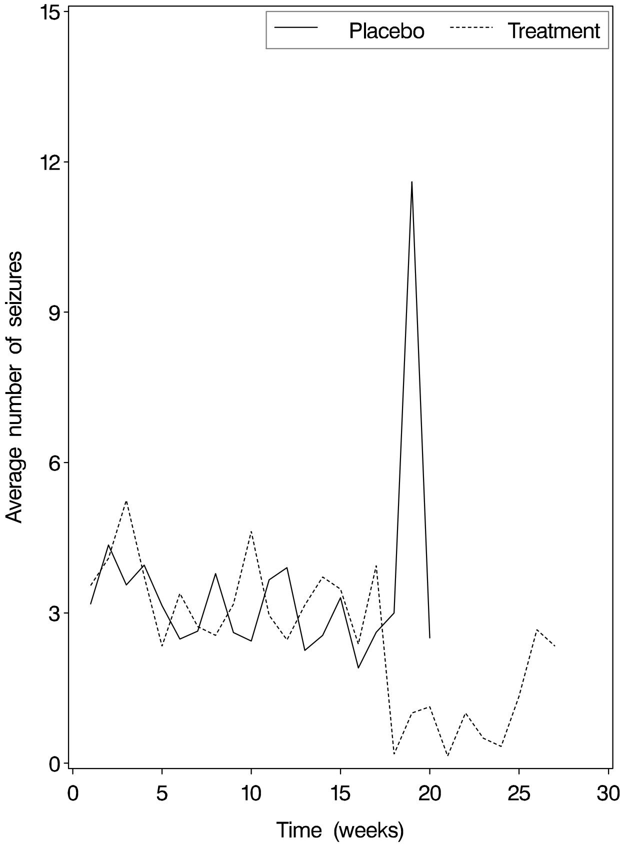

Figure 2.9: *Epilepsy Data. Average number of epileptic seizures versus time.*

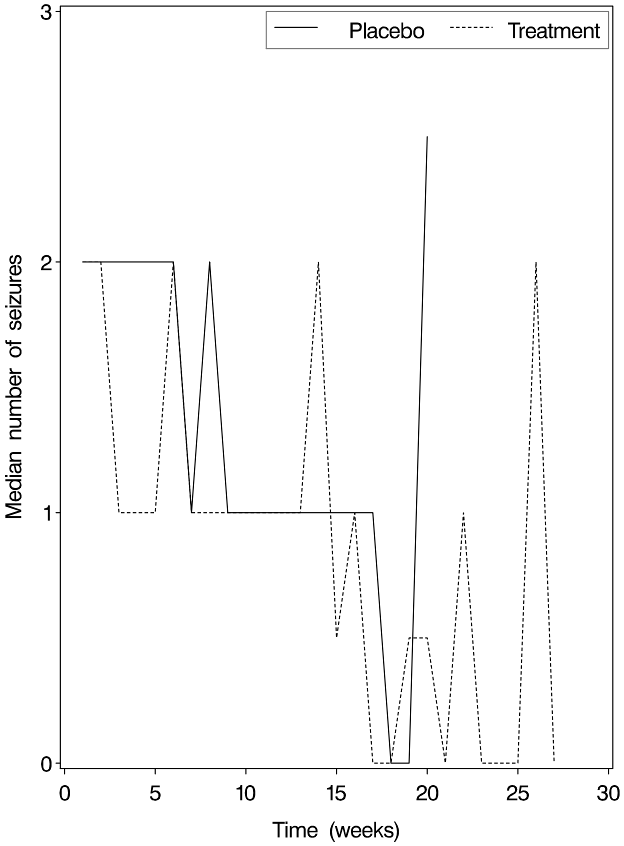

Figure 2.10: *Epilepsy Data. Median number of epileptic seizures versus time.*

Chapter 3

Basic Concepts and Models

In this chapter, we review the basic and frequently used models for continuous, binary and count data. First, we will present briefly the generalized linear models in Section 3.1, focusing on count and binary data, which is followed by overdispersion models, in Section 3.2. Linear mixed models (LMM) and generalized linear mixed models (GLMM) will be the topics of Section 3.3 and Section 3.4, respectively. Section 3.5 is devoted to models combining overdispersion and correlation. Marginalizing random effect models in modeling hierarchical count data will presented in Section 3.6. Section 3.7 deals with the commonly used models to adjust for an excess of zeros in count data analysis. In Section 3.8, generalized estimating equations (GEE) will be reviewed. Finally, in Section 3.9, a Bayesian implementation of GLMM of Section 3.4 is shown briefly.

3.1 Generalized Linear Models

Generalized linear models (GLMs) are often employed for modeling a univariate non-Gaussian data extending ordinary regression models. GLMs include a wider range of statistical models that relate outcome variables such as, counts, binary, rates and ratios, etc to a linear combination of predictor variables (McCullagh and Nelder, 1989; Agresti, 2002; Molenberghs and Verbeke, 2005). Three components specify a generalized linear model: A random component identifies a vector of observations of Y and its probability distribution; a systematic component is a specification for the vector μ in terms of a vector of p fixed unknown parameters $\boldsymbol{\xi}$; and a link function specifies the function of $\mathrm{E}(Y)$ that the model equates to the systematic component.

A family of probability density functions is called an exponential family distribution if it can be expressed as

$$f(y) \quad \equiv \quad f(y|\eta,\phi) \;=\; \exp\left\{\phi^{-1}[y\eta - \psi(\eta)] + c(y,\phi)\right\}, \qquad (3.1)$$

where η and ϕ are unknown parameters, and $\psi(\cdot)$ and $c(\cdot,\cdot)$ are known functions. η and ϕ are termed 'natural parameter' (or 'canonical parameter') and 'scale parameter,' respectively.

3.1.1 Binary Data

Suppose that for each value of the response Y, there are two possible values denoted by 0 and 1. We may write $\mathrm{pr}(Y_i = 0) = 1 - \pi$; $\mathrm{pr}(Y_i = 1) = \pi$ for 'failure' and 'success' probabilities, respectively.

For binary responses, the model of interest is: $Y \sim$ Bernoulli(π). We want to explain variability between outcome values based on covariate values with density function

$$f(y|\eta,\phi) \;=\; \pi^y(1-\pi)^{1-y} \;= \exp\left[y\ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right]. \qquad (3.2)$$

The mean is given by $\mu = \pi$ and the variance, $\mathrm{var}(\mu) = \pi(1-\pi)$ (Nelder and Wedderburn, 1972).

When collecting a set of data, let $Y_1,\ldots,Y_N$ be a set of independent binary outcomes, and let $\boldsymbol{x_1},\ldots,\boldsymbol{x_N}$ represent the corresponding p-dimensional vectors of covariate values. With a logit link function, $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x_i}'\boldsymbol{\xi}$ is the logistic regression model with $\boldsymbol{\xi}$ a vector of unknown regression coefficients.

Choices of link functions, $g(\pi_i)$, with $g(\pi_i) = \eta_i = \boldsymbol{x_i}'\boldsymbol{\xi}$ are available. The commonly used functions are:

- the logit or logistic function

$$g_1(\pi) = \ln\left(\frac{\pi}{1-\pi}\right);$$

- the probit or inverse Normal

$$g_2(\pi) = \Phi^{-1}(\pi);$$

- the complementary log-log function

$$g_3(\pi) = \ln\{-\ln(1-\pi)\}.$$

For N independent observations, $Y_1, \ldots, Y_N$, let $L_i = \log f(y_i|\eta_i, \phi)$ denote the contribution of y_i to the log likelihood. Hence, the log-likelihood function L is

$$L(\boldsymbol{\xi}) = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \log f(y_i|\eta_i, \phi),$$

where $f(y_i|\eta_i, \phi)$ is as defined by (3.2) for observation i. A general-purpose iterative methods, such as Newton-Raphson or Fisher Scoring, can be applied to obtain the maximum likelihood estimates of unknown model parameters (Agresti, 2002; McCullagh and Nelder, 1989).

3.1.2 Count Data

Count data are very common in many applications. The number of days of illness data and An. mosquito counts, described in Sections 2.1 and 2.2, respectively, are examples of such data. The Poisson distribution belongs to the exponential family and is the simplest and commonly used distribution for analysis of count data. For count responses, the model of interest is: $Y \sim$ Poisson(λ). We want to explain variability between outcome values based on covariate values with density function

$$f(y) \ = \frac{e^{-\lambda}\lambda^{y}}{y!}. \tag{3.3}$$

A key feature of the Poisson distribution is that its mean equals its variance, i.e., the mean is given by $\mu = \lambda$ and the variance, $\text{var}(\mu) = \lambda$, scale parameter $\phi = 1$.

Suppose $Y_1, \ldots, Y_N$ is a set of independent count outcomes, and let $\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}$ represent the corresponding p-dimensional vectors of covariate values. The Poisson regression model with $\boldsymbol{\xi}$ a vector of p fixed, unknown regression coefficients is given by $\log(\lambda_i) = \boldsymbol{x_i}'\boldsymbol{\xi}$.

For observation y_i the contribution to the log-likelihood L_i is $y_i \log \mu_i - \mu_i$; that is for a vector of independent observations $Y_1, \ldots, Y_N$, the log-likelihood function L becomes (McCullagh and Nelder, 1989)

$$L(\boldsymbol{\xi}) = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} (y_i \log \mu_i - \mu_i).$$

Here also, one can apply the Newton-Raphson or Fisher Scoring to obtain the maximum likelihood estimates of unknown model parameters (Agresti, 2002; McCullagh and Nelder, 1989).

3.2 Overdispersion Models

In practice, many types of outcomes using standard models within the GLMs for their analysis, such as binomial and count observations, often exhibit variability exceeding what is predicted by binomial or Poisson (Molenberghs *et al.*, 2010).

The standard Bernoulli model assumes that the mean and variance depend on a single parameter. Though a set of i.i.d. Bernoulli data cannot contradict the mean-variance relationship, it may not hold true for data having a hierarchical structure of the form z_i successes out of n_i trials, such as in cluster and longitudinal studies. To illustrate this, let us consider an example given in Agresti (2002). Suppose that in an experiment pregnant mice are exposed to a toxin and then the number of fetuses in each mouse's litter that show signs of malformations are observed after a week. Each fetus is nested in each mice. Let z_i are the number of fetuses that show signs of malformation out of n_i fetuses for mouse i. The mice also may vary according to other unmeasured characteristics, such as weight, overall health, and genetic makeup. These will then induce extra variability in the probability of malformation from litter to litter than expected for the binomial distribution. One possible way to deal with overdispersion for counts based on binary data is to allow for the overdispersion parameter $\phi \neq 1$ and only specify a relation between the mean and the variance, and then apply quasi-likelihood estimation (Wedderburn, 1974). A simple quasi-likelihood approach uses the variance function, $\text{var}(\pi_i) = \phi\pi_i\frac{(1-\pi_i)}{n_i}$. In this context, if $\phi > 1$, overdispersion is said to occur. An elegant way to account for overdispersion in clustered binary and binomial data is through inclusion of beta random-effects, leading to the so-called beta-binomial model, in which the Bernoulli model is combined with a beta distribution (Molenberghs and Verbeke, 2005; Skellam, 1948; Hinde and Demétrio, 1998a; Hinde and Demétrio, 1998b; Kleinman, 1973).

A key assumption of the GLM Poisson model is that the variance is equal to the mean, $\text{var}(\mu) = \mu = \lambda$. However, in many applications with count data, the observed variance is higher than the mean, leading to overdispersion (Agresti, 2002). In the An. mosquito data described in Section 2.2, suppose that Y_{ij} denote the number of An. gambae counts collected from house i at time j. These counts may vary from house to house based on factors such as distance of houses from the dam and month of collection (dry or wet), which in turn will induce heterogeneity, leading to more variation in the data than predicted by the Poisson model. Like the clustered binary and binomial data, one can apply quasi-likelihood estimation (Wedderburn, 1974). Here also, if $\phi > 1$, overdispersion is said to occur. An alternative approach to modeling overdispersion in count data is combining a Poisson distribution with a random effect λ_i to account for the unobserved heterogeneity. Then, $Y_i/\lambda_i \sim$ Poisson$(\lambda_i\mu_i)$. Since λ_i is unobserved, it is common to assume a gamma distribution, so that the uncondition distribution of the outcome turns out to be a negative binomial distribution (Breslow, 1984; Hinde and Demétrio, 1998a; Hinde and Demétrio, 1998b). The negative binomial distribution has mean, $\mathrm{E}(Y) = \mu$ and variance, $\mathrm{Var}(Y) = \mu(1+\sigma^2)$, where σ^2 is the variance of the unobserved term. If $\sigma^2 > 0$, the variance is larger than the mean, implying negative binomial allows for overdispersion. When $\sigma^2 = 0$, then the Poisson model results as a special case.

3.3 A Model for Longitudinal Continuous Data

For a longitudinal Gaussian outcome, the linear mixed model provides a general and flexible modeling framework based on a random-effects approach (Verbeke and Molenberghs, 2000). Suppose Y_{ij} is the jth continuous outcome measured for subject $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. A linear mixed-effects model is given as

$$Y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b_i} + \varepsilon_{ij}, \tag{3.4}$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ p-dimensional and q-dimensional vectors of known covariate values, $\boldsymbol{\beta}$ a p-dimensional vector of unknown fixed regression coefficients, $\boldsymbol{b_i}$ the q-dimensional vector of the random effects, and $\boldsymbol{\varepsilon}_i$, an n_i-dimensional vector of residual variation. The subject-specific random effect $\boldsymbol{b_i}$ and the residual error $\boldsymbol{\varepsilon}_i$ are independent, and assumed to follow a normal distribution, i.e. $\boldsymbol{b_i} \sim N(\boldsymbol{0}, D)$, and $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \Sigma_i)$, respectively.

The implied marginal model is given by $\boldsymbol{Y_i} \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i D \boldsymbol{Z}'_i + \Sigma_i)$ (Laird and Ware, 1982; Verbeke and Molenberghs, 2000).

The standard way to inference is based on maximizing the marginal likelihood function in (3.5) with respect to $\boldsymbol{\omega}$ (Verbeke and Molenberghs, 2000).

$$L(\boldsymbol{\omega}) = \prod_{i=1}^{N}\left\{(2\pi)^{-n_i/2}|V_i(\boldsymbol{\zeta})|^{-\frac{1}{2}} \times \exp\big(-\frac{1}{2}(\boldsymbol{Y_i} - \boldsymbol{X}_i\boldsymbol{\beta})' V_i{}^{-1}(\boldsymbol{\zeta})(\boldsymbol{Y_i} - \boldsymbol{X}_i\boldsymbol{\beta})\big)\right\}, \tag{3.5}$$

where $\boldsymbol{\zeta}$ is the vector of all variance and covariance components in $V_i = \boldsymbol{Z}_i D \boldsymbol{Z}'_i + \Sigma_i$, and $\boldsymbol{\omega} = (\boldsymbol{\beta}', \boldsymbol{\zeta}')$ is the vector of all parameters in the marginal model for $\boldsymbol{Y_i}$.

Maximum likelihood (ML) and restricted maximum likelihood (REML) are the commonly used parameter estimation methods (Laird and Ware, 1982; Verbeke and Molenberghs, 2000).

3.4 Generalized Linear Mixed Models

When non-Gaussian data are hierarchically organized (repeated measures or clustering), the GLM is usually extended to generalized linear mixed models (GLMMs), with a subject-specific random effect, usually a Gaussian type, added in the linear predictor to capture the correlation (Engel and Keen, 1992; Molenberghs and Verbeke, 2005; Pinheiro and Bates, 2000) or multilevel models are considered (Goldstein, 2002). GLMMs combine the properties of two statistical frameworks that are widely used, linear mixed models and generalized linear models.

Suppose that Y_{ij} is an outcome for the i^{th} subject measured at the j^{th} time point, and $\boldsymbol{b_i}$ are assumed to be normally distributed with mean $\mathbf{0}$ and variance-covariance matrix D, that is $\boldsymbol{b_i} \sim N(\mathbf{0}, D)$, with $\mathrm{E}(\boldsymbol{b_i}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{b_i}) = D$. Then, it is assumed that the conditional distribution of the response, $Y_{ij}|\boldsymbol{b_i}$ is independent and belongs to the following exponential family density

$$f_i(y_{ij}|\boldsymbol{b_i}, \phi) \;=\; \exp\left\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y, \phi)\right\}. \tag{3.6}$$

The expectation is, $\mathrm{E}(Y_{ij}|\boldsymbol{b_i}) = \mu_{ij} = \eta^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i})$, where $\eta(.)$ is a known link function, $\boldsymbol{x}_{ij}$ is a p-dimensional design matrix of the fixed effect parameters $\boldsymbol{\xi}$, and $\boldsymbol{z}_{ij}$ is a q-dimensional design matrix of the random effects $\boldsymbol{b_i}$.

The likelihood contribution of subject i is

$$f_i(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b_i}, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b_i}, \phi) f(\boldsymbol{b_i}|D) d\boldsymbol{b_i}. \tag{3.7}$$

From this the likelihood for $\boldsymbol{\xi}$, D and ϕ is given as

$$L(\boldsymbol{\xi}, D, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\xi}, \boldsymbol{b_i}, \phi) f(\boldsymbol{b_i}|D) d\boldsymbol{b_i}, \tag{3.8}$$

In general, expression (3.8) does not have analytical solution, and hence numerical approximations are needed. An extensive overview of different approximations is available in Molenberghs and Verbeke (2005) and Skrondal and Rabe-Hesketh (2004). In some cases, such as the linear mixed models for continuous outcomes, as given in Section 3.3, the expression in (3.7) takes an n_i-dimensional multivariate normal distribution, which can be solved analytically.

For the case of binary data Y_{ij}, we assume that

$$Y_{ij} \;\sim\; \text{Bernoulli}(\pi_{ij} = \kappa_{ij}), \tag{3.9}$$

$$\kappa_{ij} \;=\; \frac{\exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right)}{1 + \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right)}. \tag{3.10}$$

Turning to count data, let Y_{ij} be the value of the count variable for subject i and time point j. We assume that

$$Y_{ij} \quad \sim \quad \text{Poi}(\lambda_{ij}), \tag{3.11}$$

with the conditional mean λ_{ij} modeled as

$$\lambda_{ij} = \exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\xi} + \boldsymbol{z}_{ij}'\boldsymbol{b_i}\right). \tag{3.12}$$

3.5 Models Combining Overdispersion with Normal Random-Effects

In practice, both overdispersion and correlation can happen together, and this led Molenberghs *et al.* (2010) to formulate a flexible and unified modeling framework, which they termed the *combined model*, to simultaneously capture overdispersion and correlation for a wide range of clustered data, including count, binary and time-to-event. These authors brought together two sets of random effects. The normally distributed subject specific-random effects capture the correlation, while a conjugate measurement-specific random effect on the natural parameter, is used to accommodate overdispersion. The latter leads to the beta-binomial model for binary data and the negative-binomial model for count data. A detailed overview of the model can be found in Molenberghs *et al.* (2010).

In line with Molenberghs *et al.* (2010), the combined model having both the overdispersion and normal random effects takes the form

$$f_i(y_{ij}|\boldsymbol{b_i}, \phi) \;=\; \exp\left\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y,\phi)\right\}, \tag{3.13}$$

The expectation is, $\text{E}(Y_{ij}|\boldsymbol{b_i}) = \mu_{ij} = \theta_{ij}\kappa_{ij}$, where $\theta_{ij} \sim g_{ij}(\vartheta_{ij}, {\sigma_{ij}}^2)$, ϑ_{ij} and ${\sigma_{ij}}^2$ are mean and variances of θ_{ij}, respectively. The likelihood contribution of subject i is

$$f_i(\boldsymbol{y_i}|\boldsymbol{\vartheta}, D, \boldsymbol{\vartheta}_i, \Sigma_i) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\vartheta}, \boldsymbol{b_i}, \boldsymbol{\theta}_i)\; f(\boldsymbol{b_i}|D)\; f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \Sigma_i)\; d\boldsymbol{b_i}\; d\boldsymbol{\theta}_i.$$

From this, the likelihood is given as:

$$\begin{aligned} L(\boldsymbol{\vartheta}, D, \boldsymbol{\vartheta}, \Sigma) &= \prod_{i=1}^{N} f_i(\boldsymbol{y_i}|\boldsymbol{\vartheta}, D, \boldsymbol{\vartheta}_i, \Sigma_i) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\vartheta}, \boldsymbol{b_i}, \boldsymbol{\theta}_i)\; f(\boldsymbol{b_i}|D)\; f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \Sigma_i)\; d\boldsymbol{b_i}\; d\boldsymbol{\theta}_i. \end{aligned}$$

For the case of binary data, we assume that

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(\pi_{ij} = \theta_{ij}\kappa_{ij}), && (3.14)\\ \kappa_{ij} &= \frac{\exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right)}{1+\exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right)}. && (3.15) \end{aligned}$$

Explicitly considering $\theta_{ij} \sim \text{Beta}(\alpha, \beta)$, then $\phi = \alpha/(\alpha+\beta)$, and

$$\sigma^2_{ij} = \sigma_{i,jj} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \qquad \sigma_{i,jk} = \rho_{ijk}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The model combining overdispersion and correlation for binary outcome will be considered in more detain both in likelihood and Bayesian framework in **Chapter 4**.

For count data, let Y_{ij} be the jth outcome measured for subject $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. The Poisson model with normal and gamma random effects can be specified as

$$Y_{ij} \sim \text{Poi}(\lambda_{ij} = \theta_{ij}\kappa_{ij}), \tag{3.16}$$

with the conditional mean λ_{ij} modeled as $\theta_{ij}\kappa_{ij}$ and

$$\kappa_{ij} = \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right), \tag{3.17}$$

where $\boldsymbol{b_i} \sim N(\mathbf{0}, D)$, and $\theta_{ij} \sim \text{Gamma}(\alpha, \beta)$, $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ p-dimensional and q-dimensional vectors of known covariate values, and $\boldsymbol{\xi}$ a p-dimensional vector of unknown fixed regression coefficients.

Molenberghs *et al.* (2007) and Molenberghs *et al.* (2010) marginalized the combined model analytically over the gamma random effect, whereby this partially marginalized model takes the form:

$$\begin{aligned} f(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}) &= \int f(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) f(\theta_{ij}|\alpha_j, \beta_j) d\theta_{ij} \\ &= \begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}}. \end{aligned}$$

Then further numerical integration over the normal random effects can be made to obtain the maximum likelihood estimates.

Details of the marginal expressions for the mean vector $\text{E}(Y_{ij})$ and variance-covariance of $\boldsymbol{Y_i}$ are given in Molenberghs *et al.* (2007) and Molenberghs *et al.* (2010).

Note that the Poisson-normal GLMM results as special case of the combined model, when overdispersion random effects θ_{ij} are omitted, with a conditional mean given by:

$$Y_{ij} \sim \text{Poi}(\kappa_{ij}). \tag{3.18}$$

We apply the following notational convention. The model that brings both features together, i.e., the combined model, is denoted as (PNG), where the first symbol 'P' refers to basic Poisson model, the second symbol 'N' is for normal random effects and the final one for gamma random effects. The special case, which follows by leaving out the gamma random-effects structures, i.e., the Poisson-normal GLMM is denoted as (PN-), and omitting only the normal-random effects by (P-G). The simplest case arises when both random-effects are dropped, leading to Poisson GLM model (P--).

3.6 Marginalized Multilevel Models

The (PN-) and (PNG) models of Section 3.4 and Section 3.5, respectively, are specified conditional upon the random effects, which yield subject-specific interpretations for parameter estimates. In practice, however, interest could be on the marginal or population-averaged effects of covariates. In this section, we present marginalized versions of (PN-) and (PNG). We use the superscripts 'm' and 'c' to refer to marginal and conditional, respectively.

Marginalizing (PN-) leads to M(PN-) (Zeger *et al.*, 1988; Molenberghs *et al.*, 2007):

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \int e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}+\boldsymbol{z}'_{ij}\boldsymbol{b_i}} f(\boldsymbol{b_i})d\boldsymbol{b_i} \\ &= e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}^m+\frac{1}{2}\boldsymbol{z}'_{ij}D\boldsymbol{z}_{ij}} \\ &= \kappa^m_{ij}. \end{aligned} \tag{3.19}$$

where f is the zero-mean normal density with variance-covariance matrix $\boldsymbol{D}$. Then, the marginalized combined model, denoted as M(PNG), with slight modification of (3.19), takes a marginal mean of the form:

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \int_b \int_\theta \theta_{ij} e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}+\boldsymbol{z}'_{ij}\boldsymbol{b_i}} d\Theta_\theta f(\boldsymbol{b_i})d\boldsymbol{b_i} \\ &= \mathrm{E}(\theta_{ij}) e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}^m+\frac{1}{2}\boldsymbol{z}'_{ij}D\boldsymbol{z}_{ij}} \\ &= e^{\ln E(\theta_{ij})+\boldsymbol{x}'_{ij}\boldsymbol{\xi}^m+\frac{1}{2}\boldsymbol{z}'_{ij}D\boldsymbol{z}_{ij}} \\ &= \lambda^m_{ij}. \end{aligned} \tag{3.20}$$

Based on work of Griswold and Zeger (2004) for (PN-), employing the connector function Δ_{ij} and log-log-normal specification, we find:

$$\begin{aligned} e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}^m} &= \int e^{\Delta_{ij}+\boldsymbol{z}'_{ij}\boldsymbol{b_i}} f(\boldsymbol{b_i})d\boldsymbol{b_i} \\ &= e^{\Delta_{ij}+\frac{1}{2}\boldsymbol{z}'_{ij}D\boldsymbol{z}_{ij}}. \end{aligned} \tag{3.21}$$

From (3.19) and (3.21), Δ_{ij} becomes:

$$\Delta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\xi}^m - \frac{1}{2}\boldsymbol{z}'_{ij}D\boldsymbol{z}_{ij}. \tag{3.22}$$

As shown in Iddi and Molenberghs (2012), for (PNG) with overdispersion random effect θ_{ij} and marginal mean λ_{ij}^m, the connector follows from the integral equation given by:

$$\lambda_{ij}^m = \int_b \int_\theta \theta_{ij} e^{\Delta_{ij}+\boldsymbol{z}'_{ij}\boldsymbol{b_i}} d\Theta_\theta f(\boldsymbol{b_i}) d\boldsymbol{b_i}. \tag{3.23}$$

Then, from (3.20) and (3.23), the connector Δ_{ij} for (PNG) becomes:

$$\begin{aligned} \Delta_{ij} &= \ln E(\theta_{ij}) + \boldsymbol{x}'_{ij}\boldsymbol{\xi}^m - \frac{1}{2}\boldsymbol{z}'_{ij}D\boldsymbol{z}_{ij} \qquad (3.24) \\ &= -\ln(\alpha\beta) + \boldsymbol{x}'_{ij}\boldsymbol{\xi}^m - \frac{1}{2}\boldsymbol{z}'_{ij}D\boldsymbol{z}_{ij}. \end{aligned}$$

3.7 Models for Excessive Zero Observations

In many applications with count data, a larger proportion of zero values than what would be expected under distributional assumptions is common. Such data are often fitted by using either a hurdle model (Mullahy, 1986; Greene, 1994) or a zero-inflated model (Lambert, 1992).

The hurdle model is a way of modeling count data using a two-part approach, whereby the first part is a binary model for the count value zero or positive. Given the value is positive, a count distribution, say f_i, is truncated-at-zero and fitted for the second part. Suppose Y_i is a univariate count outcome, and π_i is probability of the i^{th} observation to be in the zero state. The hurdle model assumes Y_i fulfills a distribution given by

$$p(Y_i = y_i) = \begin{cases} \pi_i & \text{if } y_i = 0, \\ (1-\pi_i)\frac{f_i(y_i|\lambda_i)}{1-f_i(0|\lambda_i)} & \text{if } y_i > 0. \end{cases} \tag{3.25}$$

An alternative approach to account for excessive zeros is a zero-inflated model, which assumes zeros to come from two processes. The first process generates only zeros with probability, say π_i for observation i, and the second process generates counts with probability, say $(1 - \pi_i)$. In a zero-inflated model, Y_i follows a zero-inflation probability distribution given by

$$p(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)f_i(0|\lambda_i) & \text{if } y_i = 0, \\ (1-\pi_i)f_i(y_i|\lambda_i) & \text{if } y_i > 0. \end{cases} \tag{3.26}$$

π_i and λ_i are functions of covariates. Link functions, such as logit or probit, can be used for π_i, and the common log link is used for λ_i.

For clustered count data, exhibiting overdipserson, correlation, and zero-inflation, one can assemble the above concepts of the hurdle and zero-inflated models, with the (PNG) model of Section 3.5, and the M(PNG) model of Section 3.6, to further account for the excess of zeros. Hence, the hurdle and zero-inflated extensions of M(PNG) will be the topics of Section 6.1 and Section 6.2, respectively.

3.8 Generalized Estimating Equations

Generalized estimating equations (GEE) is a popular and widely used method in modeling repeated non-Gaussian data when primary interest is on marginal mean parameters. GEE was first introduced by Liang and Zeger (1986). The association between the vector of repeated measurements taken from a given subject $\boldsymbol{Y_i}$ is captured by allowing correlation within the subject through a so-called *working correlation*. Details can be found in Molenberghs and Verbeke (2005). The marginal expectations $\mathrm{E}(Y_{ij}) = \mu_{ij}$ can be directly modeled in terms of known covariates. For count data, for example, $\ln(\mu_{ij}) = x'_{ij}\boldsymbol{\xi}$.

The GEE approach assumes a working correlation matrix $R_i = R_i(\boldsymbol{\alpha})$ for $\boldsymbol{Y_i}$ where $\boldsymbol{\alpha}$ is a vector of nuisance parameters.

The score equations take the form

$$U(\boldsymbol{\xi}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\xi}'} {V_i}^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) = 0, \tag{3.27}$$

where V_i is the covariance matrix of $\boldsymbol{Y_i}$, written as $V_i = V_i(\boldsymbol{\xi}, \boldsymbol{\alpha}) = \phi {A_i}^{1/2} R_i {A_i}^{1/2}$, ϕ being an overdispersion parameter and A_i is a matrix with marginal variances on the main diagonal and zero elsewhere.

The estimator $\widehat{\boldsymbol{\xi}}$ is the solution of (3.27). Liang and Zeger (1986) showed that when the marginal mean μ_{ij} has been correctly specified and when mild regularity conditions hold, $\widehat{\boldsymbol{\xi}}$ is consistent and asymptotically normally distributed with mean $\boldsymbol{\xi}$ and variance covariance matrix

$$\mathrm{Var}(\widehat{\boldsymbol{\xi}}) = {I_0}^{-1} I_1 {I_0}^{-1}, \tag{3.28}$$

where

$$I_0 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i}'}{\partial \boldsymbol{\xi}} {V_i}^{-1} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\xi}'}, \tag{3.29}$$

$$I_1 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i}'}{\partial \boldsymbol{\xi}} V_i{}^{-1} \mathrm{Var}(\boldsymbol{Y_i}) V_i{}^{-1} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\xi}'}. \tag{3.30}$$

$I_0{}^{-1}$ and $I_0{}^{-1} I_1 I_0{}^{-1}$ are referred to as the 'model based' and 'empirically corrected' variance estimators, respectively, and the latter, also known as 'sandwich estimator', is the one to be used.

Often, working correlation matrix, such as exchangeable, autoregressive and unstructured are assumed. The advantage of GEE is that, even with incorrect specification of the variance function, one can still estimate $\boldsymbol{\xi}$ consistently. However, for missing data, bias can arise in the estimates unless the data are missing completely at random (MCAR).

3.9 Hierarchical Bayesian Model

Consider the GLMM model of Section 3.4 for Y_{ij} as an outcome for the i^{th} subject measured at the j^{th} time point, and $\boldsymbol{b_i}$ are assumed to be normally distributed with mean $\mathbf{0}$ and variance-covariance matrix D, that is $\boldsymbol{b_i} \sim N(\mathbf{0}, D)$, with $\mathrm{E}(\boldsymbol{b_i}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{b_i}) = D$. Recall that the conditional distribution of the response, $Y_{ij}|\boldsymbol{b_i}$ is independent and belongs to the following exponential family density

$$f_i(y_{ij}|\boldsymbol{b_i}, \phi) \;=\; \exp\left\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y, \phi)\right\}, \tag{3.31}$$

For the full Bayesian treatment of this model, $\boldsymbol{\xi}$ is not known and thus has its own prior distribution, $p(\boldsymbol{\xi})$. The joint prior distribution is (Gelman *et al.*, 2004)

$$p(\boldsymbol{\xi}, \boldsymbol{b_i}) = p(\boldsymbol{\xi}) p(\boldsymbol{\xi}|\boldsymbol{b_i}), \tag{3.32}$$

and the joint posterior distribution is

$$p(\boldsymbol{\xi}, \boldsymbol{b_i}|y) = p(\boldsymbol{\xi}, \boldsymbol{b_i}) p(y|\boldsymbol{\xi}). \tag{3.33}$$

For the case of binary data Y_{ij}, we assume that

$$Y_{ij} \;\sim\; \mathrm{Bernoulli}(\pi_{ij} = \kappa_{ij}), \tag{3.34}$$

$$\kappa_{ij} \;=\; \frac{\exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right)}{1 + \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right)}. \tag{3.35}$$

The following prior distributions can be used for $\boldsymbol{\xi}$ and $\boldsymbol{b_i}$: $\xi_i \sim N(0, 10^{-6})$, $\boldsymbol{b_i} \sim N(0, \tau_i)$, as also suggested in the literature (Gilks *et al.*, 1996; Gelman *et al.*, 2004). For the hyper parameters τ_i, the inverse-Gamma prior $IG(0.001, 0.001)$, can be used (Gelman *et al.*, 2004).

Estimation is based on the popular Markov chain Monte Carlo (MCMC) technique. Samples are drawn from the posterior distribution, which is defined by the prior distributions for the parameters and the likelihood function for the data.

Chapter 4

Modeling Overdispersed Longitudinal Binary Data Using a Combined Beta and Normal Random-effects Model

For hierarchical binary data, such as clustered or longitudinal, the so-called generalized linear mixed models is very popular (Engel and Keen, 1992; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005). While, i.i.d. binary data do not violate the prescribed mean-variance relationship, this is not the case in clustered or longitudinal binary data. To deal with overdispersion, Hinde and Demétrio (1998a) and Hinde and Demétrio (1998b) considered a random-effects approach, leading to beta-binomial model (Skellam, 1948; Kleinman, 1973). In practice, such data may exhibit both overdispersion and correlation aspects at once. This led Molenberghs *et al.* (2010) to propose a flexible family of models, termed as the combined model, to deal with both features simultaneously through two separate sets of random effects, not only for binary and binomial data, but also for count and time-to-event outcomes.

For binary and binomial data, Kassahun *et al.* (2012) studied the combined model in the Bayesian framework. In this setting, the possibility to specify prior distribution will be an advantage, especially when conjugate priors are used (Spiegelhalter *et al.*, 2002). The full likelihood approach similar to Molenberghs *et al.* (2010) is also considered for comparison purposes.

The outline of this chapter is as follows. In Section 4.1, the model combining overdispersion and normal random effects is reviewed. Estimation techniques both in the likelihood and Bayesian setting are provided in Section 4.2. Results of the Jimma Infant Growth data and Jimma Family Survey of Youth are presented in Sections 4.3.1 and 4.3.2, respectively, with the results of the two estimation techniques compared in Section 4.3.3. Finally, a brief discussion and some concluding remarks are provided in Section 4.4. The contribution of this chapter has been published in Kassahun *et al.* (2012).

4.1 Models Combining Conjugate and Normal Random Effects

Combining both the overdispersion effects (Section 3.2) as well as the normal random effects (Section 3.4) into the generalized linear model framework, produces the following general family (Molenberghs *et al.*, 2010):

$$f_i(y_{ij}|\boldsymbol{b_i},\boldsymbol{\xi},\theta_{ij},\phi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij},\phi)\right\}, \tag{4.1}$$

with notation similar to the one used in (3.31), but now with conditional mean

$$E(Y_{ij}|\boldsymbol{b_i},\boldsymbol{\xi},\theta_{ij}) = \mu_{ij}^c = \theta_{ij}\kappa_{ij}, \tag{4.2}$$

where the random variable $\theta_{ij} \sim \mathcal{G}_{ij}(\vartheta_{ij},\sigma_{ij}^2)$, $\kappa_{ij} = g(\boldsymbol{x}_{ij}'\boldsymbol{\xi} + \boldsymbol{z}_{ij}'\boldsymbol{b_i})$, ϑ_{ij} is the mean of θ_{ij} and σ_{ij}^2 is the corresponding variance. Finally, as before, $\boldsymbol{b_i} \sim N(\boldsymbol{0}, D)$. Write $\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\xi} + \boldsymbol{z}_{ij}'\boldsymbol{b_i}$. Unlike in Section 3.4, we now have two different notations, η_{ij} and λ_{ij}, to refer to the linear predictor and/or the natural parameter. The reason is that λ_{ij} encompasses the random variables θ_{ij}, whereas η_{ij} refers to the 'GLMM part' only. A detailed overview of the model can be found in Molenberghs *et al.* (2010).

For the case of binary data, we assume that

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(\pi_{ij} = \theta_{ij}\kappa_{ij}), & (4.3)\\ \kappa_{ij} &= \frac{\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\xi} + \boldsymbol{z}_{ij}'\boldsymbol{b_i}\right)}{1+\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\xi} + \boldsymbol{z}_{ij}'\boldsymbol{b_i}\right)}, & (4.4)\end{aligned}$$

where $\theta_{ij} \sim \text{Beta}(\alpha,\beta)$. Indeed, this model also intuitively seems useful, as overdispersion and correlation due to the data hierarchy can occur simultaneously.

The model is a two-level model with two types of random effects: (a) the b_i, to accommodate correlation among repeated measures (and some overdispersion); (b) the θ_{ij} for additional overdispersion. While (a) turns the model into a two-level model, rather than a one-level one, (b) does not further add a level, because it merely accommodates overdispersion. This is to be compared with a classical generalized linear model, where also overdispersion random effects can be taken into account (e.g., beta in the Bernoulli model to yield the beta-binomial; gamma in the Poisson model to yield the negative binomial; etc.), while keeping the so-resulting models remain one-level models.

Further, because the θ_{ij} follow a conjugate distribution, they do not have an impact on the shape of the regression function (like the normal random effects in a linear mixed model), hence there is greatly reduced sensitivity to assumptions about the random effects. This is one of the elegant properties of conjugate random effects.

4.2 Estimation

In the likelihood framework, estimation proceeds by integration. Recall from Section 3.5, the likelihood contribution of subject i is

$$f_i(\boldsymbol{y_i}|\boldsymbol{\vartheta}, D, \boldsymbol{\vartheta}_i, \Sigma_i) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\vartheta}, \boldsymbol{b_i}, \boldsymbol{\theta}_i)\ f(\boldsymbol{b_i}|D)\ f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \Sigma_i)\ d\boldsymbol{b_i}\ d\boldsymbol{\theta}_i. \tag{4.5}$$

From this, the likelihood is given as:

$$\begin{aligned} L(\boldsymbol{\vartheta}, D, \boldsymbol{\vartheta}, \Sigma) &= \prod_{i=1}^{N} f_i(\boldsymbol{y_i}|\boldsymbol{\vartheta}, D, \boldsymbol{\vartheta}_i, \Sigma_i) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\vartheta}, \boldsymbol{b_i}, \boldsymbol{\theta}_i)\ f(\boldsymbol{b_i}|D)\ f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i, \Sigma_i)\ d\boldsymbol{b_i}\ d\boldsymbol{\theta}_i. \end{aligned} \tag{4.6}$$

Here, $\boldsymbol{\vartheta}$ groups all parameters in the conditional model for $\boldsymbol{Y}_i$. In the binomial case, the expression takes the form:

$$f(z_{ij}|n_{ij}, \boldsymbol{b_i}) = \sum_{t=0}^{n_{ij}-z_{ij}} (-1)^t \kappa_{ij}^{z_{ij}+t} \frac{n_{ij}!}{z_{ij}!t!(n_{ij}-z_{ij}-t)!} \cdot \frac{B(z_{ij}+t+\alpha_j, \beta_j)}{B(\alpha_j, \beta_j)}, \tag{4.7}$$

with

$$\kappa_{ij} = \frac{\exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right)}{1 + \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right)}.$$

It is straightforward to obtain the fully marginalized probability by numerically integrating over the normal random effects, and using a tool such as the SAS procedure NLMIXED that allows for normal random effects in arbitrary, user-specified models. More details can be found in Molenberghs *et al.* (2010). As an alternative estimation method, we turn to the Bayesian paradigm, combined with the popular Markov Chain Monte Carlo (MCMC) technique, making analyses of real-world complex data feasible (Gilks *et al.*, 1996). In the Bayesian approach, prior distributions are assigned to the parameters and the random effects to adjust for parameter uncertainty. Bayesian inference for estimation of parameter θ is based on the posterior distribution, which is proportional to the likelihood multiplied with the prior distribution.

The Jimma longitudinal studies are characterized by clustering, resulting from the repeated measurements, leading to both correlation and overdispersion. When modeling such data, incorporating prior distributions for model parameters, including that of subject and observation specific random effects, will better handle the underlying uncertainties, instead of assuming that they are fixed. With the same model specification as in the likelihood framework, the parameters ξ, b_i, and θ_{ij} are taken to be a priori independent, i.e., $p(\boldsymbol{\vartheta}, D, \boldsymbol{\vartheta}_i, \Sigma_i) = p(\boldsymbol{\vartheta})p(D)p(\boldsymbol{\vartheta}_i)p(\Sigma_i)$ and the following prior distributions are used: $\xi \sim N(0, 10^{-6})$, $b_i \sim N(0, \tau_i)$, as also suggested in the literature (Gilks *et al.*, 1996; Gelman *et al.*, 2004) and $\theta_{ij} \sim \text{Beta}(\alpha, \beta)$, is unimodal and concave, when $\alpha > 1,\ \beta > 1$ (Agresti, 2002). For the hyper parameters τ_i, the inverse-Gamma prior $IG(0.001, 0.001)$, and for α and β, an improper uniform prior is used, as also suggested by Gelman *et al.* (2004).

Note that the beta-binomial distribution is a compound distribution of the binomial and its conjugate beta, which can be used to capture overdispersion in binomial data. The beta-binomial approximates the binomial distribution arbitrarily well when its two non-negative parameters, α and β, determining its shape, are sufficiently large. If one or both of these parameters are less than 1, then the probability mass function will go to infinity near its boundaries, 0 and 1, and hence not concave. As a result, the mode does not exist, leading to computational problems in MCMC. For this reason, we used the restriction $\alpha > 1,\ \beta > 1$, such that the density is always concave and unimodal whereby it is always finite over the support $[0, 1]$, as shown in Kassahun *et al.* (2012). An example SAS and WinBugs implementation of the combined model is shown in Appendix A.

Spiegelhalter *et al.* (2002) suggest use of the so-called Deviance Information Criterion for model comparison in Bayesian inference. Assume a probability model $P(y|\theta)$. The effective number of parameters with respect to a model with parameter Θ is given by $pD\{y, \Theta, \widetilde{\theta}(y)\} = E_{\theta|y}[-2\log p(y|\theta)] + 2\log[p\{y|\widetilde{\theta}(y)\}]$. We shall usually drop the arguments $\{y, \Theta, \widetilde{\theta}(y)\}$ from notation. Generally, we take $\widetilde{\theta}(y) = E(\theta|y)$, the posterior mean of the parameters. For $f(y)$ being a fully specified standardizing term that is a function of the data alone, pD, defined as a 'mean deviance minus the deviance of the means,' is given by $pD = E[D(\theta|y)] - D(E[\theta|y])$, where $D(\theta) = -2\log P(y|\theta) + 2\log f(y)$ is the Bayesian deviance, used as a measure for goodness of fit. The deviance information criterion (DIC), defined as the classical estimate of fit plus twice the effective number of parameters $DIC = D(E[\theta|y]) + 2pD = E[D(\theta|y)] + pD$ is used for model comparison. According to this criterion, the model with the smallest DIC is to be preferred. pD and DIC are easily computed using the available MCMC output by taking the posterior mean of the deviance to obtain $E[D(\theta|y)]$ and the plug-in estimate of the deviance $D(E[\theta|y])$ using the posterior means $E[\theta|y]$ of the parameter θ. In non-hierarchical models, pD approximates the effective number of parameters to be estimated. However, for hierarchical models, pD is a measure of model complexity instead of being merely the number of effective parameters to be estimated. In general, it is difficult to say what would constitute an important difference in DIC for model comparison. Spiegelhalter *et al.* (2002) suggested models receiving DIC within 1-2 of the 'best', deserve consideration, and 3-7 have considerably less support. These rules of thumb appear to work reasonably well. For the best model preferred based on DIC, the important risk factors could be identified looking the credible intervals. In the case of a single parameter and data that can be summarised in a single sufficient statistic, the credible interval and the confidence interval can be treated equivalently. Hence, to identify, the risk factor, we considered whether zero is in or outside of the credible interval.

We also attempted to fit the beta-binomial marginal density, although it is not one commonly encountered in software packages like WinBugs, where an observation x_i contributes a likelihood term L_i. We used the so-called *zero trick*, a Poi(ϕ) observation of zero has likelihood $\exp(-\phi)$, so if our observed data is a set of 0's, and ϕ_i is set to $-\log(L_i)$, we would obtain the correct likelihood contribution (Spiegelhalter *et al.*, 2003). This zero trick allows for arbitrary sampling distributions and is particularly suitable when, say, dealing with truncated distributions. However, our case studies showed that this method can be very inefficient and give a very high Monte Carlo error.

In terms of parameter interpretation, we would like to refer back to the beneficial properties that come with the conjugacy property. Indeed, because the θ_{ij} follow a conjugate distribution, the interpretation of the parameters is the same as in a classical generalized linear mixed model. Precisely, this means that the effect on the regression parameters only comes from the normal random effects in the linear predictor, a fact well documented. For a review, see, for example, Molenberghs and Verbeke (2005).

4.3 Results

For the Jimma infants study, assuming independence, the sample average probability of success and the sample variance are 0.150 and 0.128, respectively, indicating that the prescribed mean-variance link is maintained. In contrast, in the binomial setting, taking the hierarchical structure into account, the sample average and the sample variances are 0.141 and 2.107, respectively, implying that the sample contradicts the mean-variance relationship for these data.

Similar exploratory analyses on the Jimma Longitudinal Family Survey of Youth were undertaken. For the binomial response, taking the two repeated measurements results in sample average probability of success 0.919 and sample variance 0.168 indicating that the results are in line with the prescribed mean-variance relationship which is known to be always true for the Bernoulli case. This may suggest, at first sight, that these data are not prone to exhibit strong overdispersion, even in the hierarchical binomial setting. In addition to the exploratory analysis, we also conducted tests for overdispersion. The commonly used approach is to compute the ratio of the residual deviance to the residual degrees of freedom, which approximates the overdispersion parameter ($\hat{\phi}$). When the ratio is appreciably larger than 1, overdispersion is said to occur. It is pointed out that this approach could be misleading when $n_i p_i$ is not sufficiently large, where p_i is the probability of the success event. This is because it is based on asymptotic theory. As a result, a better approach is based on a quasi-binomial model, which allows extra dispersion (Skellam, 1948). The approximated overdispersion ($\hat{\phi} = 2.37$) computed as the ratio of the residual deviance to the residual degrees of freedom in the binomial, and the one estimated in the quasi-binomial model ($\hat{\phi} = 2.47$) for the Jimma Infants Growth data are very similar, both suggesting the presence of strong ovderdispersion. However, a similar analysis for the Jimma Family Survey data, does not suggest a considerable overdispersion, with values 0.765 and 1.129, approximated by the ratio of the residual deviance to the residual degrees of freedom in the binomial, and estimated by the quasi-binomial, respectively.

4.3.1 The Jimma Infant Growth Study

We will analyze the binary BMI data. The following model is assumed for the mean structure: $Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij})$, for subject i and measurement j, and

$$\begin{aligned}\text{logit}(\pi_{ij}) &= \xi_0 + b_{0i} + (b_{1i} + \xi_1)T_{ij} + \xi_2 G_i + \xi_3 P_{1i} + \xi_4 P_{2i} + \xi_5 B_{ij} \\ &\quad + \xi_6 G_i T_{ij} + \xi_7 P_{1i} T_{ij} + \xi_8 P_{2i} T_{ij} + \xi_9 B_{ij} T_{ij}, \qquad (4.8)\end{aligned}$$

where G_i is a gender indicator, P_{1i} and P_{2i} are dummy variables for place of residence corresponding to rural and urban areas and using semi-urban areas as a reference. T_{ij} is the time point at which the j^{th} measurement is taken for the i^{th} subject, which is centered at month six. B_{ij} denotes whether the i^{th} infant is breast fed or not at time j. The random intercept $b_i \sim N(0, D)$.

The Infant Growth dataset is analyzed with a simple logistic model, a beta-binomial model introducing only an overdispersion parameter, a random-effects logistic model that introduces a random-effects term to take the repeated structure of the data into account, and finally the combined model, which allows for both overdispersion and a random-effects term. Parameter estimates of the logistic model and the beta-binomial model are presented in Table 4.1 and the corresponding estimates of the logistic-normal model and the combined model are given in Table 4.2. Clearly, the logistic-normal model is an important improvement, in terms of likelihood, relative to both the ordinary logistic model and the beta-binomial. Moreover, considering the combined model, there is a very strong improvement in fit when the beta and normal random effects are simultaneously allowed for. The overdispesion term in the combined model is significant ($p < 0.001$), implying the presence of considerable extra variability due to the grouped nature of the data, which is beyond what can be accommodated by the commonly used logistic-normal model.

The logistic-normal model ignores the overdispersion that results from the grouped nature of the data. On the other hand, the beta-binomial model accommodates overdispersion which is assumed independent, implying independence between repeated measurements. Again, this is not realistic and therefore the combined model is the more viable candidate, supported further by the aforementioned likelihood comparison.

The combined model suggests that the intercept, the time effect, main effects of place of residence and breastfeeding are significant, which is also true for time interaction with rural place of residence and breast feeding. However, main effect and slope of gender were not significant, implying that proportion of overweight seems to be invariant among male and female infants over time. Infants living in rural and

Table 4.1: *Jimma Infant Growth Study. Parameter estimates, standard errors, and p-values for the regression coefficients in (1) the logistic model, (2) the beta-binomial model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

		Logistic	Beta-binomial
Effect	Parameter	Estimate (s.e., p)	Estimate (s.e., p)
Intercept	ξ_0	−1.896(0.128, 0.001)	−0.448(1.099, 0.683)
Time	ξ_1	0.127(0.031, 0.001)	0.188(0.090, 0.037)
Gender:Male	ξ_2	0.027(0.025, 0.294)	0.029(0.039, 0.456)
Place rural	ξ_3	−0.602(0.029, 0.001)	−0.949(0.501, 0.058)
Place urban	ξ_4	−0.376(0.037, 0.001)	−0.628(0.381, 0.099)
Breast feeding	ξ_5	0.545(0.128, 0.001)	0.788(0.347, 0.023)
Slope Gender:Male	ξ_6	−0.003(0.006, 0.602)	−0.007(0.011, 0.534)
Slope rural	ξ_7	0.018(0.007, 0.014)	0.029(0.020, 0.161)
Slope urban	ξ_8	0.016(0.009, 0.097)	0.026(0.022, 0.251)
Slope Breast feeding	ξ_9	−0.133(0.031, 0.001)	−0.199(0.098, 0.041)
Std. dev. random intercept	$\sqrt{d_0}$	—	—
Std. dev. random slope	$\sqrt{d_1}$	—	—
Ratio	α/β	—	1.827(1.622, 0.259)
−2log-likelihood		41,286	41,286

urban areas are at lower risk of overweight as compared to those in semi-urban ares with ($\widehat{\xi_3} = -1.058$, $p = 0.001$), and ($\widehat{\xi_4} = -0.689$, $p = 0.001$), respectively. Further, early initiation of breastfeeding has a protective effect against the risk of overweight in late infancy ($\widehat{\xi_9} = -0.167$, $p = 0.001$), as shown in Table 4.2.

4.3.2 Jimma Longitudinal Family Survey of Youth

We will now analyze current school attendance. For the logit, consider the model: $Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij})$, with

$$\text{logit}(\pi_{ij}) \quad = \quad \xi_0 + b_i + \xi_1 A_{ij} + \xi_2 G_i + \xi_3 P_{1ij} + \xi_4 P_{2ij} + \xi_5 W_{ij} + \xi_6 R_{ij}, \quad (4.9)$$

where A_{ij} is the age of the i^{th} subject at the j^{th} visit, G_i is the gender of the i^{th} subject. P_{1ij} and P_{2ij} denote the two dummy variables for place of residence of the

Table 4.2: *Jimma Infant Growth Study. Parameter estimates, standard errors, and p-values for the regression coefficients in (1) the logistic-normal model, and (2) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

		Logistic-normal	Combined
Effect	Parameter	Estimate (s.e., p)	Estimate (s.e., p)
Intercept	ξ_0	−2.741(0.186, 0.001)	−2.661(0.215, 0.001)
Time	ξ_1	0.132(0.042, 0.002)	0.147(0.049, 0.003)
Gender:Male	ξ_2	0.010(0.054, 0.852)	0.020(0.064, 0.751)
Place rural	ξ_3	−0.908(0.064, 0.001)	−1.058(0.082, 0.001)
Place urban	ξ_4	−0.581(0.082, 0.001)	−0.689(0.099, 0.001)
Breast feeding	ξ_5	0.635(0.179, 0.001)	0.764(0.209, 0.001)
Slope Gender:Male	ξ_6	−0.003(0.010, 0.728)	−0.005(0.012, 0.660)
Slope rural	ξ_7	−0.015(0.011, 0.167)	0.024(0.014, 0.085)
Slope urban	ξ_8	−0.011(0.014, 0.432)	0.015(0.017, 0.377)
Slope Breast feeding	ξ_9	−0.149(0.044, 0.001)	−0.167(0.049, 0.001)
Std. dev. random intercept	$\sqrt{d_0}$	1.774(0.034, 0.001)	2.107(0.088, 0.001)
Std. dev. random slope	$\sqrt{d_1}$	0.193(0.007, 0.001)	0.237(0.014, 0.001)
Ratio	α/β	—	0.234(0.045, 0.001)
−2log-likelihood		37,000	36,971

i^{th} subject on the j^{th} visit, which are urban, semi-urban, and rural by taking rural as a reference. W_{ij} indicates whether the i^{th} adolescent is engaged in some work for the family or help support on the j^{th} visit. Finally, R_{ij} is the j^{th} round or measurement occasion of the i^{th} subject, and $b_i \sim N(0, d)$.

Results from fitting all four models (with/without normal random effect; with/without beta random effect) can be found in Tables 4.3 and 4.4. Likelihood comparison of the beta-binomial with the standard logistic model shows no improvement in fit, implying absence of strong evidence for overdispersion. This can be noted from likelihood comparisons of the simple logistic and the beta-binomial on the one hand, as well as the logistic-normal and the combined, on the other. One can easily see, however, that the commonly used logistic-normal and the combined models are significant improvements over the standard logistic model. We further observe, while

Table 4.3: *Jimma Longitudinal Family Survey of Youth. Parameter estimates, standard errors, and p-values for the regression coefficients in (1) the logistic model, (2) the beta-binomial model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

		Logistic	Beta-binomial
Effect	Parameter	Estimate (s.e., p)	Estimate (s.e., p)
Intercept	ξ_0	1.171(0.626, 0.061)	1.155(0.702, 0.099)
Age	ξ_1	0.039(0.049, 0.414)	0.044(0.055, 0.421)
Place urban	ξ_2	0.971(0.148, 0.001)	1.089(0.266, 0.001)
Place semi-urban	ξ_3	0.979(0.159, 0.001)	1.104(0.284, 0.001)
Gender:Female	ξ_4	−1.111(0.123, 0.001)	−1.226(0.237, 0.001)
Work	ξ_5	0.134(0.122, 0.274)	0.146(0.138, 0.288)
Round	ξ_6	0.341(0.141, 0.016)	0.390(0.178, 0.029)
Std. dev. random effect	$\sqrt{d}$	—	—
Ratio	α/β	—	0.009(0.014, 0.528)
−2log-likelihood		1987.7	1987.4

the logistic-normal model suggests a significant intercept ($p = 0.045$), that the same does not emerge when the combined model is considered ($p = 0.099$) implying the beta random effect has some impact on the p-values. For these data, with two repeated measures per subject, the logistic-normal model seems adequate and the overdispersion term in the combined model is not significant ($p = 0.29$), strengthening what has been mentioned in the earlier sections. Further extension by adding random slope did not improve the fit of neither the logistic-normal nor the combined models (details not shown).

Based on the logistic-normal model in Table 4.4, adolescents living in urban and semi-urban areas have higher school attendance than those living in rural areas, with ($\widehat{\xi_2} = 1.098$, $p = 0.001$), and ($\widehat{\xi_3} = 1.092$, $p = 0.001$), respectively. Gender is also significantly associated with school attendance, where female adolescents are lower ($\widehat{\xi_4} = -1.241$, $p = 0.001$). There is evidence that school attendance increases in the second round visit compared to the first ($\widehat{\xi_6} = 0.398$, $p = 0.010$).

Table 4.4: *Jimma Longitudinal Family Survey of Youth. Parameter estimates, standard errors, and p-values for the regression coefficients in (1) the logistic-normal model, and (2) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

		Logistic-normal	Combined
Effect	Parameter	Estimate (s.e., p)	Estimate (s.e., p)
Intercept	ξ_0	1.443(0.719, 0.045)	1.463(0.888, 0.099)
Age	ξ_1	0.046(0.056, 0.408)	0.058(0.070, 0.408)
Place urban	ξ_2	1.098(0.178, 0.001)	1.379(0.393, 0.001)
Place semi-urban	ξ_3	1.092(0.189, 0.001)	1.339(0.368, 0.001)
Gender:Female	ξ_4	−1.241(0.147, 0.001)	−1.499(0.339, 0.001)
Work	ξ_5	0.153(0.144, 0.287)	0.189(0.182, 0.296)
Round	ξ_6	0.398(0.155, 0.010)	0.519(0.237, 0.028)
Std. dev. random effect	$\sqrt{d}$	1.138(0.188, 0.001)	1.342(0.318, 0.001)
Ratio	α/β	—	0.013(0.013, 0.293)
−2log-likelihood		1972.9	1972.1

4.3.3 Comparison Between Estimation Methods

For comparison with the previously applied estimation method in the likelihood framework, we again apply the same models to the two surveys, but now in the Bayesian framework. After generating 70,000 MCMC samples for the combined, and 50,000 MCMC samples for the logistic-normal, beta-binomial, and simple logistic, the first 10,000 samples are discarded and treated as so-called burn-in samples. The remaining samples are used to summarize the posterior estimates. Two distinct chains were used to check sensitivity to the initial values, and convergence was met. Convergence was checked using the Gelman-Rubin diagnostic as well as by visual inspection of the trace and QQ plots (Brooks and Gelman, 1998).

The posterior summaries of logistic and beta-binomial for the Jimma Infants Growth dataset are given in Table 4.5, while the corresponding estimates of the logistic-normal and combined models are presented in Table 4.6. Similarly, for the Jimma Longitudinal Family Survey of Youth, estimates of these four models are shown in Tables 4.7 and 4.8. The parameter estimates are fairly similar to what was obtained previously in the likelihood approach in both cases, except for differences in the case of the beta-binomial for the Jimma Infants data in Table 4.5 when compared with Table 4.1.

In terms of significance of the parameters, the same conclusion is reached for the two case studies in both approaches, except that the beta-binomial for the intercept and time effects in the Jimma infants study shows significance in the likelihood framework as given in Section 4.3.1, while the same does not emerge from the Bayesian analysis, as observed from the 95% credible interval which include zero for these effects. We compared the various models using the DIC criterion. For both studies, there is a significant reduction in the DIC of the logistic-normal and the beta-binomial, as compared to the simple logistic. We observe a rather high degree of model improvement by combining beta and normal random effects simultaneously, to allow for both the overdispersion and the data hierarchy. Moreover, the logistic and the beta-binomial ignore the correlation stemming from the data hierarchy on the one hand, and the logistic-normal does not allow for the overdispersion, on the other, which altogether make the combined model the preferred one.

According to Spiegelhalter *et al.* (2002), in comparing complex hierarchical models where the number of parameters is not clearly defined, pD is the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest, not only measures the effective number of parameters but also the model complexity. These authors further noted that the contribution pD_i of each observation i turned out its leverage, defined as the relative influence that each observation has on its own fitted value. For y_i conditionally independent given θ, pD_i, shows its interpretation as the difficulty in estimating θ with y_i. This shows the connection between the sample size, the parameters to be estimated, and the pD. The Jimma infants ($n = 7969$) and the Jimma Longitudinal family survey ($n = 2100$) data have large number of subjects followed longitudinally, where each subject was measured seven and two times, respectively. For these reasons, the pD values, as presented in Table 4.6 and Table 4.8, appeared to be larger as the by-product of the MCMC estimation to obtain leverage of each observation. The two competing models, i.e., the logistic-normal and the combined models resulted relatively in larger values of pDs in both of our case studies.

Unlike the Jimma infants study in Table 4.6, pD of the combined model for the Jimma Longitudinal Family Survey of Youth in Table 4.8, ($pD = 211.9$), is lower than that of the logistic-normal ($pD = 241.5$). This implies that, for the Jimma Longitudinal Family Survey of Youth, the combined model is less complex to fit than the logistic-normal, although this is not what we usually expect, as the combined model seems more complex, since it includes both beta and normal random effects,

Table 4.5: *Jimma Infant Growth Study. Estimated posterior mean and standard deviation in (1) the logistic model, (2) the beta-binomial model.*

		Logistic	Beta-binomial
Effect		Mean(s.d.)	Mean(s.d.)
Intercept	ξ_0	$-1.894(0.123)$	$-$ 1.486(1.488)
Time	ξ_1	0.126(0.031)	0.155(0.207)
Gender:Male	ξ_2	0.027(0.026)	0.003(0.066)
Place rural	ξ_3	$-0.602(0.029)$	$-2.486(1.290)$
Place urban	ξ_4	$-0.377(0.037)$	$-1.973(1.210)$
Breast feeding	ξ_5	0.543(0.123)	1.126(0.294)
Slope Gender:Male	ξ_6	$-0.003(0.006)$	$-0.015(0.016)$
Slope rural	ξ_7	0.018(0.007)	0.160(0.178)
Slope urban	ξ_8	0.015(0.009)	0.1610.182)
Slope Breast feeding	ξ_9	$-0.132(0.030)$	$-0.289(0.097)$
Std. dev. random intercept	$\sqrt{d_0}$	—	—
Std. dev. random slope	$\sqrt{d_1}$	—	—
Ratio	α/β	—	3.222(0.524)
DIC		41,310.0	40,390.0
pD		9.9	2511.0

while the logistic-normal includes only the normal random effects. However, for these specific data, this resulted likely because there is less conflict between the specific data set, and the prior distributions which could be associated to the conjugacy of the beta random effects, as well as the peculiar data features including number of subjects and repeated measurements per subject. The posterior densities for these two models are provided in Appendix E.

4.4 Discussion

In this chapter, we have presented a model that integrates normal and beta random effects into a single model, termed the combined model. Our work builds upon that of Molenberghs *et al.* (2010), who brought together normal random effects to induce as-

Table 4.6: *Jimma Infant Growth Study. Estimated posterior mean and standard deviation in (1) the logistic-normal model, and (2) the combined model.*

		Logistic-normal	Combined
Effect		Mean(s.d.)	Mean(s.d.)
Intercept	ξ_0	$-2.773(0.191)$	$-2.755(0.258)$
Time	ξ_1	$0.137(0.042)$	$0.169(0.062)$
Gender:Male	ξ_2	$0.020(0.054)$	$0.026(0.069)$
Place rural	ξ_3	$-0.915(0.065)$	$-1.115(0.085)$
Place urban	ξ_4	$-0.606(0.083)$	$-0.749(0.103)$
Breastfeeding	ξ_5	$0.666(0.185)$	$0.903(0.253)$
Slope Gender:Male	ξ_6	$-0.003(0.010)$	$-0.006(0.012)$
Slope rural	ξ_7	$0.015(0.011)$	$0.026(0.015)$
Slope urban	ξ_8	$0.011(0.014)$	$0.017(0.018)$
Slope Breastfeeding	ξ_9	$-0.144(0.041)$	$-0.192(0.061)$
Std. dev. random intercept	$\sqrt{d_0}$	$1.783(0.035)$	$2.212(0.074)$
Std. dev. random slope	$\sqrt{d_1}$	$0.193(0.007)$	$0.250(0.013)$
Ratio	α/β	—	$0.288(0.031)$
DIC		33,605.1	33,377.6
pD		5400.7	6218.3

sociation between repeated binary and binomial data, and a beta-binomial distributed random factor in the log-linear predictor to fine tune the overdispersion.

Maximum likelihood estimation was considered by integrating over the random effects using the SAS procedure NLMIXED.

Further, Bayesian inference has been applied. Prior information about the parameters induces correlation, which then leads to reduced effective dimensionality although the reduction depends on the available data (Spiegelhalter *et al.*, 2002). Complexity reflects the difficulty in fit and hence it seems reasonable that the measure of complexity may depend on both the prior information concerning the parameters under scrutiny and the specific data that are observed. This can be elucidated from the Jimma Longitudinal Family Survey of Youth result, where the combined model is less complex in fit, which likely results from the conjugacy of the beta random effect and the number of subjects as well as the repeated measurements per subject (Kassahun

Table 4.7: *Jimma Longitudinal Family Survey of Youth. Estimated posterior mean and standard deviation in (1) the logistic model, (2) the beta-binomial model.*

		Logistic	Beta-binomial
Effect		Mean(s.d.)	Mean(s.d.)
Intercept	ξ_0	1.185(0.624)	1.151(0.731)
Age	ξ_1	0.039(0.049)	0.047(0.057)
Place urban	ξ_2	0.977(0.148)	1.134(0.183)
Place semi-urban	ξ_3	0.987(0.161)	1.161(0.202)
Gender:Female	ξ_4	$-1.113(0.123)$	$-1.266(0.148)$
Work	ξ_5	0.133(0.122)	0.154(0.140)
Round	ξ_6	0.343(0.142)	0.404(0.165)
Std. dev. random effect	$\sqrt{d}$	—	—
Ratio	α/β	—	0.0111(0.0029)
DIC		2002.0	2001.0
pD		6.97	13.77

et al., 2012).

Analysis of the case studies show that, in the presence of overdispersion and clustering, the combined model results in improvement in model fit, which is similar to the finding in Molenberghs *et al.* (2010).

This study revealed that early breastfeeding lowers the risk of overweight at late infancy. This finding is in line with Bergmann *et al.* (2003), who showed that breastfed infants had lower BMI after 3 months from birth than bottlefed infants, though the BMIs at birth were nearly identical in both groups. Owen *et al.* (2005), who reviewed sixty-one studies, states that initial breastfeeding protects against obesity in later life, although the precise magnitude of the association remains unclear. Unlike Owen *et al.* (2005), the present study showed that infants in the breastfed group were fatter, at birth, as compared to those who were not breastfed. This is likely because of the unmeasured maternal history, such as maternal BMI, and socio-cultural aspects, which are considered to be the risk factors of overweight in children (Gillman *et al.*, 2006). In addition, it is a common practice in the study area that mothers provide additional liquid or solid food starting from early infancy, in addition to breastfeeding.

Table 4.8: *Jimma Longitudinal Family Survey of Youth. Estimated posterior mean and standard deviation in (2) the logistic-normal model, and (2) the combined model.*

		Logistic-normal	Combined
Effect		Mean(s.d.)	Mean(s.d.)
Intercept	ξ_0	1.452(0.732)	1.272(0.953)
Age	ξ_1	0.047(0.057)	0.077(0.078)
Place urban	ξ_2	1.107(0.180)	1.427(0.270)
Place semi-urban	ξ_3	1.104(0.192)	1.382(0.269)
Gender:Female	ξ_4	−1.247(0.149)	−1.528(0.214)
Work	ξ_5	0.155(0.145)	0.199(0.184)
Round	ξ_6	0.401(0.157)	0.521(0.203)
Std. dev. random effect	$\sqrt{d}$	1.148(0.203)	1.417(0.266)
Ratio	α/β	—	0.013(0.003)
DIC		1943.0	1915.0
pD		241.5	211.9

This is probably because they believe that a child with more weight is considered as healthy, which is likely to have its own impact on the BMI in the early infancy. In this study, it is also shown that place of residence does not have a long term effect in the risk of overweight, instead it is the mode of feeding which is more important. The baseline differences observed in the risk of overweight among infants living in urban, semi-urban areas might be attributable to other family related factors like social class, family income, educational level of the parents, and other socio-cultural variables, which are indicated to affect the nutrition of young children and women in Ethiopia (Macro., 2008).

In investigating school attendance among adolescents, this study showed that girls have a lower rate of current school attendance than boys, which is a common situation in most Sub-Saharan African Countries. According to the World Health Organization (WHO, 2009), there was a clear gender gap observed in primary or secondary school enrollment when the Gender Parity Index (GPI), the ratio of female to male enrollment, is considered. Between the years 1999 and 2003, GPI was found to be 0.7, indicating that there were only 7 girls enrolled at primary schools for every 10 boys. This gender gap increases with increasing level of education. This study also showed that adolescents in urban and semi-urban area have a higher rate than those in the rural areas, which is in line with report of the World Bank (2005), where it was stated that among children in rural areas with a school in the neighborhood, less than 44 % registered for school; in urban areas, the percentage is much higher up to 86 %. According to the report, distance to the nearest school, household characteristics, and learning environment were among the possible reasons of the gap in the school attendance.

Future studies on early growth of children could benefit from careful measurement of a wider range of potential confounders of overweight.

Further efforts should be made to fill the gap in school attendance among boys and girls, as well as urban and rural areas by focusing on the potential causes, such as lagging experience in primary schooling, which is then exacerbated by such factors as the practice of early marriage among Ethiopian women, families' reluctance to invest in girls' education. Situating schools closer to children's homes in rural areas, and improvement of the quality of the services is necessary. Longitudinal studies with better number of repeated measurements per subject should be conducted to get better insight on the trends of school enrollment and survival of adolescents.

Chapter 5

A Zero-Inflated Overdispersed Hierarchical Poisson Model

Count data are most commonly modeled using the Poisson model. Very often, extensions of this model are being considered, for a variety of reasons: (1) a hierarchical structure in the data, e.g., due to clustering in the data, repeated measurements of the outcome, etc.; (2) the occurrence of overdispersion, meaning that the variability in the data is not equal to the mean, as prescribed by the Poisson distribution; and (3) the occurrence of extra zeros beyond what a Poisson model allows for. The first issue is often accommodated through the inclusion of random subject-specific effects. Though not always, one conventionally assumes such random effects to be normally distributed (Engel and Keen, 1992; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005). Overdispersion is often dealt with through an overdispersion model, such as, for example, the negative-binomial model for count data (Breslow, 1984; Lawless, 1987), where the natural parameter is assumed to follow a gamma distribution. An excessive number of zeros is regularly accounted for using so-called zero-inflated models, and studied for univariate count data by Lambert (1992) and Greene (1994), with an extension for hierarchical setting studied in Min and Agresti (2005) and Lee *et al.* (2006).

This chapter proposes a general modeling framework in which correlation, overdispersion and zero-inflation (ZI) can appear together. The proposal is an extension of the modeling approach defined by Molenberghs *et al.* (2010) in which clustering and overdispersion are accommodated for through two separate sets of normal and gamma random effects in a Poisson model (PNG). Adjustment for the excessive zeros assumes that zeros may come from two processes: a point-mass or a Poisson-normal-gamma process, as a mixture, leading to ZI(PNG) model. This chapter is organized as follows. In Section 5.1, the ZI(PNG) model is described, followed by the estimation technique given in Section 5.2. The ZI(PNG) and its special cases, are applied and compared based on two real data sets, with the results given in Section 5.3. Furthermore, a simulation study to study the behaviour of the model and the bias in parameter estimates, which might result from omitting of one or more of the overdispersion, correlaion and excessive zeros is presented in Section 5.4. Some concluding remarks are given in Section 5.5. The contribution of this chapter is based on Kassahun *et al.* (2014a).

5.1 Zero-inflated Models

In zero-inflated count models, it is assumed that there are two processes that can generate zeros: zeros may come from both a point mass (process 1) as well as from the count component (process 2). It is assumed that for observation i at time j, process 1 is chosen with probability π_{ij} and process 2 with probability $1-\pi_{ij}$ (Hinde and Demétrio, 1998a; Hinde and Demétrio, 1998b). Process 1 generates only zeros, whereas process 2, $f_i(y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij})$, generates counts from a Poisson, a negative-binomial model, a Poisson-normal GLMM, or a Poisson-gamma-normal combined model. In its most general form, the zero-inflated Poisson-gamma-normal model is given as the following mixture:

$$Y_{ij} \quad \sim \quad \begin{cases} 0 & \text{with probability } \pi_{ij}, \\ f_i(y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij}) & \text{with probability } 1-\pi_{ij}, \end{cases} \tag{5.1}$$

leading to the probabilities $p(Y_{ij}=y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij},\pi_{ij})$ given by

$$p(Y_{ij}=y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij},\pi_{ij}) = \quad \begin{cases} \pi_{ij}+(1-\pi_{ij})f_i(0|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij}) & \text{if } y_{ij}=0, \\ (1-\pi_{ij})f_i(y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij}) & \text{if } y_{ij}>0. \end{cases} \tag{5.2}$$

The zero-inflation component $\pi_{ij}=\pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma}+\boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i})$ is modeled using a Bernouilli model: in the simplest case with only an intercept, but potentially containing known regressors $\boldsymbol{x}_{2ij}$ and $\boldsymbol{z}_{2ij}$, a vector of zero-inflation coefficients $\boldsymbol{\gamma}$ to be estimated, as well as random effects $\boldsymbol{b}_{2i}$. Common link functions, such as the logit or probit, can be used. Note that $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$, and $\boldsymbol{b}_i$ in Section 3.5 are now replaced by $\boldsymbol{x}_{1ij}$, $\boldsymbol{z}_{1ij}$, and $\boldsymbol{b}_{2ij}$, respectively, for the non-zero count part. The regressors in the count and zero-inflation component can either be overlapping, a subset of the regressors can be used for the zero-inflation, or entirely different regressors for the two parts can be used. In many cases, but of course not always, a simple random-intercept model is adequate, where $\boldsymbol{b}_{1i} = b_{1i}$, $\boldsymbol{b}_{2i} = b_{2i}$, and $\boldsymbol{z}_{1ij} = \boldsymbol{z}_{2ij} = 1$. Assuming that the random effects are normally distributed and possibly correlated with correlation parameter ρ, the variance-covariance matrix is

$$\boldsymbol{D} = \begin{pmatrix} d_1 & \rho\sqrt{d}_1\sqrt{d}_2 \\ \rho\sqrt{d}_1\sqrt{d}_2 & d_2 \end{pmatrix}.$$

The model is denoted as ZI(PNG), as an obvious extension with earlier notational conventions. Three obvious special cases are ZI(PN-), ZI(P-G), and ZI(P--). Also, all four models without zero inflation are special cases as well. The conditional mean and variance of the ZI(PNG) are:

$$\begin{aligned} \mathrm{E}(Y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) &= \theta_{ij}\kappa_{ij}(1-\pi_{ij}), & (5.3) \\ \mathrm{Var}(Y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) &= \theta_{ij}\kappa_{ij}(1-\pi_{ij})[1+\theta_{ij}\kappa_{ij}(\pi_{ij}+1/\alpha)]. & (5.4) \end{aligned}$$

It can be seen that the conditional variance is inflated as a result of either overdispersion in the data (parameter α), or as a result of zero-inflation (parameter π_{ij}), or both.

5.2 Estimation

Likelihood estimation of the (PNG) is done by integrating over the random effects, assembling the marginal likelihood, and maximizing it in the usual way. Molenberghs *et al.* (2007) and Molenberghs *et al.* (2010) marginalized analytically over the gamma random effect, with then further numerical integration over the normal random effects. This enables the use of a flexible normal random-effects tool such as the SAS procedure NLMIXED. From Section 3.5, the partially marginalized (PNG) takes the form:

$$\begin{aligned} f(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}) &= \int f(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) f(\theta_{ij}|\alpha_j, \beta_j) d\theta_{ij} & (5.5) \\ &= \begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}}. & (5.6) \end{aligned}$$

This idea extends in a straightforward fashion to the ZI(PNG):

$$\begin{aligned} & f(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \boldsymbol{b}_{2i}, \boldsymbol{\gamma}) \\ & \quad = I(y_{ij}=0)\pi_{ij} \\ & \qquad +(1-\pi_{ij}) \begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}}, \end{aligned}$$

with $\pi_{ij} = \pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma} + \boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i})$. A sample SAS implementation code is given in Appendix B.

5.3 Results

5.3.1 The Jimma Infant Growth Study

We will fit the ZI(PNG) to the data, introduced in Section 2.1, and compare it to its special cases: (P--), (P-G), (PN-) (PNG), ZI(P--), ZI(PN-), and ZI(P-G). We model κ_{ij} as

$$\begin{aligned}\ln(\kappa_{ij}) &= \xi_0 + b_{1i} + \xi_1 R_i + \xi_2 U_i + \xi_3 T_{ij} + \xi_4 G_i + \xi_5 B_{ij} + \xi_6 H_{ij} + \xi_7 R_i T_{ij} \\ &\quad + \xi_8 U_i T_{ij} + \xi_9 G_i T_{ij} + \xi_{10} B_{ij} T_{ij} + \xi_{11} H_{ij} T_{ij}\end{aligned}$$

and the zero-inflation probability (π_{ij}) as

$$\text{logit}(\pi_{ij}) = \gamma_0 + b_{2i} + \gamma_1 R_i + \gamma_2 U_i + \gamma_3 T_{ij} + \gamma_4 G_i + \gamma_5 B_{ij} + \gamma_6 H_{ij},$$

with R_i an indicator for rural residence and U_i for urban residence. The semi-urban residence category is taken as the reference. Further, G_i is a gender indicator and T_{ij} is the time point at which the j^{th} measurement is taken for the i^{th} subject; B_{ij} and H_{ij} denote, respectively, whether or not the i^{th} infant is breastfed and given any medication between the $(j-1)^{st}$ and j^{th} measurement occasions.

Clearly, as can be observed from Tables 5.1 and 5.2, the zero-inflated models performed much better, resulting in a substantial improvement in fit, hence implying that the extra zeros need to be accommodated, which is expected given the excessive zero counts in these data as shown in Section 2.1. The ZI(PN-) model is an important improvement, in terms of likelihood, relative to the ZI(P--), while much more improvement is gained in the case of the ZI(P-G). Moreover, considering the ZI(PNG), there is a strong improvement in fit when the gamma and normal random effects, in addition to zero-inflation, are simultaneously included. A similar observation can be made for the non-zero-inflated models. There is a very strong improvement in fit of the ZI(P-G), when compared to the ZI(PN-). It points to the fact that overdispersion is more important an effect than the repeated-measures nature, hence the ZI(P-G) is able to perform better from the start. It underscores, once more, that overdispersion with count data is a very common situation. Eventually, both are needed. The zero-inflation regression coefficients are similar in all models, statistically significant, and can be interpreted as model coefficients for the proportion of extra zeros. ZI(PNG) and ZI(P-G) exhibit similar fits, not only in terms of parameter estimates but also in inference, except that gender is significant in the former ($p = 0.0311$) while this is not the case for the latter ($p = 0.0922$). Both models suggest that medical help, breast feeding, main effect of rural place of residence are significant; the same is true for time interactions with breast feeding and urban place of residence.

5.3.2 Epilepsy Data

We analyze the epilepsy data, introduced in Section 2.4. Let Y_{ij} represent the number of epileptic seizures that patient i experiences during week j of the follow-up period. Also, let t_{ij} be the time-point at which Y_{ij} has been recorded. Consider the combined model (3.16)–(7.3), with parameterization similar to the one in Molenberghs *et al.* (2010), but now accounting for zero inflation, assuming that counts are generated from a (PN-) process with mean λ_{ij}:

$$\ln(\lambda_{ij}) = \begin{cases} (\xi_{00} + b_{1i}) + \xi_{01}t_{ij} & \text{if placebo,} \\ (\xi_{10} + b_{1i}) + \xi_{11}t_{ij} & \text{if treated.} \end{cases} \tag{5.7}$$

or from a (PNG) process with mean $\lambda_{ij} = \theta_{ij}\kappa_{ij}$:

$$\ln(\kappa_{ij}) = \begin{cases} (\xi_{00} + b_{1i}) + \xi_{01}t_{ij} & \text{if placebo,} \\ (\xi_{10} + b_{1i}) + \xi_{11}t_{ij} & \text{if treated,} \end{cases} \tag{5.8}$$

The zero-inflation probability (π_{ij}) is modeled as $\text{logit}(\pi_{ij}) = \gamma_0 + b_{2i} + \gamma_1 t_{ij}$. The data are analyzed with the ZI(PNG), ZI(P-G), ZI(PN-), ZI(P--). For the sake of comparison, also the non-zero-inflated counterparts are fitted. Parameter estimates and predicted probabilities of zeros are presented in Table 5.3. Clearly, in terms of likelihood comparison, the zero-inflated versions performed much better, resulting in a substantial improvement in fit.

The ZI(P-G) is an important improvement relative to the ZI(P--), while much more improvement is gained in the case of the ZI(PN-). Moreover, the ZI(PNG) leads to a substantially improved fit. Further, we observe that, omitting either the overdispersion or the correlation underestimates the predicted probability of zeros, which becomes worse when both are omitted at the same time. The ZI(PNG), fitted without random effects in the zero-inflation part, results in -2log-likelihood of 5386.8, and predicted probability of zeros equal to 0.3271. This implies that inclusion of random effects in the zero-inflation part tends to have little impact on the predicted probability of zeros. However, based on likelihood comparison, model fit improves considerably. This same phenomenon is also evident in the ZI(PN-) fitted with random effects included only in the non-zero count part (-2log-likelihood is 5971.9, and predicted probability of zeros 0.3112).

None of the zero-inflated models suggests evidence of significance in slope difference and slope ratio, except for the ZI(P--), where significance is maintained for the slope difference ($p = 0.004$). However, the latter, unrealistically, omits correlation and overdispersion. The zero-inflation regression coefficients can be interpreted as model coefficients for the proportion of extra zeros, and are statistically significant.

5.4 Simulation Study

In this section, we report on a simulation study set up to examine the bias in estimating the regression parameters when dealing with overdispersed, longitudinal count data with excess zeros. For such data, the bias is likely to result from not appropriately accounting for the excess zero counts, misspecification of the overdispersion, which is a very common situation for count data in a way that the prescribed mean-variance link is violated and misspecification of the correlation results from the repeated-measurements nature of the data.

5.4.1 Simulation Setting

Data are generated along a design inspired by the Jimma Infant Study. Age in months, status of getting medical help, and breast feeding behavior were among the covariates of interest in the study, and are used in the simulation study as well.

We randomly generated 200 data sets from the zero-inflated combined model for 2000 subjects with 10 measurements per subject. The response vector $\boldsymbol{y}_i$ for the i^{th} subject was generated as a correlated and overdispersed count from a negative-binomial process subject to zero-inflation. That is, for each subject, $Y_{ij} \sim \text{NB}(\psi_{ij}, \theta)$, where $\theta = 1$ with $\psi_{ij} = (1 + \kappa_{ij}/\theta)^{-1}$ and where $\kappa_{ij} = \exp\{\xi_0 + b_i + \xi_1 t_{ij} + \xi_2 H_{ij}\}$ for $i = 1, \dots, 2000$ and $j = 1, \dots, 10$. Further, t_{ij} represents the time point at which the j^{th} measurement is recorded for the i^{th} subject and H_{ij} denotes whether or not the i^{th} subject is given any medication help at the j^{th} measurement occasion, generated from a Bernoulli process with $p = 0.9$. Correlation is induced via a subject-specific random intercept b_i generated from a normal distribution with mean 0 and variance 0.8. Then, zero inflation is added by defining the final response vector $\boldsymbol{Y}_i^*$ to have components ${Y_{ij}}^* = (1 - u_{ij})Y_{ij}$, where the u_{ij} are Bernoulli random variables with parameters π_{ij} and $\text{logit}(\pi_{ij}) = \gamma_0 + \gamma_1 t_{ij}$.

Three different scenarios were considered for data generation: S_1: without excess zeros; S_2: with an excess of zeros of around 20%; S_3: with an excess of zeros of roughly 40%. The corresponding total zero percentages are 48%, 68%, and 88%, respectively. This was achieved, for each scenario, by appropriately choosing the zero-inflation coefficients. The true parameter values used to generate the data were $\xi = (1.12, 0.13, -1.89)^T$. Similarly, for the zero-inflation part, $\gamma = (-1, -1)^T$, $\gamma = (1, -0.25)^T$ and $\gamma = (1.8, -0.1)^T$ were used for S_1, S_2, and S_3, respectively.

5.4.2 Simulation Results

The simulated data are analyzed by the ZI(PNG), ZI(P-G), ZI(PN-), and ZI(P--), as well as by their non-zero-inflated counterparts. Mean, relative bias (rbias) and predicted probabilities of zero counts are summarized for the three scenarios in Tables 5.4–7.10, respectively.

Parameter estimates of the ZI(PNG) were in agreement with their true model in all scenarios. This shows that the different components: zero-inflation, overdispersion, and correlation, can be well separated in practice, in settings like the ones considered here. The zero-inflated model converged for almost all simulated sets of data.

Under S_1, as shown in Table 5.4, the ZI(PNG) and the (PNG) performed well and fairly similar in terms of relative bias, except for the intercept ξ_0 for which a larger bias is observed in the (PNG). The percentage of zero counts (48%) is nearly equally predicted in both cases. But, severe impact starts to emerge in the non zero-inflation models when excess zero counts are present, but not accounted for, as evidenced in Tables 7.9 and 7.10. The predicted number of zero counts is largely underestimated in the non-zero-inflated models. When many zeros are allowed for, as in S_3, the effect is more pronounced in the intercept term and the negative-binomial parameter α as compared to S_2. Moreover, the bias in the standard deviation of the random-effects, for instance, in the 'true' model tends to increase in S_3, which gets substantially higher for models with neglected zero-inflation component, such as the (PNG) and (PN-).

The impact of omitting the overdispersion is remarkable. This can be clearly observed, for example, from the considerable increment in the relative bias of the ZI(PN-). When overdispersion is omitted, the zero-inflation component will try to recover part of the overdispersion.

When the correlation stemming from the repeated measurements is misspecified, substantial impact appears in inferences of the ZI(P-G), which gets even worse in the (P-G), as evidenced quite clearly from the larger relative bias of the intercept term. When correlation is omitted from the model, the overdispersion term will try to recover for this misspecification.

Unlike in S_1, the ZI(PNG) significantly beats the (PNG), confirming the importance of accounting for the excess zeros in addition to the repeated measures nature and the overdispersion.

We conclude that failure to account for excess zeros, overdispersion, and/or correlation has a substantial impact on bias and predicted probabilities. This was clearly shown on such key model parameters as the intercept term, the overdispersion parameter, and the variance of the random effects. All scenarios suggest that the zero-inflated combined model is the preferred one in terms of relative bias and predicted probabilities of zeros.

5.5 Discussion

In this chapter, we have described a modeling strategy for a hierarchical count data where excessive zeros correlation and overdispersion can happen together and assembled in one single model. Our work extends Molenberghs *et al.* (2010) who combined gamma and normal random effects to account for overdispersion and correlation. Such extension to further deal with zero-inflatation provides a parsimonious yet useful approach. Molenberghs *et al.* (2007) and Molenberghs *et al.* (2010), brought together normal random effects to induce association between repeated Poisson data, and a gamma distributed random factor in the log-linear predictor to fine-tune the overdispersion. Their model produces the standard negative-binomial and Poisson-normal models as special cases, when there are repeated measures as well as with univariate outcomes.

In terms of estimation, we have focused on maximum likelihood estimation. This can be done by integrating over the random effects, either fully analytically, using the explicit expressions derived, or by combining analytic and numeric techniques. The latter has been implemented in the SAS procedure NLMIXED, for the Poisson, binary, and survival cases, and applied to a case study (Molenberghs *et al.*, 2010).

Of course, with the considerations of not only one but multiple sets of random effects comes the obligation to reflect on the precise nature of such latent structures. As underscored by Verbeke and Molenberghs (2010), full verification of the adequacy of a random-effects structure is not possible based on statistical considerations alone, because there is a many-to-one map from hierarchical models to the implied marginal model. Of course, this should not stop the user from considering such models, but rather issues a word of caution.

Two real data sets with count outcome characterized by zero-inflation, overdispersion and correlation features were studied, one with higher proportion of zeros (Jimma Infant Growth Study) and another with moderate zero percentages (Epilepsy Study). Both case studies suggest that the model assembling all these features at once is the most preferred one. In addition, a simulation study was conducted to further investigate the impact of omitting each or a combination of zero-inflation, overdispersion and correlation. We learned that omitting such features, while actually preset, introduced considerable bias in parameter estimates and hence may lead to incorrect inferences.

Table 5.1: *Jimma Infant Growth Study. Parameter estimates and standard errors for the regression coefficients in (P--), (P-G), (PN-), and (PNG).*

		(P--)	(PN-)	(P-G)	(PNG)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	3.4198(0.0648)	2.0652(0.0744)	3.6443(0.3573)	5.7541(0.4567)
Rural	ξ_1	0.2209(0.0229)	0.2209(0.0291)	0.1674(0.0906)	−0.0733(0.1231)
Urban	ξ_2	−0.1850(0.0331)	−0.5266(0.0399)	−0.1185(0.1157)	−0.3000(0.1600)
Time	ξ_3	−0.1477(0.0073)	−0.1307(0.0078)	−0.1870(0.0425)	−0.3287(0.0506)
Gender	ξ_4	0.1681(0.0182)	0.2478(0.0241)	0.2351(0.0767)	0.2444(0.1041)
Breast feeding	ξ_5	−1.5710(0.0614)	−1.4554(0.0664)	−1.8120(0.3066)	−3.1539(0.4151)
Help	ξ_6	−3.2198(0.0196)	−2.9870(0.0230)	−3.7025(0.1784)	−6.1493(0.1896)
Slope Rural	ξ_7	−0.0085(0.0027)	−0.0090(0.0029)	−0.0033(0.0139)	0.0182(0.0158)
Slope Urban	ξ_8	0.0461(0.0037)	0.0542(0.0039)	0.0397(0.0174)	0.0797(0.0202)
Slope Gender	ξ_9	−0.0011(0.0021)	−0.0061(0.0023)	−0.0033(0.0114)	0.0063(0.0129)
Slope Breast feeding	ξ_{10}	0.1583(0.0069)	0.1441(0.0072)	0.1988(0.0359)	0.3213(0.0453)
Slope Help	ξ_{11}	0.1641(0.0023)	0.1324(0.0081)	0.2326(0.0221)	0.3448(0.0219)
Std. dev random effect	$\sqrt{d}$	—	1.9612(0.0267)	—	1.6847(0.0433)
Negative-binomial parameter	α	—	—	0.0641(0.0009)	0.1045(0.0021)
−2log-likelihood		281,126	203,981	91,370	90,274

Table 5.2: *Jimma Infant Growth Study. Parameter estimates and standard errors for the regression coefficients in ZI(P--), ZI(P-G), ZI(PN-), and ZI(PNG).*

		ZI(P--)	ZI(PN-)	ZI(P-G)	ZI(PNG)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	2.2148(0.0636)	1.3877(0.1205)	2.2200(0.1571)	2.0388(0.1616)
Rural	ξ_1	0.2610(0.0252)	0.3880(0.0400)	0.2536(0.0577)	0.2804(0.0586)
Urban	ξ_2	−0.1049(0.0364)	−0.0945(0.0549)	−0.1096(0.0842)	−0.1301(0.0858)
Time	ξ_3	−0.0289(0.0072)	0.0331(0.0119)	−0.0302(0.0176)	−0.0213(0.0178)
Gender	ξ_4	0.0835(0.0199)	0.1338(0.0321)	0.0797(0.0473)	0.1027(0.0477)
Breast feeding	ξ_5	−0.3430(0.0593)	0.0644(0.1138)	−0.3370(0.1481)	−0.3384(0.1528)
Help	ξ_6	0.2378(0.0211)	0.3312(0.0298)	0.2028(0.0498)	0.2225(0.0507)
Slope Rural	ξ_7	−0.0047(0.0030)	−0.0202(0.0042)	−0.0043(0.0071)	− 0.0060(0.0070)
Slope Urban	ξ_8	0.0222(0.0041)	0.0178(0.0059)	0.0223(0.0096)	0.0227(0.0096)
Slope Gender	ξ_9	−0.0010(0.0023)	−0.0100(0.0032)	−0.0003(0.0056)	− 0.0035(0.0054)
Slope Breast feeding	ξ_{10}	0.0372(0.0066)	− 0.0011(0.0113)	0.0375(0.0164)	0.0345(0.0167)
Slope Help	ξ_{11}	0.0087(0.0059)	0.0019(0.0035)	0.0087(0.0059)	0.0084(0.0058)
Std. dev. non-zero part random effect	$\sqrt{d_1}$	—	0.5856(0.0075)	—	0.4311(0.0112)
Negative-binomial parameter	α	—	—	0.4797(0.0099)	0.2807(0.0086)
Inflation intercept	γ_0	−6.0412(0.6933)	−6.0163(0.5759)	−6.0608(0.6255)	−6.0241(0.5656)
Inflation Rural	γ_1	0.1231(0.0396)	0.1222(0.0467)	0.1331(0.0398)	0.1306(0.0469)
Inflation Urban	γ_2	−0.1380(0.0475)	−0.1578(0.0569)	−0.1368(0.0478)	−0.1578(0.0571)
Inflation Time	γ_3	−0.1835(0.0045)	−0.1941(0.0048)	−0.1834(0.0045)	−0.1942(0.0048)
Inflation Gender	γ_4	−0.1606(0.0328)	−0.1658(0.0388)	−0.1582(0.0329)	−0.1675(0.0389)
Inflation Breast feeding	γ_5	0.2056(0.0814)	0.2394(0.0940)	0.1960(0.0821)	0.2285(0.0945)
Inflation Help	γ_6	9.3894(0.6877)	9.6095(0.5680)	9.3833(0.6192)	9.6145(0.5576)
Std. dev. zero part random effect	$\sqrt{d_2}$	—	0.7575(0.0333)	—	0.7604(0.0335)
Correlation of random effects	ρ	—	−0.0907(0.0402)	—	−0.1127(0.0566)
−2log-likelihood		100,780	80,555	74,489	73,570

Table 5.3: *Epilepsy Study. Parameter estimates and standard error in ZI(P--), ZI(P-G), ZI(PN-), ZI(PNG), (P--), (P-G), (PN-), and (PNG).*

		ZI(PNG)	(PNG)	ZI(P-G)	(P-G)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Intercept placebo	ξ_{00}	0.9467(0.1665)	0.9113(0.1755)	1.2361(0.1100)	1.2594(0.0.1119)
Slope placebo	ξ_{01}	$-0.0162(0.0075)$	$-0.0248(0.0077)$	$-0.0072(0.0113)$	$-0.0126(0.0111)$
Intercept treatment	ξ_{10}	0.8361(0.1716)	0.6557(0.1782)	1.3974(0.1098)	1.4750(0.1093)
Slope treatment	ξ_{11}	$-0.0061(0.0074)$	$-0.0118(0.0075)$	$-0.0219(0.0112)$	$-0.0352(0.0101)$
Negative-binomial parameter	α_1	0.2449(0.0253)	2.4640(0.2113)	1.7874(0.1004)	0.5274(0.0255)
Std. dev. non-zero part random effect	$\sqrt{d_1}$	0.9974(0.0854)	1.0625(0.0871)	–	–
Inflation intercept	γ_0	$-4.5813(0.6405)$	–	$-7.1064(1.3344)$	–
Inflation slope	γ_1	0.0921(0.0339)	–	0.2921(0.0655)	–
Std. dev. zero part random effect	$\sqrt{d_2}$	2.5327(0.4396)	-	–	–
Correlation of random effects	ρ	$-0.0961(0.1534)$	–	–	–
Predicted prob. zeros		0.3522	0.3206	0.1849	0.1583
-2log-likelihood		5317.9	5417.0	6318.9	6326.1

		ZI(PN-)	(PN-)	ZI(P--)	(P--)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Intercept placebo	ξ_{00}	0.9027(0.1552)	0.8179(0.1677)	1.4205(0.0439)	1.2662(0.0424)
Slope placebo	ξ_{01}	$-0.0042(0.0047)$	$-0.0143(0.0044)$	0.0061(0.0045)	$-0.0134(0.0043)$
Intercept treatment	ξ_{10}	0.9078(0.1590)	0.6475(0.1701)	1.7608(0.0402)	1.4531(0.0383)
Slope treatment	ξ_{11}	$-0.0074(0.0045)$	$-0.0120(0.0043)$	$-0.0153(0.0041)$	$-0.0328(0.0038)$
Std. dev. non-zero part random effect	$\sqrt{d_1}$	0.9713(0.0824)	1.0755(0.0857)	–	–
Inflation intercept	γ_0	$-3.7123(0.5003)$	–	$-1.2879(0.1203)$	–
Inflation slope	γ_1	0.0952(0.0249)	–	0.0593(0.0109)	–
Std. dev. zero part random effect	$\sqrt{d_2}$	2.2215(0.3434)	–	–	–
Correlation of random effects	ρ	$-0.1541(0.1574)$	–	–	–
Predicted prob. zeros		0.3384	0.2627	0.3316	0.0459
-2log-likelihood		5845.1	6271.9	9760	11590

Table 5.4: *Simulation study under scenario S_1. Mean, standard error, and relative bias of the parameter estimates in ZI(PNG), ZI(P-G), ZI(PN-), ZI(P--), and its non-zero-inflated counterparts.*

			ZI(PNG)		(PNG)		ZI(P-G)		(P-G)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.068(0.003)	0.046	0.991(0.004)	0.115	1.277(0.003)	0.139	2.404(0.005)	1.147
Time	ξ_1	0.13	0.125(0.001)	0.040	0.136(0.001)	0.046	0.125(0.001)	0.040	0.133(0.001)	0.026
Help	ξ_2	-1.89	$-1.794(0.002)$	0.051	$-1.796(0.002)$	0.049	$-1.705(0.002)$	0.098	$-1.708(0.002)$	0.096
Negative-binomial parameter	α	1.00	0.953(0.002)	0.047	0.995(0.002)	0.005	1.774(0.003)	0.774	0.552(0.001)	0.448
Std. dev random effect	$\sqrt{d}$	0.80	0.780(0.001)	0.025	0.779(0.001)	0.026	–	–	–	–
Inflation intercept	γ_0	-1.00	$-0.856(0.099)$	0.104	–	–	$-0.265(0.123)$	0.725	–	–
Inflation time	γ_1	-1.00	$-1.049(0.098)$	0.049	–	–	$-1.698(0.122)$	0.687	–	–
Predicted prob. zeros		0.48	0.493		0.481		0.359		0.291	
Frequency of convergence			199		200		200		200	

			ZI(PN-)		(P-N)		ZI(P--)		(P--)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.216(0.003)	0.086	0.892(0.003)	0.204	1.661(0.003)	0.483	1.250(0.003)	0.116
Time	ξ_1	0.13	0.101(0.001)	0.225	0.127(0.001)	0.026	0.089(0.001)	0.318	0.124(0.001)	0.043
Help	ξ_2	-1.89	$-1.467(0.002)$	0.224	$-1.693(0.002)$	0.104	$-1.275(0.002)$	0.326	$-1.682(0.002)$	0.109
Std. dev random effect	$\sqrt{d}$	0.80	0.796(0.001)	0.005	0.861(0.001)	0.076	–	–	–	–
Inflation intercept	γ_0	-1.00	$-0.386(0.005)$	0.614	–	–	0.247(0.003)	1.247	–	–
Inflation time	γ_1	$-.00$	$-0.094(0.001)$	0.906	–	–	$-0.094(0.001)$	0.906	–	–
Predicted prob. zeros		0.48	0.473		0.365		0.483		0.255	
Frequency of convergence			200		200		200		200	

Table 5.5: *Simulation study under scenario S_2. Mean, standard error, and relative bias of the parameter estimates in ZI(PNG), ZI(P-G), ZI(PN-), ZI(P--), and its non-zero-inflated counterparts.*

			ZI(PNG)		(PNG)		ZI(P-G)		(P-G)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.079(0.004)	0.037	1.833(0.005)	0.637	1.089(0.005)	0.027	2.796(0.006)	1.497
Time	ξ_1	0.13	0.123(0.001)	0.052	0.239(0.001)	0.839	0.125(0.001)	0.040	0.225(0.001)	0.730
Help	ξ_2	−1.89	−1.766(0.003)	0.066	−1.776(0.003)	0.060	−1.671(0.003)	0.116	−1.703(0.003)	0.099
Negative-binomial parameter	α	1.00	0.908(0.004)	0.093	0.372(0.001)	0.628	2.379(0.008)	1.379	0.266(0.001)	0.734
Std. dev random effect	$\sqrt{d}$	0.80	0.772(0.002)	0.035	0.754(0.002)	0.058	–	–	–	–
Inflation intercept	γ_0	1.00	1.056(0.005)	0.056	–	–	0.993(0.006)	0.003	–	–
Inflation time	γ_1	−0.25	−0.246(0.001)	0.014	–	–	−0.354(0.001)	0.416	–	–
Predicted prob. zeros		0.68	0.696		0.398		0.549		0.367	
Frequency of convergence			200		200		200		200	

			ZI(PN-)		(P-N)		ZI(P--)		(P--)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.183(0.004)	0.056	−0.235(0.004)	1.210	1.666(0.004)	0.488	0.215(0.004)	0.808
Time	ξ_1	0.13	0.099(0.001)	0.235	0.212(0.001)	0.633	0.087(0.001)	0.329	0.210(0.001)	0.613
Help	ξ_2	−1.89	−1.444(0.003)	0.236	−1.679(0.003)	0.112	−1.261(0.002)	0.420	−1.664(0.003)	0.120
Std. dev random effect	$\sqrt{d}$	0.80	0.834(0.001)	0.042	0.976(0.001)	0.220	–	–	–	–
Inflation intercept	γ_0	1.00	1.473(0.004)	0.473	–	–	1.816(0.003)	0.816	–	–
Inflation time	γ_1	−0.25	−0.209(0.001)	0.163	–	–	−0.202(0.001)	0.193	–	–
Predicted prob. zeros		0.68	0.677		0.520		0.682		0.422	
Frequency of convergence			200		200		200		200	

Table 5.6: *Simulation study under scenario S_3. Mean, standard error, and relative bias of the parameter estimates in ZI(PNG), ZI(P-G), ZI(PN-), ZI(P--), and its non-zero-inflated counterparts.*

			ZI(PNG)		(PNG)		ZI(P-G)		(P-G)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.076(0.007)	0.039	4.005(0.009)	2.828	0.980(0.009)	0.125	4.494(0.008)	3.012
Time	ξ_1	0.13	0.125(0.001)	0.042	0.216(0.001)	0.658	0.121(0.001)	0.070	0.202(0.001)	0.554
Help	ξ_2	−1.89	−1.757(0.005)	0.070	−1.765(0.005)	0.067	−1.676(0.005)	0.113	−1.701(0.005)	0.100
Negative-binomial parameter	α	1.00	0.887(0.007)	0.112	0.088(0.001)	0.912	3.041(0.034)	2.041	0.076(0.001)	0.924
Std. dev random effect	$\sqrt{d}$	0.80	0.765(0.003)	0.043	0.609(0.004)	0.239	–	–	–	–
Inflation intercept	γ_0	1.80	1.862(0.006)	0.034	–	–	1.487(0.009)	0.174	–	–
Inflation time	γ_1	−0.10	−0.102(0.001)	0.017	–	–	−0.118(0.001)	0.177	–	–
Predicted prob. zeros		0.88	0.884		0.590		0.807		0.604	
Frequency of convergence			200		200		200		200	

			ZI(PN-)		(P-N)		ZI(P--)		(P--)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.051(0.008)	0.061	−1.515(0.007)	2.353	1.660(0.006)	0.482	−0.631(0.007)	1.563
Time	ξ_1	0.13	0.104(0.001)	0.203	0.195(0.001)	0.502	0.088(0.001)	0.323	0.193(0.001)	0.487
Help	ξ_2	−1.89	−1.473(0.005)	0.221	−1.669(0.005)	0.117	−1.257(0.004)	0.335	−1.661(0.005)	0.121
Std. dev random effect	$\sqrt{d}$	0.80	0.941(0.002)	0.176	1.416(0.002)	0.769	–	–	–	–
Inflation intercept	γ_0	1.80	2.205(0.005)	0.225	–	–	2.629(0.004)	0.382	–	–
Inflation time	γ_1	−0.10	−0.112(0.001)	0.122	–	–	−0.127(0.001)	0.271	–	–
Predicted prob. zeros		0.88	0.876		0.756		0.877		0.675	
Frequency of convergence			200		200		200		200	

Chapter 6

Marginalized Multilevel Hurdle and Zero-Inflated Models for Overdispersed and Correlated Count Data with Excess Zeros

In statistical modeling of hierarchical non-Gaussian outcomes, such as clustered or longitudinal count data, generalized linear mixed models (GLMMs) are very popular (Engel and Keen, 1992; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005). In most practical settings, such hierarchically organized count data are characterized not only by association, resulting from the repeated measures per subject or clustering of observations with in a subject, but also by overdispersion and excessive zero features. In Chapter 5, a ZI(PNG) model was studied to simultaneously deal with correlation, overdispersion and excessive zeros simultaneously. In this modeling strategy, we assume that zeros may come from two sources: from a point-mass, which generates only zeros, and from a Poisson-normal-gamma model, where counts are generated from a Poisson-normal-gamma process. An alternative route to deal with excessive zeros in hierarchcial and overdispersed count data is to combine the hurdle idea of Mullahy (1986) with (PNG) model of Molenberghs *et al.* (2010). The hurdle specification is based on a two-part conditional model, using a zero mass and a truncated-at-zero count distribution. Adjustment based on the hurdle specification proceeds by using a truncated Poisson-normal-gamma process for non-zero counts, leading to H(PNG) model as a hurdle counterpart of ZI(PNG) of Chapter 5.

The GLMM (Engel and Keen, 1992; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005) and its extention for overdispersion (Molenberghs *et al.*, 2010) and data hierarchy and zero-inflation (Min and Agresti, 2005) do not automatically provide population averaged interpretation for regression parameters, while such results are very often needed in practice. A marginalized multilevel model (MMM) is proposed by Heagerty (1999) and Heagerty and Zeger (2000) by simultaneously defining a marginal mean and a conditional mean by making use of so-called connector function, yielding marginally interpretable covariate effects. Iddi and Molenberghs (2012) extend the the combined modeling concepts of Molenberghs *et al.* (2010) with marginalized multilevel model (MMM) idea of Heagerty (1999) and proposed a marginalized combined model. Furthermore, the connections between bridge distributions, marginalized multilevel models, and generalized linear mixed models is explored in Molenberghs *et al.* (2013), by placing particular attention on binary and count data, for several commonly used link function choices.

In this Chapter, we will employ the combined model idea of Molenberghs *et al.* (2010) and the marginalized multilevel model of Heagerty (1999), in conjunction with concepts of hurdle or zero-inflated models, and we will present a marginalized hurdle combined model as well as a marginalized zero-inflated combined model, as two alternative modeling strategies for overdispersed and correlated count data with excess zeros. The former was also studied by Lee *et al.* (2011), where the logit link function was used for the zero-inflation. We considered both logit and probit link functions, whereby the latter leads to closed-form expressions. In addition, instead of using only one of the logit or probit links, we make use of them simultaneously (Griswold and Zeger, 2004), and specify a logit link for the marginal model, and a probit for the conditional model, so that the odds ratio interpretation is still retained, while taking computational advantage of the probit link. Population averaged interpenetration is possible not only for the positive count component, but also for zero-inflation component. This chapter is organized as follows. In Section 6.1, the marginalized version of H(PNG) model, denoted as MH(PNG) is described, followed by marginalized ZI version, MZI(PNG) model in Section 6.2, with the estimation technique given in Section 6.3. The MH(PNG) and MZI(PNG) and their special cases, are applied and compared based on three real data sets, namely: the An. mosquito data sets collected through IRC and CDC techniques, and the Jimma Longitudinal Family Survey of Youth, as described in Sections 2.2.1, 2.2.2, 2.3, respectively, with the results presented in Section 6.4. Finally, some concluding remarks are provided in Section 6.5. The contribution of this chapter is based on Kassahun *et al.* (2014b).

6.1 Marginalized Hurdle Combined Model

Merging ideas of the combined model of (Molenberghs *et al.*, 2010) and the hurdle model (Mullahy, 1986), a two-part hurdle combined model is considered to deal with zero-inflated overdispersed clustered count data. While the first part models only the zero state with probability π_{ij}^c or π_{ij}^m, the second part handles non-zero counts, which are assumed to follow a truncated-at-zero probability mass function, such as, in this case, a truncated Poisson-normal-gamma model. The superscripts c and m are used to emphasize the conditional and marginal nature of the specifications to follow. The zero-inflation component π_{ij}^c is modeled using a Bernoulli model, through an appropriate link function, such as the probit or logit, potentially containing known regressors as well as random effects. For clustered binary data, it is well documented that, unlike the logit link, the probit link leads to closed form solutions (Zeger *et al.*, 1988; Griswold and Zeger, 2004; Molenberghs *et al.*, 2010). This leads to a conditional model specified as:

$$p(Y_{ij}=y_{ij}|\boldsymbol{b}_i,\boldsymbol{\xi},\theta_{ij},\phi,\pi_{ij}^c)=\begin{cases}\pi_{ij}^c & \text{if } y_{ij}=0,\\ (1-\pi_{ij}^c)\frac{f_i(y_{ij}|\lambda_{ij}^c,\theta_{ij})}{1-f_i(0|\lambda_{ij}^c,\theta_{ij})} & \text{if } y_{ij}>0,\end{cases}$$

where $\pi_{ij}^c = \Phi(\Delta_{ij1}+\boldsymbol{z}_{ij1}'\boldsymbol{b_{i}}_1)$, $\lambda_{ij}^c = \theta_{ij}\exp(\Delta_{ij2}+\boldsymbol{z}_{ij2}'\boldsymbol{b_{i}}_2)$, $\boldsymbol{b_i}=(\boldsymbol{b_i}_1,\boldsymbol{b_i}_2)' \sim N(\boldsymbol{0},D)$, and $\theta_{ij}\sim\text{Gamma}(\alpha,\beta)$. Further, Δ_{ij1} and Δ_{ij2} are connector functions of the zero part and the positive count part, corresponding to the random vectors $\boldsymbol{b_i}_1$ and $\boldsymbol{b_i}_2$ and regressors $\boldsymbol{z}_{ij1}$ and $\boldsymbol{z}_{ij2}$, respectively. The marginal specification is

$$p(Y_{ij}=y_{ij})=\begin{cases}\pi_{ij}^m & \text{if } y_{ij}=0,\\ (1-\pi_{ij}^m)\frac{f_i(y_{ij}|\lambda_{ij}^m)}{1-f_i(0|\lambda_{ij}^m)} & \text{if } y_{ij}>0,\end{cases}$$

where $\text{logit}(\pi_{ij}^m)=\boldsymbol{x}_{ij1}'\boldsymbol{\gamma}^m$ and $\ln(\lambda_{ij}^m)=\boldsymbol{x}_{ij2}'\boldsymbol{\xi}^m$, with known regressors $\boldsymbol{x}_{ij1}$ and $\boldsymbol{x}_{ij2}$ and a vector of zero-inflation coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$. Based on Griswold and Zeger (2004) and Iddi and Molenberghs (2012), specifying a logit link for the marginal model and a probit link for the conditional model leads to computational advantages from the probit-normal relationship, with the marginal parameters still having the odds-ratio interpretation. Hence, the connector functions, as shown in Iddi and Molenberghs (2012) are as follows. For the logit:

$$\Delta_{ij1\ell}=\sqrt{1+\boldsymbol{z}_{ij1}'D\boldsymbol{z}_{ij1}}\,\Phi^{-1}[\text{expit}(\boldsymbol{x}_{ij1}'\boldsymbol{\gamma}^m)],$$

with

$$\text{expit}(\boldsymbol{x}'_{ij1}\boldsymbol{\gamma}^m) = \int \Phi(\Delta_{ij1} + \boldsymbol{z}'_{ij1}\boldsymbol{b_{i}}_1) f(\boldsymbol{b_{i}}_1) d\boldsymbol{b_{i}}_1;$$

for the probit:

$$\Delta_{ij1p} = \sqrt{1 + \boldsymbol{z}'_{ij1} D \boldsymbol{z}_{ij1}} \Phi^{-1}[\Phi(\boldsymbol{x}'_{ij1}\boldsymbol{\gamma}^m)].$$

In line with Section 3.6, the connector function for the positive counts part is:

$$\Delta_{ij2} = \ln E(\theta_{ij}) + \boldsymbol{x}'_{ij2}\boldsymbol{\xi}^m - \frac{1}{2}\boldsymbol{z}'_{ij2} D \boldsymbol{z}_{ij2}. \tag{6.1}$$

In cases where a simple random-intercept model is adequate, i.e., when $\boldsymbol{b_{1i}} = b_{1i}$, $\boldsymbol{b_{2i}} = b_{2i}$, and $\boldsymbol{z}_{1ij} = \boldsymbol{z}_{2ij} = 1$, D takes the simple form:

$$D = \begin{pmatrix} d_1^2 & \rho d_1 d_2 \\ \rho d_1 d_2 & d_2^2 \end{pmatrix}. \tag{6.2}$$

We denote the conditional hurdle combined model by H(PNG)$_\ell$, H(PNG)$_p$, and the marginal as MH(PNG)$_\ell$, MH(PNG)$_p$, where the subscripts 'ℓ' and 'p' refer to the logit and probit link functions used in the zero inflation part, respectively, and 'M' stands for the marginalized version. Some special cases, such as the MH(PN-)$_\ell$, M(PNG), (PNG), M(PN-), and (PN-) immediately follow.

6.2 Marginalized Zero-Inflated Combined Model

As a counterpart to the hurdle formulation of the previous section, the zero-inflated model assumes that data are generated from two processes as a mixture, in line with Section 3.7. Recall that the first process generates only zeros, while count observations are generated from the second (Lambert, 1992; Greene, 1994). These ideas can be well extended to the combined model of Molenberghs *et al.* (2010) from Section 3.5, assuming a mixing probability π^c_{ij} of zeros from process one and counts from a Poisson-gamma-normal process with probability $1 - \pi^c_{ij}$. This leads to a conditional zero-inflated combined model given by:

$$p(Y_{ij} = y_{ij} | \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}, \phi, \pi^c_{ij}) = \begin{cases} \pi^c_{ij} + (1 - \pi^c_{ij}) f_i(0 | \lambda^c_{ij}, \theta_{ij}) & \text{if } y_{ij} = 0, \\ (1 - \pi^c_{ij}) f_i(y_{ij} | \lambda^c_{ij}, \theta_{ij}) & \text{if } y_{ij} > 0. \end{cases}$$

The marginal formulation is:

$$p(Y_{ij} = y_{ij}) = \begin{cases} \pi^m_{ij} + (1 - \pi^m_{ij}) f_i(0 | \lambda^m_{ij}) & \text{if } y_{ij} = 0, \\ (1 - \pi^m_{ij}) f_i(y_{ij} | \lambda^m_{ij}) & \text{if } y_{ij} > 0, \end{cases}$$

where π_{ij}^c, λ_{ij}^c, $\boldsymbol{b}_i$, π_{ij}^m, λ_{ij}^m, as well as their corresponding connector functions, $\Delta_{ij1\ell}$, Δ_{ij1p} and Δ_{ij2}, have similar expressions as those presented in Section 6.1.

The notational convention employed is as outlined at the end of the previous section, with now ‘ZI’ replacing ‘H’. For example, the conditional zero-inflated combined model is denoted as ZI(PNG)$_\ell$ or ZI(PNG)$_p$, depending on the first-process link function used.

6.3 Estimation

Consider first the models with zero-inflation but without marginalization. Then, the probability resulting from H(PNG)$_\ell$ or H(PNG)$_p$, marginal over θ_{ij} but still conditional upon the normal random effect $\boldsymbol{b}_i$, is:

$$f(y_{ij}|\boldsymbol{b}_i,\boldsymbol{\xi},\theta_{ij},\phi,\pi_{ij}) = I(y_{ij}=0)\pi_{ij}^c + (1-\pi_{ij}^c)g_1(\boldsymbol{b}_i), \tag{6.3}$$

where

$$g_1(\boldsymbol{b_i}) = \left(\begin{array}{c}\alpha_j + y_{ij} - 1\\ \alpha_j - 1\end{array}\right)\cdot\left(\frac{\beta_j}{1+\kappa_{ij}^c\beta_j}\right)^{y_{ij}}\cdot\left(\frac{1}{1+\kappa_{ij}^c\beta_j}\right)^{\alpha_j}\kappa_{ij}^{c\;y_{ij}}\cdot\frac{1}{1-\left(\frac{1}{1+\kappa_{ij}^c\beta_j}\right)^{\alpha_j}},$$

either

$$\text{logit}(\pi_{ij}^c) = \boldsymbol{x}_{1ij}'\boldsymbol{\gamma} + \boldsymbol{z}_{ij1}'\boldsymbol{b_i}_1 \tag{6.4}$$

or

$$\text{probit}(\pi_{ij}^c) = \boldsymbol{x}_{1ij}'\boldsymbol{\gamma} + \boldsymbol{z}_{ij1}'\boldsymbol{b_i}_1, \tag{6.5}$$

and

$$\kappa_{ij}^c = \exp\left(\boldsymbol{x}_{2ij}'\boldsymbol{\xi} + \boldsymbol{z}_{ij2}'\boldsymbol{b_i}_2\right). \tag{6.6}$$

The likelihood function for H(PNG) is given by:

$$L(\boldsymbol{\xi},\boldsymbol{\gamma},D,\phi) = \prod_{i=1}^{N}\int\prod_{j=1}^{n_i}\pi_{ij}^c(\boldsymbol{b_i})^{I(y_{ij}=0)}\left\{(1-\pi_{ij}^c)(\boldsymbol{b_i})g_1(\boldsymbol{b_i})\right\}^{1-I(y_{ij}=0)} \tag{6.7}$$
$$\times\phi(\boldsymbol{b_i}|D)d\boldsymbol{b_i}.$$

Because (6.7) does not have a closed-form solution, we propose the application of numerical techniques to obtain the maximum likelihood estimates, using the adaptive Gauss-Hermite quadrature in SAS NLMIXED. Equivalent flexible non-linear mixed model optimizers can be employed as well, of course.

Turning to the MH(PNG)$_\ell$ or MH(PNG)$_p$, we make use of the connector functions $\Delta_{ij1\ell}$, Δ_{ij1p}, and/or Δ_{ij2}, as shown in Section 6.1. The above approach for maximum likelihood estimation can be used, upon replacing $\boldsymbol{x}'_{1ij}\boldsymbol{\gamma}$ with $\Delta_{ij1\ell}$ or Δ_{ij1p} in (6.4) and (6.5), respectively, and $\boldsymbol{x}'_{2ij}\boldsymbol{\xi}$ with Δ_{ij2} in (6.6).

For the ZI(PNG)$_\ell$ or ZI(PNG)$_p$, we slightly modify (6.3) to

$$f(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}, \phi, \pi_{ij}) = I(y_{ij}=0)\pi^c_{ij} + (1-\pi^c_{ij})g_2(\boldsymbol{b}_i), \tag{6.8}$$

with likelihood

$$\begin{aligned} L(\boldsymbol{\xi}, \boldsymbol{\gamma}, D, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} & \Big\{ \pi^c_{ij}(\boldsymbol{b_i}) + (1-\pi_{ij})^c(\boldsymbol{b_i}) g_3(\boldsymbol{b_i}) \Big\}^{I(y_{ij}=0)} \\ & \times \Big\{ (1-\pi^c_{ij})(\boldsymbol{b_i}) g_2(\boldsymbol{b_i}) \Big\}^{1-I(y_{ij}=0)} \phi(\boldsymbol{b_i}, D) d\boldsymbol{b_i}, \end{aligned}$$

where

$$\begin{aligned} g_2(\boldsymbol{b_i}) &= \begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1 + \kappa^c_{ij}\beta_j} \right)^{y_{ij}} \cdot \left(\frac{1}{1 + \kappa^c_{ij}\beta_j} \right)^{\alpha_j} {\kappa^c_{ij}}^{y_{ij}}, \\ g_3(\boldsymbol{b_i}) &= \left(\frac{1}{1 + \kappa^c_{ij}\beta_j} \right)^{\alpha_j}. \end{aligned}$$

Here also, it is straightforward to get MZI(PNG)$_\ell$ or MZI(PNG)$_p$, by employing the connector functions and making similar replacements in π^c_{ij} and κ^c_{ij}, like that in the hurdle combined model. The SAS implementation for the more general situations are given in Appendix C.

6.4 Data Analysis

6.4.1 IRC Data

We will analyze the IRC data as described in Section 2.2.1. Let Y_{ij} represent the number of An. gambiae counts for house i during month j of the follow-up period. Also, let t_{ij} be the time point (months) at which Y_{ij} has been measured, $t_{ij} = 1, 2, \ldots n_i$ until at most $n_i = 32$, and s_{ij} denote season coded as (1: wet; 0: dry) for house i during month j. Further, v_i is the village of house i coded as (1: at risk; 0: control). We transformed t_{ij} to $t'_{ij} = t_{ij}/12$ (years). The marginal mean model for the Poisson process is given by

$$\ln(\kappa^m_{ij}) = \xi_0 + \xi_1 t'_{ij} + \xi_2 v_i + \xi_3 s_{ij} + \xi_4 t'_{ij} v_i.$$

The combined model assuming that counts are generated from a Poisson-normal-gamma process has mean $\lambda^c_{ij} = \theta_{ij}\kappa_{ij}$ with $\theta_{ij} \sim \text{Gamma}(\alpha, 1/\alpha)$. The marginal

Table 6.1: *IRC Data. Parameter estimates (standard errors) for the regression coefficients in (1) ZI(PNG)$_\ell$, (2) MZI(PNG)$_\ell$ with logit link for zero-inflation.*

		ZI(PNG)$_\ell$	MZI(PNG)$_\ell$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	−0.0085(0.1639)	0.0340(0.1594)
Time	ξ_1	0.0061(0.0988)	0.0051(0.0983)
Village	ξ_2	0.9824(0.1688)	0.9851(0.1684)
Season	ξ_3	2.2418(0.0982)	2.2427(0.0979)
Village×Time	ξ_4	−0.0524(0.1163)	−0.0528(0.1159)
Overdispersion	α	1.4472(0.1126)	1.4390(0.1112)
Std. dev. random intercept count	d_1	0.2856(0.0677)	0.2840(0.0669)
Inflation intercept	γ_0	2.2771(0.1634)	2.0152(0.1475)
Inflation time	γ_1	0.0172(0.0568)	0.0118(0.0495)
Inflation village	γ_2	−1.0145(0.1486)	−0.8975(0.1287)
Inflation season	γ_3	−1.3816(0.1039)	−1.2174(0.0917)
Std. dev. random intercept inflation	d_2	0.8597(0.0814)	0.5107(0.0471)
Corr. random effects	ρ	−0.4583(0.2058)	−0.4430(0.2084)
−2log-likelihood		12,817	12,815
AIC		12,843	12,841

model for the zero-inflation probability is modeled as a function of time, village, and season:

$$F(\pi_{ij}^m) = \gamma_0 + \gamma_1 t'_{ij} + \gamma_2 v_i + \gamma_3 s_{ij},$$

where $F(\cdot)$ is either the logit or probit function. The corresponding conditional models introduce a normally distributed random intercept, b_{1i} in the Poisson model, and b_{2i} in the binomial model, such that the random effects are assumed jointly normal and possibly correlated ρ, the variance-covariance matrix is given by (6.2).

The data were first analyzed using the zero-inflation models ZI(PNG)$_\ell$, ZI(PNG)$_p$, MZI(PNG)$_\ell$, and MZI(PNG)$_p$. Results are shown in Tables 6.1 and 6.2. Comparing ZI(PNG)$_\ell$ and MZI(PNG)$_\ell$ as well as ZI(PNG)$_p$ and MZI(PNG)$_p$, parameter estimates and the corresponding standard errors of the count part appear similar except for slight differences in ξ_0. This follows from the nature of the connector function.

Table 6.2: *IRC Data. Parameter estimates (standard errors) for the regression coefficients in (1) ZI(PNG)$_p$, (2) MZI(PNG)$_p$ with probit link for zero-inflation.*

		ZI(PNG)$_p$	MZI(PNG)$_p$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	−0.0080(0.1640)	0.0338(0.1598)
Time	ξ_1	0.0098(0.0989)	0.0090(0.0989)
Village	ξ_2	0.9816(0.1687)	0.9797(0.1687)
Season	ξ_3	2.2426(0.0983)	2.2423(0.0983)
Village×Time	ξ_4	−0.0553(0.1164)	−0.0541(0.1164)
Overdispersion	α	1.4442(0.1119)	1.4441(0.1119)
Std. dev. random intercept count	d_1	0.2835(0.0669)	0.2834(0.0669)
Inflation intercept	γ_0	1.3387(0.0936)	1.1917(0.0833)
Inflation time	γ_1	0.0152(0.0340)	0.0135(0.0303)
Inflation village	γ_2	−0.5954(0.0879)	−0.5300(0.0772)
Inflation season	γ_3	−0.8099(0.0592)	−0.7210(0.0527)
Std. dev. random intercept inflation	d_2	0.5119(0.0474)	0.5120(0.0474)
Corr. random effects	ρ	−0.4394(0.2097)	−0.4388(0.2098)
−2log-likelihood		12,817	12,817
AIC		12,843	12,843

However, estimates corresponding to the zero-inflation component, such as γ_0, γ_2, γ_3 and d_2 show some difference. This is expected due to the change of link function. For example, the ratio of the logit-based parameters and their probit-based counterparts is almost everywhere close to $16/15 \cdot \pi/\sqrt{3} \simeq 1.70$ (Molenberghs and Verbeke, 2005). Further, the marginalized counterparts, for zero-inflation, employ a combination of logit and probit links, explaining that marginalization induces differences in the ZI portion of the models. ZI(PNG)$_\ell$, ZI(PNG)$_p$, MZI(PNG)$_\ell$, and MZI(PNG)$_p$ produce very similar fits, as follows from both the deviance and the AIC. While models are formally non-nested, they have the same number of parameters, so the deviance offers some indication as well. As expected, the zero-inflation estimates are affected by the link function used. Observe that ZI(PNG)$_p$ and MZI(PNG)$_p$ yield very similar estimates for d_2, as shown in Table 6.2, which is also true for MZI(PNG)$_\ell$, but the corresponding estimate in ZI(PNG)$_\ell$ is relatively larger as given in Table 6.1. This same phenomenon appears for the correlation parameter ρ. In terms of parameter significance, all four models give similar results such that all parameters are found to be significant, except time and village-time interaction in the counts, and time effect in the zero-inflation. In addition, all four models suggest that standard deviations of the random intercepts of the positive counts and the excess zeros, overdispersion parameter, zero-inflation intercept and zero-inflation coefficients of village and season are statistically significant, implying strong evidence for all phenomena: correlation stemming from the data hierarchy, overdispersion, and excess zeros beyond what the (PNG) allows for. In addition, the correlation parameter ρ is negative and statistically significant across the models, suggesting that the two processes generating counts on the one hand and merely zeros on the other are in an inverse relationship.

Second, as an alternative modeling strategy, the hurdle models H(PNG)$_\ell$, H(PNG)$_p$, MH(PNG)$_\ell$, and MH(PNG)$_p$ were fitted. Results are shown in Tables 6.3 and 6.4. Similar to what was observed for the ZI models, some differences are observed among H(PNG)$_\ell$ and MH(PNG)$_\ell$ as well as among H(PNG)$_p$ and MH(PNG)$_p$ in the ZI parameter estimates, while the estimates of the count part remain similar, except for the intercept. Again, based on AIC and deviance, all four models fit the data equivalently and suggest strong evidence of correlation, overdispersion and excess zeros. Also here, ρ is negative.

While few new messages emanate from the hurdle models next to their ZI counterparts, or vice versa, the agreement is actually comforting and can be seen as a simple form of sensitivity analysis. That said, there is a small reduction in fit statistics in the hurdle version, which evidently may not always be the case. Parameter estimates as well as the associated inferences for the count process are similar, except for a small difference in the value of ξ_0. However, estimates for the zero-inflation part, such as γ_0, γ_2, γ_3, and d_2 show some differences, even with the same link function, which stems from the fact that the ZI models combine a model for the count with an atom at zero as a mixture, while the hurdle model separately handles the zero observations and the positive counts. Min and Agresti (2005) showed that the hurdle model works well both in zero-inflation and zero-deflation settings, while zero-inflated models are suitable only for handling zero-inflation. These authors further indicated that, when a data set is subject to zero-deflation at a level of a factor, the estimates of the corresponding parameter may be unstable in ZI models. To investigate this, we make use of a special form of the hurdle model, also known as the zero-altered model, which requires the same covariates as well as the same distributional forms in the two parts, as suggested by Min and Agresti (2005). The model is:

$$\begin{aligned}\ln[-\ln(1-\pi_{ij})] &= a_1 + a_2(\xi_0 + \xi_1 t_{ij} + \xi_2 v_i + \xi_3 s_{ij}),\\ \ln(\kappa_{ij}^m) &= \xi_0 + \xi_1 t_{ij} + \xi_2 v_i + \xi_3 s_{ij}.\end{aligned}$$

By setting $a_2 = 1$, and testing whether $a_1 = 0$, one can test for zero-inflation. If $a_1 < 0$ ($a_1 > 0$), then the data are zero-inflated (zero-deflated). Fitting the zero-altered combined model shows strong evidence of zero-inflation ($a_1 = -2.0771$, likelihood ratio test statistic 13 on 1 degree of freedom, $p < 0.001$). On the other hand, Todem *et al.* (2012) propose an extension of the ZI models to handle both zero-inflation and deflation. Details can be found in Todem *et al.* (2012). Briefly, in our context, the proposal is based on a suitable transformation of π_{ij} to, say, ζ_{ij} of the form $\pi_{ij} = (\zeta_{ij} - f(0))(1 - f(0))^{-1}$, where f is a Poisson-normal-gamma model, and ζ_{ij} is specified as function of covariates, $\zeta_{ij} = (1 + \alpha\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma}))^{\frac{-1}{\alpha}}$, and α, $\boldsymbol{x}_{ij}$ and $\boldsymbol{\gamma}$ as before. We applied this extended zero-modified model on the IRC data to test for zero-inflation. Again, strong evidence of zero-inflation results ($a_1 = -2.6901$, likelihood ratio test statistic 186 on 1 degree of freedom, $p < 0.001$). Todem *et al.* (2012) noted that the mixing probability π_{ij} is allowed to take both positive (zero-inflation) and negative (zero-deflation) values only when specified marginally, and negative π_{ij} do not allow for hierarchical interpretation of the mixture model, as π_{ij} have a probability definition in the latter. In general and in line with common wisdom, it is suggested that when choosing among zero-inflated and hurdle models, in addition to statistical fit criteria, it is prudent to reflect upon the data generation processes. In this regard, if zeros are expected from both parts, then the zero-inflated model could be preferred (Neelon *et al.*, 2010). Min and Agresti (2005) listed a number of advantages of the hurdle model: it works well both in zero-inflation and zero-deflation situations, as stated above; it can be used to test for evidence of zero-inflation; it is easier to fit as it separately handles the count and the zero processes.

Generally, the marginalized models led to estimates relatively superior in precision. The MZI(PNG)$_\ell$ and MH(PNG)$_\ell$ models, as given in Tables 6.1 and 6.3, respectively, provide marginally meaningful estimates, with the additional advantage of an odds-ratio interpretation for the ZI parameters. Further, MH(PNG)$_\ell$ results in the smallest AIC and deviance values, though differences are modest. The MH(PNG)$_\ell$ model as shown in Table 6.3 suggests that villages at risk had higher expected An. gambaie log-counts (1.0553, $p < 0.0001$) as compared to the controls. Further, log-counts in the wet season were higher than in the dry season (2.2496, $p < 0.0001$). However, no statistically significant association was observed for time effect ($p = 0.5703$); the same is true for the village-time interaction ($p = 0.4109$). The zero-inflation estimate

Table 6.3: *IRC Data. Parameter estimates (standard errors) for the regression coefficients in (1) H(PNG)$_\ell$, (2) MH(PNG)$_\ell$ with logit link for zero-inflation.*

		H(PNG)$_\ell$	MH(PNG)$_\ell$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	$-0.0823(0.1666)$	$-0.0423(0.1626)$
Time	ξ_1	$0.0548(0.0987)$	$0.0561(0.0986)$
Village	ξ_2	$1.0548(0.1704)$	$1.0553(0.1703)$
Season	ξ_3	$2.2479(0.0970)$	$2.2496(0.0972)$
Village×Time	ξ_4	$-0.0958(0.1182)$	$-0.0974(0.1181)$
Overdispersion	α	$1.4343(0.1105)$	$1.4348(0.1105)$
Std. dev. random intercept count	d_1	$0.2888(0.0687)$	$0.2878(0.0686)$
Inflation intercept	γ_0	$2.9793(0.1402)$	$2.6807(0.1303)$
Inflation time	γ_1	$0.0094(0.0488)$	$0.0067(0.0438)$
Inflation village	γ_2	$-1.1114(0.1361)$	$-0.9970(0.1205)$
Inflation season	γ_3	$-1.7893(0.0841)$	$-1.6221(0.0783)$
Std. dev. random intercept inflation	d_2	$0.8006(0.0707)$	$0.4630(0.0401)$
Corr. random effects	ρ	$-0.4732(0.1882)$	$-0.4708(0.1890)$
−2log-likelihood		12,814	12,810
AIC		12,840	12,836

corresponding to village ($\widehat{\gamma}_2 = -0.9970$, $p < 0.0001$) with $\exp(\widehat{\gamma}_2) = 0.37$ implies that the odds of zeros in the at-risk villages is nearly one third of what is expected in the control villages. In addition, it was found that the odds of zeros in the wet season is much smaller than that of the dry season ($\widehat{\gamma}_3 = -1.6221$, $p < 0.0001$). The correlation of the random effects is negative and significant ($\widehat{\rho} = -0.4708$, $p = 0.0137$), suggesting the presence of a strong inverse relationship between the count and zero-inflation processes. Further, both processes are influenced by covariates such as season and village, differently, in such a way that parameter estimates corresponding to the positive counts are positive in sign, while negative for the zero-inflation part. Altogether, these results suggest that village type (at risk versus control), classified based on distance from the dam and season (wet versus dry), belong to the potential operating factors affecting An. gambaie density.

Two special cases were considered for comparison's purpose. These are (PN-)

Table 6.4: *IRC Data. Parameter estimates (standard errors) for the regression coefficients in (1) H(PNG)$_p$, (2) MH(PNG)$_p$ with probit link for zero-inflation.*

		H(PNG)$_p$	MH(PNG)$_p$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	$-0.0851(0.1670)$	$-0.0435(0.1627)$
Time	ξ_1	$0.0560(0.0986)$	$0.0560(0.0986)$
Village	ξ_2	$1.0566(0.1704)$	$1.0566(0.1704)$
Season	ξ_3	$2.2499(0.0972)$	$2.2498(0.0972)$
Village×Time	ξ_4	$-0.0973(0.1181)$	$-0.0973(0.1181)$
Overdispersion	α	$1.4347(0.1105)$	$1.4347(0.1105)$
Std. dev. random intercept count	d_1	$0.2881(0.0686)$	$0.2881(0.0686)$
Inflation intercept	γ_0	$1.7058(0.0780)$	$1.5473(0.0697)$
Inflation time	γ_1	$0.0144(0.0286)$	$0.0131(0.0259)$
Inflation village	γ_2	$-0.6358(0.0785)$	$-0.5767(0.0702)$
Inflation season	γ_3	$-1.0208(0.0463)$	$-0.9260(0.0426)$
Std. dev. random intercept inflation	d_2	$0.4639(0.0402)$	$0.4639(0.0402)$
Corr. random effects	ρ	$-0.4742(0.1888)$	$-0.4742(0.1888)$
−2log-likelihood		12,812	12,812
AIC		12,838	12,838

and (PNG), with their marginalized counterparts; recall that the first omits overdispersion and ZI, while the second omits ZI only. Results are shown in Tables 6.5 and 6.6. Clearly, we observe that both models fit the data poorly as compared to the MZI(PNG)$_\ell$, MZI(PNG)$_p$, MH(PNG)$_\ell$ and MH(PNG)$_p$ models. Because of the marginal interpretation, M(PN-) and M(PNG) can be compared to MH(PNG)$_\ell$. We see that both fixed-effect and variance component estimates were severely affected due to the simplified models' misspecification; evidently also, inferences are affected. For example, the village effect in M(PN-) is 2.6861 ($p < 0.0001$), which is above twofold of the estimate from MH(PNG)$_\ell$. In addition, the intercept ξ_0, season effect ξ_3, and standard deviation of the random effects d_1 are highly affected. We observe similar issues for the M(PNG), even though it is more general than the M(PN-). Impact is severe on ξ_0, ξ_2, ξ_3, and d_1. As a result, based on M(PNG), ξ_0 is highly significant, and based on M(PN-), ξ_0 are ξ_1 are highly significant, which was not the case with the

Table 6.5: *IRC Data. Parameter estimates (standard errors) for the regression coefficients in (1) (PNG), (2) M(PNG).*

		(PNG)	M(PNG)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	−2.3547(0.2113)	−1.8221(0.2051)
Time	ξ_1	−0.1320(0.1095)	−0.1321(0.1095)
Village	ξ_2	1.6262(0.2592)	1.6264(0.2592)
Season	ξ_3	3.2641(0.1030)	3.2640(0.1030)
Village×Time	ξ_4	0.0783(0.1436)	0.0782(0.1436)
Std. dev. random intercept count	d_1	1.0320(0.1086)	1.0319(0.1085)
Overdispersion	α	7.7575(0.3102)	7.7567(0.3102)
−2log-likelihood		13,188	13,188
AIC		13,202	13,202

MZI(PNG)$_\ell$ and MH(PNG)$_\ell$. Further, the estimate of the overdispersion parameter ($\widehat{\alpha} = 7.7567$) in the M(PNG) is relatively large as compared to the corresponding value of MH(PNG)$_\ell$, underscoring strong zero-inflation present in the data. Because of this, α is trying to recover from this misspecification. In addition, from Table 6.6, parameter estimates of GEE are affected as a result of these misspecification, and appear similar to the corresponding estimates of M(PN-), except ξ_2, though the estimates in M(PN-) are superior in precision. Altogether, these results imply that when dealing with longitudinal count data where correlation, overdispersion, and excess zeros are likely to appear at the same time, failure to model these three features simultaneously can have a serious impact on the marginal parameter estimates. Inevitably, incorrect inferences may follow.

Predicted probability of zeros among the control and at risk villages is computed for (PN-); M(PNG); ZI(PNG)$_\ell$, MZI(PNG)$_\ell$, ZI(PNG)$_p$, MZI(PNG)$_p$; H(PNG)$_\ell$, MH(PNG)$_\ell$, H(PNG)$_p$, MHPNG)$_p$. The Observed percentage of zeros and the ones predicted by the aforementioned models are summarized in Table 6.7. Clearly, the hurdle versions lead to predicted probabilities almost the same as the observed ones. The ZI versions are also doing well, though the zero percentages are slightly under predicted. These results, once again, imply that the hurdle models are doing better in these data, in line with what has been suggested by the model fit statistics. For

Table 6.6: *IRC Data. Parameter estimates (standard errors) for the regression coefficients in (1) (PN-), (2) M(PN-), (3) Generalized estimating equations (GEE).*

		(PN-)	M(PN-)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	−3.1636(0.1321)	−2.1254(0.1739)
Time	ξ_1	0.0484(0.0189)	0.0484(0.0189)
Village	ξ_2	2.6862(0.0926)	2.6861(0.0926)
Season	ξ_3	3.0704(0.0331)	3.0704(0.0331)
Village×Time	ξ_4	−0.0231(0.0212)	−0.0232(0.0212)
Std. dev. random intercept count	d_1	1.4409(0.0943)	1.4410(0.0943)
−2log-likelihood		53,791	53,791
AIC		53,803	53,803

		GEE
Effect	Parameter	Estimate (s.e.)
Intercept	ξ_0	−2.0809(0.1836)
Time	ξ_1	0.0763(0.0635)
Village	ξ_2	1.6040(0.1722)
Season	ξ_3	3.0716(0.1144)
Village×Time	ξ_4	−0.0588(0.0747)

(PN-) and M(PNG), both of which ignore excessive zeros, the percentage of zeros are poorly predicted.

The observed and predicted percentage of zeros by MH(PNG)$_\ell$ model and MZI(PNG)$_\ell$ model versus month of collection are displayed in Figure 6.1, and for M(PNG) Model and M(PN-) Model, are plotted in Figure 6.3. Clearly, MH(PNG)$_\ell$ model and MZI(PNG)$_\ell$ model behave very similar and predict percentage of zeros well. This same phenomenon also holds true for MH(PNG)$_\ell$ model and MZI(PNG)$_\ell$ model, as shown in Figure 6.2. However, M(PNG) model and M(PN-) model predict very poorly in such a way that the percentages of zeros are highly under predicted.

Table 6.8 shows predicted mean and standard deviation by M(PN-), M(PNG), MZI(PNG)$_\ell$ with logit link for zero-inflation, MZI(PNG)$_p$ with probit link for zero-inflation, MH(PNG)$_\ell$ with logit link for zero-inflation, MHPNG)$_p$ with probit link for

Table 6.7: *IRC Data. Observed and Predicted Probability of Zeros in M(PN-); M(PNG); ZI(PNG)$_\ell$, MZI(PNG)$_\ell$; ZI(PNG)$_p$, MZI(PNG)$_p$; H(PNG)$_\ell$, MH(PNG)$_\ell$; H(PNG)$_p$, MHPNG)$_p$.*

	control		at risk	
Model	observed	predicted	observed	predicted
M(PN-)	83.6	58.0503	69.2	32.3362
M(PNG)	83.6	55.5497	69.2	34.6017
ZI(PNG)$_\ell$	83.6	81.9047	69.2	66.4544
MZI(PNG)$_\ell$	83.6	82.0633	69.2	66.6333
ZI(PNG)$_p$	83.6	81.9591	69.2	66.6191
MZI(PNG)$_p$	83.6	81.9600	69.2	66.6193
H(PNG)$_\ell$	83.6	83.8038	69.2	69.3673
MH(PNG)$_\ell$	83.6	83.9369	69.2	69.5185
H(PNG)$_p$	83.6	83.8615	69.2	69.5688
MH(PNG)$_p$	83.6	83.8615	69.2	69.5691

zero-inflation. Clearly, M(PN-) and M(PNG) over predict the mean, especially in the at risk village. This is likely because these models under predict zeros, which in turn will affect the mean. The hurdle and ZI models predict the mean very well, with the standard deviations getting smaller as compared to the observed ones. This is mainly because the extra variability in the observed data is captured by the overdispersion parameter, the random effects as well as the zero-inflation component.

6.4.2 CDC Data

In this section, results of the CDC An. mosquito data, as described in Section 2.2.2, are presented. We considered ZI(PNG)$_\ell$, MZI(PNG)$_\ell$, ZI(PNG)$_p$, and MZI(PNG)$_p$, H(PNG)$_\ell$, MH(PNG)$_\ell$, H(PNG)$_p$, and MH(PNG)$_p$. With the same model parameterizations like Section 6.4.1, the results for ZI(PNG)$_\ell$, MZI(PNG)$_\ell$ and ZI(PNG)$_p$, MZI(PNG)$_p$ are shown in Table 6.9 and Table 6.10, respectively. Similarly, the results for H(PNG)$_\ell$, MH(PNG)$_\ell$ and H(PNG)$_p$, MH(PNG)$_p$ are shown in Table 6.11 and Table 6.12, respectively. Similar to the IRC data, the hurdle versions resulted in smallest fit statistics. Again, in line with MH(PNG)$_\ell$ of the IRC data in Table 6.3, the CDC data in Table 6.11 also suggested that at risk village and wet season resulted in higher

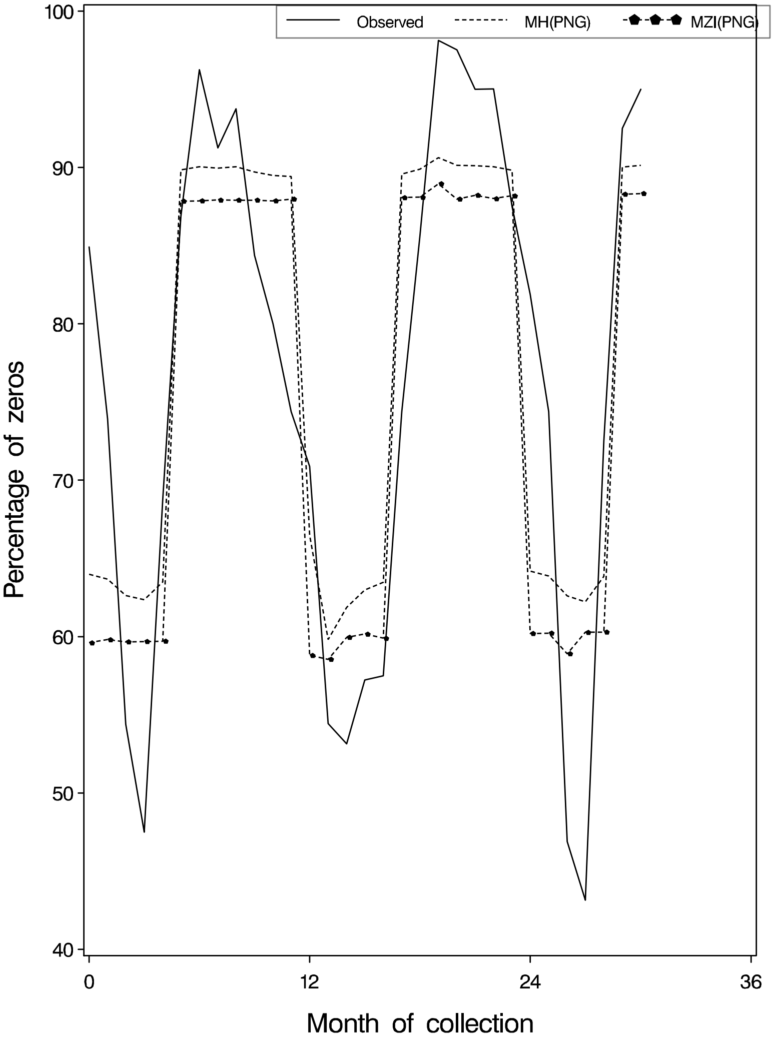

Figure 6.1: *IRC Data. Observed and predicted percentage of zeros by MH(PNG)$_\ell$ and MZI(PNG)$_\ell$ versus Month of collection.*

expected An. gambaie log-counts. These results obtained from two different collection types, once again, underscore that village type and season are important covariates of An. gambaie density around the Gilgel-Gibe dam. Furthermore, overdispersion, standard deviations of the positive counts and excessive zero parts, and correlation parameters were found to be statistically significant. MH(PNG)$_\ell$ in the CDC data of Table 6.11 suggest that the intercept, time and time-village interaction were not

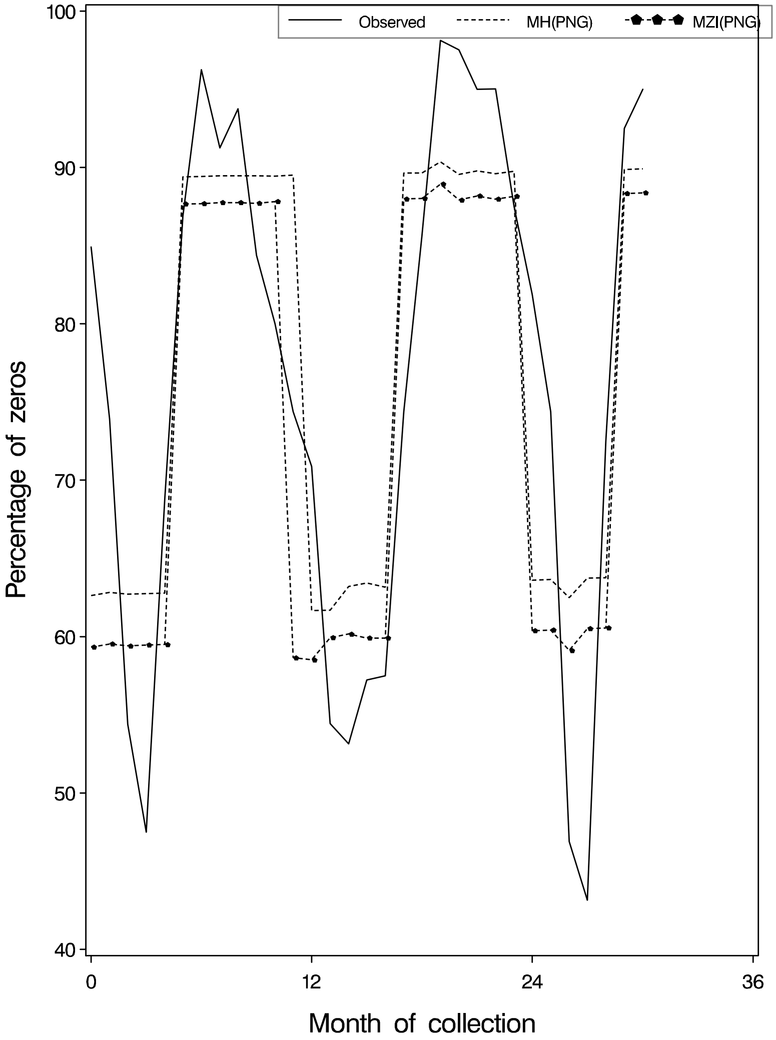

Figure 6.2: *IRC Data. Observed and predicted percentage of zeros by MH(PNG)$_p$ and MZI(PNG)$_p$ versus Month of collection.*

statistically significant, and this is similar to what we have obtained in Table 6.3 of the IRC data. The only exception is that, while the effect of time in the zero-inflation part γ_1 in the CDC data is significant, this is not the case in the IRC collection.

As shown in Tables 6.13 and 6.14, the special cases, M(PN-) and M(PNG), not only fit the data poorly relative to MH(PNG)$_\ell$ in Table 6.11 and MH(PNG)$_p$ in Table 6.12, but also the corresponding parameter estimates and the associated inferences are

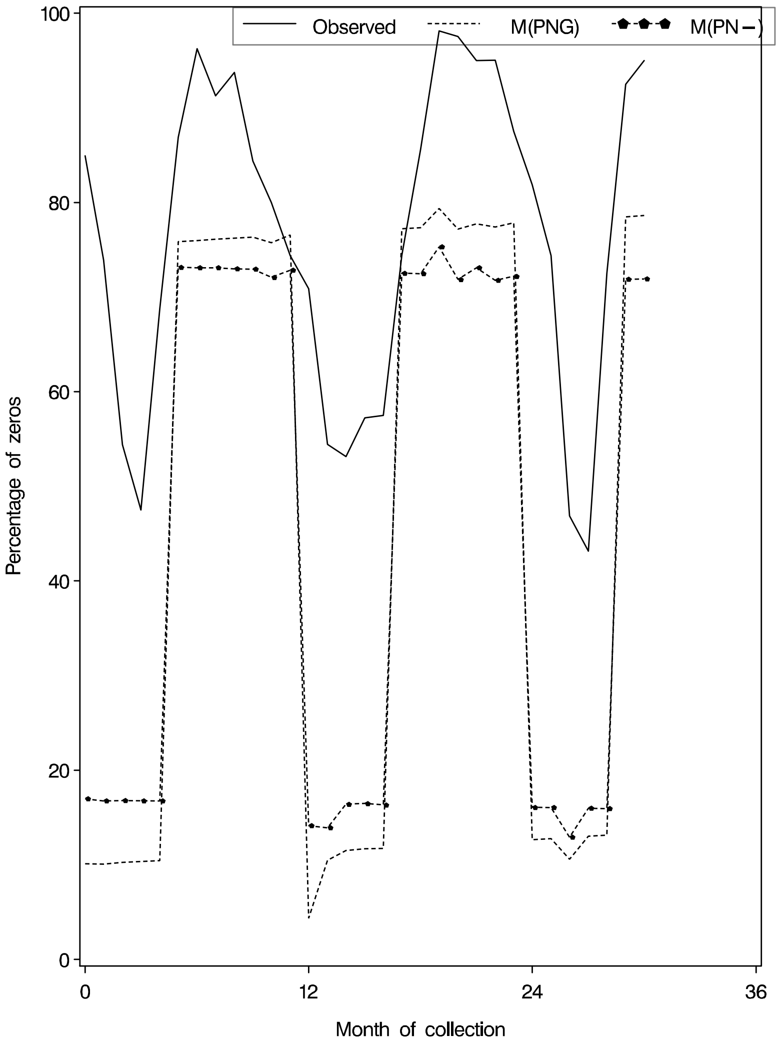

Figure 6.3: *IRC Data. Observed and predicted percentage of zeros by M(PNG) and M(PN-) versus Month of collection.*

highly affected, similar to the IRC data, as shown in Section 6.4.1. Furthermore, from Table 6.14, we observe the impact of this misspecification on the GEE estimates as well, which appear similar with the corresponding estimates of M(PN-) model, thought the latter led to estimates better in precision. These results, once again, convey the important message that overdispersion and/or ZI are common phenomenon of An. mosquito count data, whereby omitting any one of them or both may lead to

Table 6.8: *IRC Data. Observed and Predicted Mean and Standard deviation in M(PN-), M(PNG), MZI(PNG)$_\ell$, MZI(PNG)$_p$, MH(PNG)$_\ell$, MHPNG)$_p$.*

	control		at risk	
	Mean	Std. dev.	Mean	Std. dev.
Observed	1.616	6.427	6.916	18.760
M(PN-)	1.379	1.300	19.764	18.542
M(PNG)	1.829	1.776	10.183	9.742
MZI(PNG)$_\ell$	1.516	1.446	6.565	6.136
MZI(PNG)$_p$	1.548	1.482	6.504	6.062
MH(PNG)$_\ell$	1.513	1.435	6.756	6.349
MH(PNG)$_p$	1.540	1.473	6.685	6.254

erroneous conclusion.

Table 6.15 contains the observed percentage of zeros and the ones predicted by (PN-); M(PNG); ZI(PNG)$_\ell$, MZI(PNG)$_\ell$, ZI(PNG)$_p$, MZI(PNG)$_p$; H(PNG)$_\ell$, MH(PNG)$_\ell$, H(PNG)$_p$, MHPNG)$_p$. In line with results of the IRC data in Table 6.7, the hurdle versions are doing better than the ZI models, where the zero percentages are again under predicted in the latter.

6.4.3 Jimma Longitudinal Family Survey of Youth

Turning to the Jimma Longitudinal Family survey of Youth, described in Section 2.3, let Y_{ij} represent the number of days of work for subject i during time j (year). Also, let $t_{ij} = 1, 2, 3$ be the time point (in years) at which Y_{ij} has been measured, a_{ij} the age (in years) of subject i at time j, and s_i sex (1: male; 0: female). The marginal mean models for the Poisson process and marginal ZI probability are given by:

$$\begin{aligned} \ln(\kappa_{ij}^m) &= \xi_0 + \xi_1 t_{ij} + \xi_2 s_i + \xi_3 a_{ij}, \\ \text{logit}(\pi_{ij}^m) &= \gamma_0 + \gamma_1 t_{ij} + \gamma_2 s_i + \gamma_3 a_{ij}. \end{aligned}$$

Results from fitting the ZI(PN-)$_\ell$ and MZI(PN-)$_\ell$ are shown in Table 6.17, and from the H(PN-)$_\ell$ and MH(PN-)$_\ell$ in Table 6.18. Likely because of lack of overdispersion, the (PNG) and M(PNG) do not fit well. Further, in contrast to the previous case study, there is a much smaller number of repeated measures per subject in this case.

Table 6.9: *CDC Data. Parameter estimates (standard errors) for the regression coefficients in (1) ZI(PNG)$_\ell$, (2) MZI(PNG)$_\ell$ with logit link for zero-inflation.*

		ZI(PNG)$_\ell$	MZI(PNG)$_\ell$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	−0.2440(0.3696)	−0.1484(0.3635)
Time	ξ_1	−0.2219(0.1524)	−0.2233(0.1525)
Village	ξ_2	1.0292(0.3629)	1.0261(0.3617)
Season	ξ_3	1.2046(0.2183)	1.2067(0.2188)
Village×Time	ξ_4	−0.1291(0.1937)	−0.1302(0.1939)
Overdispersion	α	2.1351(0.3558)	2.1654(0.3596)
Std. dev. random intercept count	d_1	0.4533(0.1308)	0.4507(0.1306)
Inflation intercept	γ_0	2.4515(0.4972)	2.1370(0.4345)
Inflation time	γ_1	−1.0794(0.2975)	−0.9728(0.2668)
Inflation village	γ_2	−0.4539(0.4839)	−0.3488(0.4287)
Inflation season	γ_3	−2.5478(0.4781)	−2.2458(0.4037)
Std. dev. random intercept inflation	d_2	0.8957(0.2767)	0.5391(0.1629)
Corr. random effects	ρ	−0.7667(0.3811)	−0.7738(0.3786)
−2log-likelihood		2495.2	2494.9
AIC		2521.2	2520.9

This phenomenon is further scrutinized by fitting and comparing ZI(P--)$_\ell$ with ZI(P-G)$_\ell$, as well as H(P--)$_\ell$ with H(P-G)$_\ell$. Estimates of the overdispersion parameter, its p-value, and model fit statistics are summarized in Table 6.16. The larger p-values of the overdispersion term as well as the poor model fit statistic of ZI(P-G)$_\ell$ and H(P-G)$_\ell$, compared to the ZI(P--)$_\ell$ and H(P--)$_\ell$, suggest the relative unimportance of the overdispersion term, once the excess zeros have been taken in to account. This implies that the normal random effects, combined with either a hurdle or zero-inflation component, are sufficient to describe the data. The random effects' standard deviation change substantially among conditional and marginal models, as can be seen from both Tables 6.17 and 6.18, leading to non-negligible differences in the zero-inflation estimates as well. MZI(PNG)$_\ell$ and MH(PNG)$_\ell$ suggest similar inferences for all covariate effects, such that all were found to be statistically significant, except for the correlation ρ. We observed that, though the effect of time is positive both in

Table 6.10: *CDC Data. Parameter estimates (standard errors) for the regression coefficients in (1) ZI(PNG)$_p$, (2) MZI(PNG)$_p$ with probit link for zero-inflation.*

		ZI(PNG)$_p$	MZI(PNG)$_p$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	$-0.2430(0.3676)$	$-0.1426(0.3613)$
Time	ξ_1	$-0.2244(0.1526)$	$-0.2242(0.1525)$
Village	ξ_2	$1.0273(0.3620)$	$1.0273(0.3620)$
Season	ξ_3	$1.2027(0.2177)$	$1.2029(0.2177)$
Village×Time	ξ_4	$-0.1286(0.1939)$	$-0.1288(0.1939)$
Overdispersion	α	$2.1623(0.3553)$	$2.1611(0.3550)$
Std. dev. random intercept count	d_1	$0.4491(0.1304)$	$0.4489(0.1303)$
Inflation intercept	γ_0	$1.4713(0.2877)$	$1.2954(0.2499)$
Inflation time	γ_1	$-0.6499(0.1705)$	$-0.5718(0.1480)$
Inflation village	γ_2	$-0.2777(0.2869)$	$-0.2446(0.2517)$
Inflation season	γ_3	$-1.5339(0.2722)$	$-1.3503(0.2275)$
Std. dev. random intercept inflation	d_2	$0.5375(0.1616)$	$0.5374(0.1616)$
Corr. random effects	ρ	$-0.7824(0.3758)$	$-0.7816(0.3760)$
−2log-likelihood		2494.6	2494.6
AIC		2520.6	2520.6

the non-zero counts and the extra zeros, the covariates age and sex have a positive effect on the non-zero counts model, while negative in the zero-inflation component. Based on MH(PN-)$_\ell$, male adolescents have higher involvement in work as compared to their female counterparts ($p = 0.0023$) and higher age has a positive effect for work involvement ($p = 0.018$).

Next, we omitted the zero-inflation components and considered (PN-) and M(PN-). Results are shown in Table 6.19. Clearly, considering likelihood and AIC values, these models fit the data poorly, implying that accommodating for extra zeros cannot be circumvented. Further, parameter estimates for ξ_0, ξ_2, ξ_3, and d_1 are highly affected by this misspecification.

The observed percentages of zeros among the three years and those predicted by (PN-); M(PN-); ZI(PN-)$_\ell$, MZI(PN-)$_\ell$; H(PN-)$_\ell$, MH(PN-)$_\ell$ are summarized in Table 6.20. We observe that both the ZI and hurdle model versions give similar

Table 6.11: *CDC Data. Parameter estimates (standard errors) for the regression coefficients in (1) H(PNG)$_\ell$, (2) MH(PNG)$_\ell$ with logit link for zero-inflation.*

		H(PNG)$_\ell$	MH(PNG)$_\ell$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	−0.5412(0.4766)	−0.4783(0.4806)
Time	ξ_1	−0.1685(0.1950)	−0.1688(0.1950)
Village	ξ_2	1.1770(0.4123)	1.1962(0.4119)
Season	ξ_3	1.2189(0.2459)	1.2222(0.2460)
Village×Time	ξ_4	−0.3213(0.2524)	−0.3220(0.2524)
Overdispersion	α	3.7028(1.5452)	3.7105(1.5558)
Std. dev. random intercept count	d_1	0.3970(0.1482)	0.3959(0.1492)
Inflation intercept	γ_0	2.9483(0.2870)	2.7615(0.2714)
Inflation time	γ_1	−0.4095(0.1037)	−0.3787(0.0951)
Inflation village	γ_2	−0.8590(0.2787)	−0.8275(0.2587)
Inflation season	γ_3	−1.9473(0.1706)	−1.8112(0.1593)
Std. dev. random intercept inflation	d_2	0.6544(0.1281)	0.3892(0.0748)
Corr. random effects	ρ	−0.8324(0.2834)	−0.8348(0.2838)
−2log-likelihood		2488.9	2487.7
AIC		2514.9	2513.7

results, and hence are working well equally. These results further strengthen the fact that the hurdle and the ZI models fit to the data equivalently, observing the similar fit statistics values, as shown in Tables 6.17 and 6.18.

6.5 Discussion

In this chapter, we have presented marginalized modeling strategies for hierarchical count data, characterized by correlation, overdispersion, and excess zeros: a marginalized hurdle Poisson-normal-gamma model and a marginalized zero-inflated Poisson-normal. These models bring together the marginalization concept that Heagerty (1999) applied to multilevel models, adjustment for an excess of zero counts based on the hurdle model or the zero-inflated model (Mullahy, 1986; Lambert, 1992) and

Table 6.12: *CDC Data. Parameter estimates (standard errors) for the regression coefficients in (1) H(PNG)$_p$, (2) MH(PNG)$_p$ with probit link for zero-inflation.*

		H(PNG)$_p$	MH(PNG)$_p$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	−0.5439(0.4772)	−0.4648(0.4796)
Time	ξ_1	−0.1696(0.1950)	−0.1696(0.1950)
Village	ξ_2	1.1774(0.4125)	1.1773(0.4125)
Season	ξ_3	1.2221(0.2461)	1.2222(0.2461)
Village×Time	ξ_4	−0.3207(0.2525)	−0.3297(0.2525)
Overdispersion	α	3.7090(1.5502)	3.7090(1.5502)
Std. dev. random intercept count	d_1	0.3979(0.1486)	0.3979(0.1486)
Inflation intercept	γ_0	1.7363(0.1626)	1.6189(0.1480)
Inflation time	γ_1	−0.2452(0.0612)	−0.2286(0.0570)
Inflation village	γ_2	−0.4912(0.1638)	−0.4580(0.1515)
Inflation season	γ_3	−1.1469(0.0965)	−1.0693(0.0901)
Std. dev. random intercept inflation	d_2	0.3877(0.0745)	0.3878(0.0745)
Corr. random effects	ρ	−0.8319(0.2838)	−0.8319(0.2837)
−2log-likelihood		2488.6	2488.6
AIC		2514.6	2514.6

the combined modeling framework for overdispersion and correlation (Molenberghs *et al.*, 2010). Two normally distributed random-effects vectors, possibly correlated, were included such that one of them captures the correlation in the positive counts profile, while the other does the same in the excess zero model component. The correlation between these random effects can be interpreted as the correlation between the data generation processes of the zero state and the positive counts. A marginal, population-averaged interpretation is possible not only for the positive counts part, but also for the zero-inflation component, where the latter has the usual odds ratio interpretation. Link functions, such as the logit or probit link, can be used for the zero-inflation part. The marginal density using the latter has a closed-form solution, while this is not the case for the former and an iterative numerical approximation may be required. Based on Griswold and Zeger (2004), instead of using only one of the logit or the probit links, we considered a logit link for the marginal model,

Table 6.13: *CDC Data. Parameter estimates (standard errors) for the regression coefficients in (1) (PNG), (2) M(PNG).*

		(PNG)	M(PNG)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	−1.9183(0.3157)	−1.6405(0.3172)
Time	ξ_1	0.1109(0.1513)	0.1109(0.1513)
Village	ξ_2	0.9664(0.3915)	0.9664(0.3915)
Season	ξ_3	2.2044(0.1685)	2.2044(0.1685)
Village×Time	ξ_4	−0.0612(0.2025)	−0.0612(0.2025)
Std. dev. random intercept count	d_1	0.7455(0.1355)	0.7454(0.1355)
Overdispersion	α	3.7465(0.3297)	3.7465(0.3297)
−2log-likelihood		2540.0	2540.0
AIC		2554.0	2554.0

and a probit one for the conditional model, thus retaining the odds-ratio interpretation of the covariate effects, while taking computational advantage of the probit link. Though these models seem relatively complex, they can be conveniently and effectively implemented in available software packages, such as the SAS NLMIXED procedure.

We analyzed the IRC and CDC data described in Section 2.2.1 and Section 2.2.2, respectively. Both the MH(PNG) model and MZI(PNG) model worked well. The two models differ mainly in their zero-inflation parameter estimates, which is likely because of the difference in the way they handle extra zeros. Furthermore, the MH(PNG) model performed better in prediction of the zero percentages than that of MZI(PNG) in both the IRC and CDC data when prediction is made in terms of village type (at risk versus control). On the other hand, both models performed similar in predicating percentage of zeros with month of collection. This difference is likely because the effect of village type on percentage of zeros is much stronger than that of time. The data analysis showed that distance from the dam and season were the two main operating factors of the An. gambaie density around the Gilgel-Gibe hydroelectric dam. However, when the excess zeros and overdispersion are unrealistically omitted, the covariate time showed a significant association, though this was not the case based on the MH(PNG)$_\ell$, which was the most preferred one. On the other hand, in

Table 6.14: *CDC Data. Parameter estimates (standard errors) for the regression coefficients in (1) (PN-), (2) M(PN-), (3) Generalized estimating equations (GEE).*

		(PN-)	M(PN-)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	$-2.0430(0.2394)$	$-1.6740(0.2402)$
Time	ξ_1	$-0.0225(0.0577)$	$-0.0225(0.0577)$
Village	ξ_2	$1.5194(0.3023)$	$1.5194(0.3023)$
Season	ξ_3	$2.1517(0.0762)$	$2.1517(0.0762)$
Village×Time	ξ_4	$-0.2882(0.0677)$	$-0.2882(0.0677)$
Std. dev. random intercept count	d_1	$0.8590(0.1174)$	$0.8590(0.1174)$
-2log-likelihood		4892.8	4892.8
AIC		4904.8	4904.8

		GEE
Effect	Parameter	Estimate (s.e.)
Intercept	ξ_0	$-1.6875(0.3747)$
Time	ξ_1	$-0.0425(0.1747)$
Village	ξ_2	$1.4569(0.4348)$
Season	ξ_3	$2.1545(0.2282)$
Village×Time	ξ_4	$-0.2223(0.2331)$

the presence of modest overdispersion, as observed in the Jimma Longitudinal Family Survey of Youth, it appears that overdispersion and the excess zero aspects may not be well separated, and hence the Poisson-normal GLMM with only zero-inflation adjustment was sufficient. This might be due to the relatively modest overdispersion present in the data, once accounting for excess zeros. Relatively larger sample sizes and higher number of repeated measures per subject (like in the IRC data), evidently facilitate model fitting. In addition, analysis of the case studies showed that the model omitting the ZI feature, i.e., M(PN-) model, predicted the proportion of zeros poorly as compared to the MH(PN-) model and MZI(PN-) model.

Building upon Molenberghs *et al.* (2010), we argue that the normal and non-normal random effects, the latter often of a gamma type, can usefully be integrated together into a single model to induce association between repeated Poisson data and

Table 6.15: *CDC Data. Observed and Predicted Probability of Zeros in (PN-); M(PNG); ZI(PNG)$_\ell$, MZI(PNG)$_\ell$; ZI(PNG)$_p$, MZI(PNG)$_p$; H(PNG)$_\ell$, MH(PNG)$_\ell$; H(PNG)$_p$, MHPNG)$_p$.*

	control		at risk	
Model	observed	predicted	observed	predicted
M(PN-)	75.3	56.9345	62.7	37.3252
M(PNG)	75.3	56.5168	62.7	39.8834
ZI(PNG)$_\ell$	75.3	65.4972	62.7	52.4098
MZI(PNG)$_\ell$	75.3	65.2125	62.7	52.1915
ZI(PNG)$_p$	75.3	65.3748	62.7	52.0533
MZI(PNG)$_p$	75.3	65.3838	62.7	52.0639
H(PNG)$_\ell$	75.3	75.5102	62.7	62.8561
MH(PNG)$_\ell$	75.3	75.6469	62.7	63.0136
H(PNG)$_p$	75.3	75.4138	62.7	63.1263
MH(PNG)$_p$	75.3	75.4138	62.7	63.1263

Table 6.16: *Jimma Longitudinal Family Survey of Youth. Overdispersion parameter, and model fit statistics for ZI(P--)$_\ell$, ZI(P-G)$_\ell$, H(P--)$_\ell$, and H(P-G)$_\ell$.*

	ZI(P--)$_\ell$	ZI(P-G)$_\ell$	H(P--)$_\ell$	H(P-G)$_\ell$
Overdispersion (p_{value})	-	0.0037(0.7470)	-	0.0062(0.6227)
-2log-likelihood	13302	13417	13302	14751
AIC	13318	13435	13318	14769

to correct for the overdispersion, in addition to the ‘Hurdle’ or the ‘ZI’ adjustments to account for an excess of zeros. We considered AIC and deviance statistics for model comparison. As a result, the hurdle models showed better fit to the data, and lead to parameter estimates relatively superior in precision. Further, in terms of predicted percentage of zeros, in general, the hurdle versions performed better, regardless of the link function used, either probit or logit. In a univariate setting, one might employ a likelihood based approach as proposed by Vuong (1989) for comparison of non-nested models, like H(P--) and ZI(P--). Min and Agresti (2005), based on their simulation

Table 6.17: *Jimma Longitudinal Family Survey of Youth. Parameter estimates (standard errors) for the regression coefficients in (1) ZI(PN-)$_\ell$, (2) MZI(PN-)$_\ell$.*

		ZI(PN-)$_\ell$	MZI(PN-)$_\ell$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	0.8620(0.1440)	0.8685(0.1441)
Time	ξ_1	0.0673(0.0179)	0.0672(0.0179)
Sex	ξ_2	0.0840(0.0272)	0.0842(0.0273)
Age	ξ_3	0.0288(0.0092)	0.0288(0.0092)
Std. dev. random intercept count	d_1	0.1197(0.0321)	0.1198(0.0321)
Inflation intercept	γ_0	2.2627(0.3858)	2.0132(0.3472)
Inflation time	γ_1	0.1862(0.0463)	0.1634(0.0417)
Inflation sex	γ_2	−0.4227(0.0731)	−0.3807(0.0654)
Inflation age	γ_3	−0.0605(0.0251)	−0.0540(0.0225)
Std. dev. random intercept inflation		0.8010(0.0689)	0.4719(0.0403)
Corr. random effects	ρ	−0.1329(0.2888)	−0.1334(0.2944)
−2log-likelihood		13,242	13,242
AIC		13,264	13,264

study, mentioned a number of advantages of the hurdle model: it works well both in zero-inflation and zero-deflation situations, it can be used to test for evidence of zero-inflation, it is easier to fit as it separately handles the count part and the zero part. On the other hand, the extended ZI model based on Todem *et al.* (2012) can handle both zero-inflation and zero-deflation, though the proposed link function does not have the computational flexibility of the probit link. Furthermore, this model can be used to assess assumption of excessive zeros in a given data, and might lead to comparable results with the zero-altered model studied by Min and Agresti (2005). One might expect a marked difference among the hurdle and the ZI models when the percentage of excess zeros is moderate or low, whereby the former is likely to perform better. In general, it is suggested that in choosing among zero-inflated and hurdle models, in addition to statistical fit criteria, , such as, deviance and AIC, it is better to consider the data generation processes. According to Neelon *et al.* (2010), if zeros are expected from both parts, a zero-inflated model is preferred.

Table 6.18: *Jimma Longitudinal Family Survey of Youth. Parameter estimates (standard errors) for the regression coefficients in (1) H(PN-)$_\ell$, (2) MH(PN-)$_\ell$.*

		H(PN-)$_\ell$	MH(PN-)$_\ell$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	0.8657(0.1437)	0.8720(0.1438)
Time	ξ_1	0.0675(0.0179)	0.0674(0.0178)
Sex	ξ_2	0.0829(0.0271)	0.0831(0.0272)
Age	ξ_3	0.0286(0.0092)	0.0287(0.0092)
Std. dev. random intercept count	d_1	0.1201(0.0320)	0.1202(0.0320)
Inflation intercept	γ_0	2.3408(0.3830)	2.0853(0.3451)
Inflation time	γ_1	0.1792(0.0460)	0.1573(0.0415)
Inflation sex	γ_2	−0.4285(0.0726)	−0.3864(0.0651)
Inflation age	γ_3	−0.0632(0.0249)	−0.0565(0.0224)
Std. dev. random intercept inflation	d_2	0.7987(0.0681)	0.4692(0.0397)
Corr. random effects	ρ	−0.1316(0.2807)	−0.1362(0.2893)
−2log-likelihood		13,241	13,242
AIC		13,263	13,264

Table 6.19: *Jimma Longitudinal Family Survey of Youth. Parameter estimates (standard errors) for the regression coefficients in (1) (PN-)$_\ell$, (2) M(PN-)$_\ell$.*

		(PN-)$_\ell$	M(PN-)$_\ell$
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	−0.5359(0.3149)	0.8642(0.3300)
Time	ξ_1	0.0559(0.0253)	0.0559(0.0253)
Sex	ξ_2	0.5365(0.0869)	0.5365(0.0869)
Age	ξ_3	−0.0465(0.0206)	−0.0465(0.0206)
Std. dev. random intercept count	d_1	1.6734(0.0485)	1.6734(0.0485)
−2log-likelihood		20,109	20,109
AIC		20,119	20,119

Table 6.20: *Jimma Longitudinal Family Survey of Youth. Observed and Predicted Probability of Zeros in (PN-); M(PN-); ZI(PN-)$_\ell$, MZI(PN-)$_\ell$ with logit link for zero-inflation; H(PN-)$_\ell$, MH(PN-)$_\ell$ with logit link for zero-inflation.*

	year one		year two		year three	
Model	observed	predicted	observed	predicted	observed	predicted
M(PN-)	72.4	54.4197	77.9	55.1338	76.2	54.0556
M(PNG)	72.4	54.4197	77.9	55.1338	76.2	54.0556
ZI(PN-)$_\ell$	72.4	75.2293	77.9	76.1590	76.2	79.1744
MZI(PN-)$_\ell$	72.4	75.2621	77.9	76.1361	76.2	79.1000
H(PN-)$_\ell$	72.4	75.2650	77.9	76.1975	76.2	79.1832
MH(PN-)$_\ell$	72.4	75.2984	77.9	76.1734	76.2	79.1069

Chapter 7

A Joint Model for Hierarchical Continuous and Zero Inflated Overdispersed Count Data

Many applications in public health, medical and biomedical or other studies demand modelling of two or more longitudinal outcomes jointly to get better insight into their joint evolution. In this regard, a joint model for a longitudinal continuous and a count sequence, the latter possibly overdispersed and zero-inflated (ZI), will be specified that assembles aspects coming from each one of them into one single model. For the continuous sequence, the linear mixed models (LMM) provide a general and flexible modeling framework where a subject-specific random effect, assumed to follow a normal distribution, is included to account for the correlation (Laird and Ware, 1982; Verbeke and Molenberghs, 2000). On the other hand, for the count outcome, clustering and overdispersion are accommodated through two distinct sets of random effects in a generalized linear model as proposed by Molenberghs *et al.* (2010); one is normally distributed, the other conjugate to the outcome distribution. An excessive number of zero counts is often accounted for by using a so-called ZI or hurdle model. ZI models combine either a Poisson or negative-binomial model with an atom at zero as a mixture, while the hurdle model separately handles the zero observations and the positive counts. A unified ZI(PNG) model to simultaneously allow for the correlation, overdispersion and zero-inflation in the count sequence was studied in Chapter 5. The marginalized versions of ZI(PNG) and a hurdle counterpart H(PNG), denoted as MZI(PNG) and MH(PNG) were the focus of Chapter 6. Recently Kassahun *et al.* (2013), extended Molenberghs *et al.* (2010) and studied a joint modeling strategy to simultaneously deal with: (1) correlation coming from the continuous sequence; (2) correlation, overdispersion and zero-inflation of the count sequence, where for the zero-inflation either the ZI or the hurdle models are considered as alternative approaches.

In this chapter, we propose a general joint modelling framework in which correlation from the count sequence as well as correlation, overdispersion and zero-inflation from the count sequence can appear together. The association among the two sequences is captured by correlating the normal random effects describing the continuous and count outcome sequences, respectively. This chapter is organized as follows. In Section 7.1, two alternative joint modeling strategies are described, followed by their marginalized versions in Section 7.2, with the estimation technique given in Section 7.3. The models are applied and compared based on the Jimma Infant data as introduced in Section 2.1, with the results presented in Section 7.4. In addition, these models are further studied in a simulation study which is presented in Section 7.5, Finally, some concluding remarks are given in Section 7.6. The contribution of this chapter has been published in Kassahun *et al.* (2013).

7.1 A Joint Combined Model for Continuous and Zero-inflated Count Data

Let Y_{ij} denote a longitudinal continuous outcome, and Z_{ik} an overdispersed count outcome with excessive number of zeros, with densities $f_{1i}(y_{ij})$ and $f_{2i}(z_{ik})$, respectively ($i = 1, \ldots, N$, $j = 1, \ldots, n_{1i}$, and $k = 1, \ldots, n_{2i}$). Formulation of a joint model could be based on the random-effects approach. Y_{ij} and Z_{ik} are modeled separately by including subject-specific random-effects $\boldsymbol{b_{1i}}$ and $\boldsymbol{b_{2i}}$, respectively. Conditionally upon the random-effects, the two outcomes are assumed independent. Hence, the association between Y_{ij} and Z_{ik} is captured by letting $\boldsymbol{b_{1i}}$ and $\boldsymbol{b_{2i}}$ to correlate (Molenberghs and Verbeke, 2005). A special case is the so-called shared parameter model, where the same set of random-effects is assumed for all outcomes. However, this approach has the disadvantage that it is based on strong assumptions about the association of the two outcomes, and hence may not be valid (Molenberghs and Verbeke, 2005).

Combining model elements from the linear mixed model of Sections 3.3, and the zero-inflated combined model of Section 3.7, in one single model, the so resulting zero-inflated joint combined model, conditional upon the random effects, has the following distribution:

$$f_i(y_{ij}, z_{ik}|\boldsymbol{b_{1i}}, \boldsymbol{b_{2i}}, \boldsymbol{\beta}, \boldsymbol{\xi}, \theta_{ik}, \phi, \pi_{ik}) = f_{1i}(y_{ij}|\boldsymbol{b_{1i}}, \boldsymbol{\beta}) \times f_{2i}(z_{ik}|\boldsymbol{b_{2i}}, \boldsymbol{\xi}, \theta_{ik}, \phi, \pi_{ik}), \quad (7.1)$$

where $\boldsymbol{b_{1i}}$ and $\boldsymbol{b_{2i}}$ are assumed to follow a multivariate normal distribution and correlated,

$$\boldsymbol{b_i} = (\boldsymbol{b_{1i}}, \boldsymbol{b_{2i}})' \sim MVN\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{D_{11}} & \boldsymbol{D_{12}} \\ \boldsymbol{D_{12}}' & \boldsymbol{D_{22}} \end{bmatrix}\right)$$

$\boldsymbol{D_{11}}$, $\boldsymbol{D_{12}}$, and $\boldsymbol{D_{22}}$ are unknown positive-definite matrices. For the continuous outcome, $f_{1i}(y_{ij}|\boldsymbol{b_{1i}}, \boldsymbol{\beta})$, is the linear mixed model, as discussed in Section 3.3, and for the count sequence, $f_{2i}(z_{ik}|\boldsymbol{b_{2i}}, \boldsymbol{\xi}, \theta_{ik}, \phi, \pi_{ik})$ is the zero-inflated combined model given as:

$$Z_{ik} \quad \sim \quad \begin{cases} 0 & \text{with probability } \pi_{ik}, \\ f_i(z_{ik}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{with probability } 1 - \pi_{ik}, \end{cases} \quad (7.2)$$

leading to the probabilities $p(Z_{ik} = z_{ik}|\boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik})$ given by

$$p(Z_{ik} = y_{ik}|\boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik}) = \begin{cases} \pi_{ik} + (1 - \pi_{ik})f_i(0|\boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}) & \text{if } z_{ik} = 0, \\ (1 - \pi_{ik})f_i(z_{ik}|\boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}) & \text{if } z_{ik} > 0. \end{cases} \quad (7.3)$$

as defined in Section 5.1. Similarly, for the hurdle joint combined model, we combine the linear mixed model and the hurdle combined model, where $f_{2i}(z_{ik}|\boldsymbol{b_{2i}}, \boldsymbol{\xi}, \theta_{ik}, \phi, \pi_{ik})$ is:

$$Z_{ik} \quad \sim \quad \begin{cases} 0 & \text{with probability } \pi_{ik}, \\ {f_i}^*(z_{ik}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{with probability } 1 - \pi_{ik}, \end{cases} \quad (7.4)$$

where ${f_i}^*$ is a truncated-at-zero Poisson-gamma-normal model, leading to probabilities given by:

$$p(Z_{ik} = z_{ik}|\boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}, \phi, \pi_{ik}) = \begin{cases} \pi_{ik} & \text{if } z_{ik} = 0, \\ (1 - \pi_{ik})\frac{f_i(z_{ik}|\lambda_{ik}, \theta_{ik})}{1 - f_i(0|\lambda_{ik}, \theta_{ik})} & \text{if } z_{ik} > 0, \end{cases}$$

as shown in Section 6.1. The resulting model becomes ZI(NN-)(PNG) or H(NN-)(PNG).

7.2 Marginalized Joint H(NN)(PNG) and ZI(NN-)(PNG) Models

In this section a marginalized version of the joint model for hierarchical continuous and overdisperesed and zero-inflated count data, shown in Section 7.1 will be considered.

First, for a longitudinal Gaussian outcome, the linear mixed models, as shown in Section 3.3 is very popular. The implied marginal model is given by:

$$\boldsymbol{Y_i} \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \Sigma_i)$$

(Laird and Ware, 1982; Verbeke and Molenberghs, 2000). Note that

$$\mathrm{E}(y_{ij}) = \mathrm{E}(y_{ij}/\boldsymbol{b}_{1i}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}$$

and hence conditional and marginal parameters of $\boldsymbol{\beta}$ in (7.1) are equal (Verbeke and Molenberghs, 2000).

Second, for the count sequence (PNG) to deal with correlation and overdispersion, with extensions H(PNG) and ZI(PNG) to further account for excessive zeros were shown in Section 7.1. A marginalized formulation of (PNG), denoted as M(PNG) in Section 3.6. Similarly, details of marginal expressions and estimation techniques for H(PNG) and ZI(PNG), with notations MH(PNG) and MZI(PNG) are shown in Sections 6.1– 6.3.

Fully marginalized joint models, i.e., MH(NN-)(PNG) and MZI(NN-)(PNG) easily follow in a straight forward fashion by simply replacing H(PNG) and ZI(PNG) models in f_{2i} of (7.1) with MH(PNG) of Section 6.1 and MZI(PNG) of Section 6.2.

7.3 Estimation

Let us consider the count component. We will make use of the partial marginalization for parameter estimation, as presented in Molenberghs *et al.* (2010). By this we refer to integrating the likelihood over the gamma random effects only, leaving the normal random effects untouched. The corresponding conditional probability for the combined model of Section 4.1 model is:

$$\begin{aligned} f(y_{ij}|\boldsymbol{b_i},\boldsymbol{\xi},\theta_{ij},\phi) &= \int f(y_{ij}|\boldsymbol{b_i},\boldsymbol{\xi},\theta_{ij},\phi)f(\theta_{ij}|\alpha_j,\beta_j)d\theta_{ij} \\ &= \begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}}, \end{aligned}$$

where κ_{ij} is as in (3.17). For the zero-inflated Poisson-gamma-normal combined case:

$$\begin{aligned} & f(y_{ij}|\boldsymbol{b_i},\boldsymbol{\xi},\theta_{ij},\phi,\pi_{ij}) \\ &= I(y_{ij}=0)\pi_{ij} \\ &\quad +(1-\pi_{ij}) \begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}}, \end{aligned}$$

with $\pi_{ij} = \pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma})$. Note that, with this approach, we assume that the gamma random effects are independent within a subject. This is fine, given the correlation is induced by the normal random effects.

Applying the above result to the joint combined model in (7.1), the zero-inflated joint combined model, conditional upon the random effects, is:

$$\begin{aligned}
&f_i(y_{ij}, z_{ik}|\boldsymbol{b_{1i}}, \boldsymbol{b_{2i}}, \boldsymbol{\beta}, \boldsymbol{\xi}, \theta_{ik}, \phi, \pi_{ik}) \\
&= \quad \frac{1}{(2\pi)^{\frac{n_i}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_{1i})'\Sigma_i{}^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_{1i})} \\
&\qquad \times \prod_k f(z_{ik}|\boldsymbol{b_{2i}}, \boldsymbol{\xi}, \theta_{ik}, \phi, \pi_{ik}).
\end{aligned}$$

Similarly, the hurdle Poisson-gamma-normal combined model, say $f^*(y_{ij})$, using a truncated-at-zero Poisson-gamma-normal model, as discussed in Section 6.1 is:

$$\begin{aligned}
&f^*(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\xi}, \theta_{ij}, \phi, \pi_{ij}) \\
&= \quad I(y_{ij} = 0)\pi_{ij} \\
&\quad +(1-\pi_{ij})\begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}} \\
&\quad \times \frac{1}{1-\left(\frac{1}{1+\kappa_{ij}\beta_j}\right)^{\alpha_j}}.
\end{aligned}$$

Consequently, the partially marginalized form for the H(NN-)(PNG) is:

$$\begin{aligned}
&f_i(y_{ij}, y_{ik}|\boldsymbol{b_{1i}}, \boldsymbol{b_{2i}}, \boldsymbol{\beta}, \boldsymbol{\xi}, \theta_{ik}, \phi, \pi_{ik}) \\
&= \quad \frac{1}{(2\pi)^{\frac{n_i}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_{1i})'\Sigma_i{}^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_{1i})} \\
&\qquad \times \prod_k f^*(y_{ik}|\boldsymbol{b_{2i}}, \boldsymbol{\xi}, \theta_{ik}, \phi, \pi_{ik}).
\end{aligned}$$

Estimation of the fully marginalized joint models is straightforward by replacing H(PNG) by MH(PNG), leading to MH(NN-)(PNG) model and ZI(PNG) by MZI(PNG), leading to MZI(NN-)(PNG) model.

For all of these, it is straightforward to obtain the fully marginalized probability by numerically integrating over the normal random effects, and using a tool such as the SAS procedure NLMIXED that allows for normal random effects in arbitrary, user-specified models. While the SAS procedure NLMIXED is equipped with default starting values, it is advisable to provide user-defined starting values instead. These can be obtained, for example, from models without random effects, with some trial and error. It is equally wise to ensure that both the outcome values as well as the covariates have magnitudes that are neither very large nor extremely small, because this may jeopardize stability of the iterative process. Also, it is useful, for example, to first first fit the individual models and use the output as starting values for the joint model. Against this background, our data analysis proceeded without difficulty. Example NLMIXED code is provided in Appendix D for both the data analysis and the simulation study.

7.4 Analysis of the Jimma Infant Growth Study

We analyze the Jimma Infant data as introduced in Section 2.1, where body weight as well as number of days of diarrheal illnesses were measured repeatedly for each infant. The two outcomes will be modeled jointly to capture association between them. Denote by Y_{ij} and Y_{ik} weight and number of days of illness measurements for the i^{th} infant at the j^{th} and k^{th} visit. We formulate a ZI(NN-)(PNG) or a H(NN-)(PNG) model for these data. The means are μ_{ij} and κ_{ik}, respectively. We model these as $\mu_{ij} = \beta_0 + b_{1i} + \beta_1 A_{ij} + \beta_2 {A_{ij}}^2$, and κ_{ik} as $\ln(\kappa_{ik}) = \xi_0 + b_{2i} + \xi_1 A_{ik}$. To account for excess zeros, the zero-inflation probability π_{ik} is written as $\text{logit}(\pi_{ik}) = \gamma_0 + \gamma_1 A_{ik}$. Here, A_{ij} is the age of the i^{th} infant at the j^{th} visit. Further, b_{1i} and b_{2i} represent subject-specific intercepts, assumed normally distributed and possibly correlated with mean and variance-covariance matrix given by

$$(b_{1i}, b_{2i})' \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} d_1 & \rho\sqrt{d}_1\sqrt{d}_2) \\ \rho\sqrt{d}_1\sqrt{d}_2 & d_2 \end{pmatrix}\right].$$

We examine the zero-inflation as well as the overdispersion aspect. The first issue can be addressed using a special type of the hurdle model, also known as the zero-altered model (Min and Agresti, 2005). These authors consider the testing problem for zero-inflation. This model requires the same covariates as well as the same distributional forms in the two parts. Explicitly, we assume,

$$\begin{aligned} \ln[-\ln(1-\pi_{ik})] &= a_1 + a_2(\xi_0 + \xi_1 A_{ik}) + b_{2i}, \\ \ln(\kappa_{ik}) &= \xi_0 + \xi_1 A_{ik} + b_{2i}. \end{aligned}$$

By setting, $a_2 = 1$, and testing whether $a_1 = 0$, one can test for zero-inflation. If $a_1 < 0$, then the data are zero-inflated; If $a_1 > 0$, the data are zero-deflated. Fitting the zero-altered combined model showed strong evidence of zero-inflation ($a_1 = -2.0689$, likelihood ratio test statistic = 3343, on one degree of freedom).

We also fitted the H(NN-)(PN-) and compared it with the H(NN-)(PNG). The difference in deviance is $94,181 - 86,379$, which evidently is extremely significant. This strongly underscores the presence of the overdispersion parameter α.

Parameter estimates for the (NN-), (PNG), and (NN-)(PNG) models are presented in Table 7.1. Technically, the separate models were fitted for the two outcomes together anyway, but assuming that $\rho = 0$, which is entirely equivalent to fitting the models separately. Clearly, body weight and number of days of diarrheal illness show strong inverse relationship as evidenced by the correlation of the random effects in the (NN-)(PNG). In addition, likelihood comparison shows a convincing improvement in model fit, when random effects are allowed to correlate. Comparing the separate and joint models, while parameter estimates for the continuous outcome remain the same, small changes are observed in the count part. All parameters are statistically significant in all models considered.

Table 7.2 gives the hurdle counterparts to the models in Table 7.1, i.e., (NN-), H(PNG), and H(NN-)(PNG). Evidently, the linear mixed model (NN-) is left unaltered because the hurdle aspect applies to the count process only. Similarly, Table 7.3 shows the ZI counterparts, (NN-), ZI(PNG), and ZI(NN-)(PNG). We deduce that the fit is improved quite a bit further, implying that the excess zeros need to be accommodated in the model. The aforementioned inverse relationship remains. While accounting for the excess zeros does not bring changes in the parameters for the continuous process, this is not the case for the counts, where the estimates change, with their corresponding standard errors getting relatively smaller. When the excess zeros are accounted for, the negative binomial parameter α gets much smaller in Tables 7.2 and 7.3, as compared to the corresponding value in Table 7.1. This underscores, once again, the connection between the zero-inflation and overdispersion phenomena. Indeed, when ZI is unaccounted for, the overdispersion aspect of the models captures a certain amount of this effect.

Turning to Tables 7.1 − 7.3 with an eye on the overdispersion parameter α, we see that it drops drastically when comparing Table 7.1 with Tables 7.2 and 7.3. Because in our parameterization larger values for α imply more pronounced overdispersion effects, it is clear that accounting for excess zeros explains a good amount of apparent overdispersion. The amount explained is relatively invariant to whether the two processes are estimated jointly or rather separately. The reduction also holds regardless of whether either a ZI or a hurdle correction is made. Finally, even after correction for excess zeros, there convincingly remains an amount of overdispersion. This underscores that both overdispersion as well as excess zeros need to be accounted for.

Table 7.1: *Jimma Infant Growth Study. (NN-), (PNG), and (NN-)(PNG) models.*

Effect	Parameter	(NN-) & (PNG)	(NN-)(PNG)
		Estimate (s.e.)	Estimate (s.e.)
Continuous process (Weight)			
Intercept	β_0	3.2767(0.0112)	3.2768(0.0112)
Age	β_1	0.7680(0.0026)	0.7680(0.0026)
Age×Age	β_2	$-0.0335(0.0002)$	$-0.0335(0.0002)$
Std. dev error	σ	0.6298(0.0022)	0.6298(0.0022)
Std. dev random effect	$\sqrt{d_1}$	0.8298(0.0072)	0.8299(0.0072)
Count process (Days of illness)			
Intercept	ξ_0	$-1.1567(0.0557)$	$-1.2094(0.0551)$
Age	ξ_1	0.2246(0.0072)	0.2279(0.0072)
Std. dev random effect	$\sqrt{d_2}$	0.3654(0.0507)	0.4611(0.0435)
Negative-binomial parameter	α	17.9605(0.2914)	17.6345(0.2858)
Common parameter			
Corr. random effect	ρ	—	$-0.6282(0.0565)$
-2log-likelihood		172,241	172,054

From Tables 7.2 and 7.3, we further observe that the H(NN-)(PNG) and the ZI(NN-)(PNG) are very similar, not only in terms of parameter estimates but also as far as resulting inferences go. In count data modeling, the choice among zero-inflated and hurdle models should be based not only on model fit but also on assumptions about the underlying data generation process (Neelon *et al.*, 2010). If zeros are expected to come from both the point mass and the count component, then zero-inflated models may be preferable. In addition, in the presence of strong evidence of zero-inflation, zero-inflation models may provide better fit. On the other hand, Min and Agresti (2005) discuss why the hurdle model is, in general, preferred to the zero-inflated model: it allows to test for zero-inflation, it works also in the zero-deflation setting, and its two parts are separate.

The fully marginalized models, i.e., (NN-) & M(PNG), M(NN-)(PNG), (NN-) & MH(PNG), (NN-) & MZI(PNG), and MZI(NN-)(PNG), are also fitted, leading to

Table 7.2: *Jimma Infant Growth Study. (NN-), H(PNG), and H(NN-)(PNG) models.*

Effect	Parameter	(NN-) & H(PNG)	H(NN-)(PNG)
		Estimate (s.e.)	Estimate (s.e.)
Continuous process (Weight)			
Intercept	β_0	3.2767(0.0112)	3.2767(0.0112)
Age	β_1	0.7680(0.0026)	0.7680(0.0026)
Age×Age	β_2	$-0.0335(0.0002)$	$-0.0335(0.0002)$
Std. dev error	σ	0.6298(0.0022)	0.6298(0.0022)
Std. dev random effect	$\sqrt{d_1}$	0.8298(0.0072)	0.8299(0.0072)
Count process (Days of illness)			
Intercept	ξ_0	2.0437(0.0251)	2.0225(0.0251)
Age	ξ_1	0.0185(0.0028)	0.0199(0.0028)
Std. dev random effect	$\sqrt{d_2}$	0.4374(0.0118)	0.4392(0.0118)
Negative-binomial parameter	α	0.3271(0.0104)	0.3259(0.0103)
Inflation intercept	γ_0	2.8383(0.0290)	2.8383(0.0290)
Inflation Age	γ_1	$-0.1687(0.0035)$	$-0.1687(0.0035)$
Common parameter			
Corr. random effect	ρ	—	$-0.2255(0.0246)$
-2log-likelihood		165,475	165,395

population-average interpretation for all fixed effects.

Parameter estimates for (NN-) & M(PNG), M(NN-)(PNG) are shown in Table 7.4, the results being very similar to that of (NN-) & (PNG), (NN-)(PNG) in Table 7.1, except a slight difference in ξ_0. Similarly, marginalized hurdle versions (NN-) & MH(PNG) and (NN-) & MH(PNG), as shown in Table 7.5, again, suggested that these data are subject to overdispersion, correlation and zero-inflation features, with parameter estimates as well as model fit statistics very similar to (NN-) & H(PNG) and (NN-) & H(PNG) as reported in Table 7.2, except some difference in ξ_0. This same phenomenon also holds true for (NN-) & MZI(PNG) and (NN-) & MZI(PNG) given in Table 7.6 as compared to (NN-) & H(PNG) and (NN-) & H(PNG) of Table 7.2. Across the models, we observe that the intercept of the count sequence ξ_0 shows

Table 7.3: *Jimma Infant Growth Study. (NN-), ZI(PNG), and ZI(NN-)(PNG) models.*

Effect	Parameter	(NN-) & ZI(PNG)	ZI(NN-)(PNG)
		Estimate (s.e.)	Estimate (s.e.)
Continuous process (Weight)			
Intercept	β_0	3.2767(0.0112)	3.2767(0.0112)
Age	β_1	0.7680(0.0026)	0.7680(0.0026)
Age×Age	β_2	-0.0335(0.0002)	-0.0335(0.0002)
Std. dev. error	σ	0.6298(0.0022)	0.6298(0.0022)
Std. dev. random effect	$\sqrt{d_1}$	0.8298(0.0072)	0.8299(0.0072)
Count process (Days of illness)			
Intercept	ξ_0	2.0270(0.0259)	2.0020(0.0260)
Age	ξ_1	0.0190(0.0029)	0.0205(0.0029)
Std. dev. random effect	$\sqrt{d_2}$	0.4464(0.0123)	0.4500(0.0123)
Negative-binomial parameter	α	0.3259(0.0104)	0.3261(0.0104)
Inflation intercept	γ_0	2.8071(0.0292)	2.8051(0.0292)
Inflation Age	γ_1	-0.1686(0.0035)	-0.1685(0.0035)
Common parameter			
Corr. random effect	ρ	—	-0.2409(0.0245)
-2log-likelihood		165,461	165,369

little difference as a result of the marginalization. On the other hand all parameter estimates and standard errors in the continuous sequence are strikingly very similar, implying that whether a marginal or hierarchical formulation employed, parameters retain both their interpretation as well as their magnitude. Further, marginally, the correlation parameter of the random effects is, once again negative and significant, suggesting that body weight and number of days of illness are inversely associated, with the estimates similar to the corresponding conditional models.

Table 7.4: *Jimma Infant Growth Study. (NN-), M(PNG), and M(NN-)(PNG) models.*

Effect	Parameter	(NN-) & M(PNG)	M (NN-)(PNG)
		Estimate (s.e.)	Estimate (s.e.)
		Continuous process (Weight)	
Intercept	β_0	3.2767(0.0112)	3.2768(0.0112)
Age	β_1	0.7680(0.0026)	0.7680(0.0026)
Age×Age	β_2	−0.0335(0.0002)	−0.0335(0.0002)
Std. dev error	σ	0.6298(0.0022)	0.6298(0.0022)
Std. dev random effect	$\sqrt{d_1}$	0.8298(0.0072)	0.8299(0.0072)
		Count process (Days of illness)	
Intercept	ξ_0	−1.0899(0.0459)	−1.1031(0.0454)
Age	ξ_1	0.2246(0.0072)	0.2279(0.0072)
Std. dev random effect	$\sqrt{d_2}$	0.3654(0.0507)	0.4611(0.0435)
Negative-binomial parameter	α	17.9606(0.2914)	17.6342(0.2857)
		Common parameter	
Corr. random effect	ρ	—	−0.6282(0.0565)
−2log-likelihood		172,241	172,054

7.5 Simulation Study

A simulation study is conducted to assess the impact of not appropriately accounting for the excess zero counts as well as misspecification of the overdispersion in joint modeling of hierarchical continuous and count outcome. We choose to conduct this study following three different settings.

7.5.1 Simulation Settings

Data are generated in the spirit of the design and outcomes of the data in Section 2.1, which consist of body weight and counts of the number of days of diarrheal disease illnesses among infants measured repeatedly over time. Age in months is considered as the time variable.

A random sample of 250 data sets are generated under three scenarios. S_1:

Table 7.5: *Jimma Infant Growth Study. (NN-), MH(PNG), and MH(NN-)(PNG) models.*

Effect	Parameter	(NN-) & MH(PNG)	MH(NN-)(PNG)
		Estimate (s.e.)	Estimate (s.e.)
		Continuous process (Weight)	
Intercept	β_0	3.2767(0.0112)	3.2767(0.0112)
Age	β_1	0.7680(0.0026)	0.7680(0.0026)
Age×Age	β_2	$-0.0335(0.0002)$	$-0.0335(0.0002)$
Std. dev error	σ	0.6298(0.0022)	0.6298(0.0022)
Std. dev random effect	$\sqrt{d_1}$	0.8298(0.0072)	0.8299(0.0072)
		Count process (Days of illness)	
Intercept	ξ_0	2.1421(0.0250)	2.1189(0.0249)
Age	ξ_1	0.0183(0.0028)	0.0199(0.0028)
Std. dev random effect	$\sqrt{d_2}$	0.4378(0.0118)	0.4392(0.0118)
Negative-binomial parameter	α	0.3270(0.0104)	0.3259(0.0103)
Inflation intercept	γ_0	2.8378(0.0290)	2.8383(0.0290)
Inflation Age	γ_1	$-0.1687(0.0035)$	$-0.1687(0.0035)$
		Common parameter	
Corr. random effect	ρ	—	$-0.2256(0.0246)$
-2log-likelihood		165,475	165,395

from a ZI(NN-)(PNG); S_2 from a ZI(NN-)(PN-); S_3 from a (NN-)(PNG). Model fitting is based on these three models, supplemented with others: ZI(NN-)(PNG), H(NN-)(PNG), or (NN-)(PNG); also, the versions without overdispersion are considered: ZI(NN-)(PN-), H(NN-)(PN-), or (NN-)(PN-).

We consider 200 subjects with 10 measurements per subject. The continuous response Y_{ij} is modeled as $Y_{ij} = \beta_0 + \beta_1 A_{ij} + \beta_2 {A_{ij}}^2 + b_{1i} + \varepsilon_{ij}$. The subject-specific random intercept b_{1i} and the residual error ε_i are assumed independent, and generated from normal distribution with mean 0 and standard deviations 2 and 0.6, respectively. The count outcome Y_{ik} is modeled using predictor function $\kappa_{ik} = \exp\{\xi_0 + b_{2i} + \xi_1 A_{ik}\}$. When there is overdispersion, the outcome is generated directly from a negative-binomial process with $Y_{ik} \sim \text{NB}(\psi_{ik}, \theta)$, where $\theta = 1$ and

Table 7.6: *Jimma Infant Growth Study. (NN-), MZI(PNG), and MZI(NN-)(PNG) models.*

Effect	Parameter	(NN-) & MZI(PNG)	MZI(NN-)(PNG)
		Estimate (s.e.)	Estimate (s.e.)
		Continuous process (Weight)	
Intercept	β_0	3.2767(0.0112)	3.2767(0.0112)
Age	β_1	0.7680(0.0026)	0.7680(0.0026)
Age×Age	β_2	$-0.0335(0.0002)$	$-0.0335(0.0002)$
Std. dev. error	σ	0.6298(0.0022)	0.6298(0.0022)
Std. dev. random effect	$\sqrt{d_1}$	0.8298(0.0072)	0.8299(0.0072)
		Count process (Days of illness)	
Intercept	ξ_0	2.1266(0.0256)	2.1027(0.0256)
Age	ξ_1	0.0190(0.0029)	0.0206(0.0029)
Std. dev. random effect	$\sqrt{d_2}$	0.4464(0.0123)	0.4500(0.0123)
Negative-binomial parameter	α	0.3259(0.0104)	0.3261(0.0104)
Inflation intercept	γ_0	2.8071(0.0292)	2.8053(0.0292)
Inflation Age	γ_1	$-0.1686(0.0035)$	$-0.1685(0.0035)$
		Common parameter	
Corr. random effect	ρ	—	$-0.2405(0.0245)$
-2log-likelihood		165,461	165,369

$\psi_{ik} = (1 + \kappa_{ik}/\theta)^{-1}$. As before, A_{ij} represents the age at which the j^{th} measurement is taken for the i^{th} subject. Practically, age is generated from the empirical distribution observed in the Jimma Infact Growth Study. The random intercept b_{2i} follows a mean-zero normal with variance 1.5. When zero-inflation is present, this is added by defining the final response vector $\boldsymbol{Y}_i^*$ with components ${Y_{ik}}^* = (1 - u_{ik})Y_{ik}$, where the u_{ik} are Bernoulli random variables with parameters π_{ik} and $\text{logit}(\pi_{ik}) = \gamma_0 + \gamma_1 A_{ik}$. To correlate the two processes, the random intercepts b_{1i} and b_{2i} are allowed to correlate with one another, with $\rho = -0.5$. When generating data, the true parameter values were $\boldsymbol{\beta} = (3.3, 0.77, -0.03)^T$, $\boldsymbol{\xi} = (2, 0.02)^T$, and $\boldsymbol{\gamma} = (2, -0.2)^T$.

7.5.2 Simulation Results

The results under S_1 are summarized in Tables 7.7 and 7.8. Clearly, the ZI(NN-)(PNG) and the H(NN-)(PNG) result in estimates very close to the true values. However, as can be seen from the (NN-)(PNG), omitting zero-inflation highly affects the estimates in the count component, with a non-negligible amount of bias loaded on the correlation parameter. Further, when both zero-inflation and overdispersion are mis-specified, by fiting the (NN-)(PN-), as shown in Table 7.8, a similar phenomenon is evident, where now the random-effects variance tries to recover from mis-specifying the overdispersion.

Under scenario S_2, the results of which are presented in Table 7.9, the impact of omitting the extra zeros is still evident, though the overdispersion parameter α in the (NN-)(PNG) seems to help recover from misspecification. Further, we also note that correlation is overestimated as a result of the misspecification. These results, once more, underscore the necessity of models appropriately accounting for the excessive zeros.

Scenario 1 leads to about 75% of zeros, with a similar fraction (72%) in Scenario 2. Scenario 3 is qualitatively different, with roughly 18% of zeros. Comparing mean and standard deviation shows that all three are overdispersed. Under Scenarios 1 and 2, this stems to a large part from extra zeros, whereas in Scenario 3 this is “pure” overdispersion. When data are overdispersed, but not subject to considerable zero-inflation as in S3, fitting models allowing for extra zeros is less important. As shown in Table 7.10, the (NN-)(PNG), which is the true model, performs well. In addition, we observe that the (NN-)(PN-) model is also doing well, but this probably may not be the case when data are subject to much higher levels of overdispersion than those considered here.

In addition, across our simulation study, we learned that the zero-inflated models are relatively harder to fit when compared to the hurdle models, where convergence of models is never an issue, with convergence guaranteed for the latter. Further, in scenarios S_1 and S_2, though data are generated from the ZI(NN-)(PNG), the H(NN-)(PNG) is also performing very well. In addition, across our simulation study, we learned that the zero-inflated models are relatively harder to fit when compared to the hurdle models, where convergence of models is never an issue, with convergence guaranteed for the latter. Further, in scenarios S_1 and S_2, though data are generated from the ZI(NN-)(PNG), the H(NN-)(PNG) is also performing very well.

Table 7.7: *Simulation study under scenario S_1. Mean and Relative bias (RB) of the parameter estimates in the ZI(NN-)(PNG), H(NN-)(PNG), and (NN-)(PNG).*

Effect	Parameter	True	ZI(NN-)(PNG)	H(NN-)(PNG)	(NN-)(PNG)
			Mean (RB)	Mean (RB)	Mean (RB)
		Continuous process			
Intercept	β_0	3.3	3.297(-0.001)	3.297(-0.001)	3.297(-0.001)
Age	β_1	0.77	0.772(0.002)	0.772(0.002)	0.772(0.002)
Age×Age	β_2	−0.03	−0.030(0.003)	−0.030(0.003)	−0.030(0.003)
Std. dev. error	σ	0.6	0.599(-0.001)	0.599(-0.001)	0.599(-0.001)
Std. dev. random effect	$\sqrt{d_1}$	2	1.988(-0.006)	1.987(-0.006)	1.988(-0.006)
		Count process			
Intercept	ξ_0	2	1.977(-0.012)	2.145(0.073)	−0.169(-1.085)
Age	ξ_1	0.02	0.023(0.145)	0.022(0.100)	0.218(9.885)
Std. dev. random effect	$\sqrt{d_2}$	1.5	1.478(-0.015)	1.341(-0.106)	1.311(-0.126)
Negative-binomial parameter	α	1	0.992(-0.008)	1.051(0.051)	11.895(10.895)
Inflation intercept	γ_0	2	1.996(-0.002)	2.209(0.105)	—
Inflation Age	γ_1	−0.2	−0.198(-0.009)	−0.188(-0.062)	—
		Common parameter			
Corr. random effect	ρ	−0.5	−0.503(0.006)	−0.501(0.001)	−0.563(0.126)
Frequency of convergence			250	250	250

7.6 Discussion

In this chapter, we have presented a joint modeling strategy for a hierarchical continuous and count outcome, where the latter is subject to zero-inflation as well as overdispersion. We show that zero-inflation, and overdispersion features in count data modeling could also be well extended to a joint modeling framework such that model fit is improved and inference refined. However, any failure to appropriately account for such features appropriately may result in a substantial impact on the parameter estimates and precision estimates. When zero-inflation is omitted in the model, the overdispersion term will try to recover for this mis-specification, though both are eventually needed.

Fitting a ZI(NN-)(PNG) model, even when correctly specified, is relatively more complex than a H(NN-)(PNG). The latter has several additional advantages, in particular the possibility to test for zero inflation. In the real data analysis, there were no model convergence issues, which is reassuring. Of course, as we learned from analyzing these data, both overdispersion as well as additional zero inflation were

Table 7.8: *Simulation study under scenario S_1. Mean and Relative bias (RB) of the parameter estimates in the ZI(NN-)(PN-), H(NN-)(PN-), and (NN-)(PN-).*

Effect	Parameter	True	ZI(NN-)(PN-)	H(NN-)(PN-)	(NN-)(PN-)
			Mean (RB)	Mean (RB)	Mean (RB)
		Continuous process			
Intercept	β_0	3.3	3.284(-0.005)	3.292(-0.002)	3.296(-0.0012)
Age	β_1	0.77	0.772(0.002)	0.772(0.002)	0.772(0.0021)
Age×Age	β_2	-0.03	$-0.030(0.007)$	$-0.030(0.003)$	$-0.030(0.003)$
Std. dev. error	σ	0.6	0.599(-0.001)	0.599(-0.001)	0.599(-0.001)
Std. dev. random effect	$\sqrt{d_1}$	2	1.988(-0.006)	1.988(-0.006)	1.988(-0.006)
		Count process			
Intercept	ξ_0	2	1.134(-0.433)	2.016(0.008)	$-0.4531(-1.2266)$
Age	ξ_1	0.02	0.025(0.270)	0.026(0.290)	0.162(7.085)
Std. dev. random effect	$\sqrt{d_2}$	1.5	1.566(0.044)	1.426(-0.049)	1.794(0.196)
Negative-binomial parameter	α	1	—	—	—
Inflation intercept	γ_0	2	1.972(-0.014)	2.209(0.1044)	—
Inflation Age	γ_1	-0.2	$-0.195(-0.027)$	$-0.188(-0.062)$	—
		Common parameter			
Corr. random effect	ρ	-0.5	$-0.446(0.108)$	$-0.436(0.128)$	$-0.406(-0.188)$
Frequency of convergence			233	250	250

convincingly present. Furthermore, the set of data was very large. These are comfortable conditions to reach convergence. In the simulation study, the hurdle model was at an advantage when it came to model fit. Practically, readers may consider both approaches, hurdle and ZI, and perhaps use the relevant parameters from the hurdle model as starting values for the ZI fit.

In terms of estimation, we have focused on maximum likelihood estimation. This can be done by integrating over the random effects, using a combination of analytical and numerical techniques. Precisely, the likelihood was integrated analytically over the conjugate (gamma) random effect, using techniques outlined in Molenberghs *et al.* (2010). The so-resulting likelihood, still conditional on the normal random effect, is integrated numerically over the said random effect using the SAS procedure NLMIXED.

In conclusion, we note that our approach corrects for overdispersion and/or allows for joint modeling. In our example, both phenomena were present, although overdispersion results in a larger deviance reduction than joint modeling. One lesson

Table 7.9: *Simulation study under scenario S_2. Mean and Relative bias (RB) of the parameter estimates in the ZI(NN-)(PN-), H(NN-)(PN-), and (NN-)(PNG).*

Effect	Parameter	True	ZI(NN-)(PN-)	H(NN-)(PN-)	(NN-)(PNG)
			Mean (RB)	Mean (RB)	Mean (RB)
			Continuous process		
Intercept	β_0	3.3	3.325(0.008)	3.294(-0.002)	3.296(-0.001)
Age	β_1	0.77	0.769(-0.001)	0.7711(0.001)	0.771(0.001)
Age×Age	β_2	−0.03	−0.030(0.000)	−0.030(0.003)	−0.030(0.003)
Std. dev. error	σ	0.6	0.599(-0.001)	0.599(-0.001)	0.599(-0.001)
Std. dev. random effect	$\sqrt{d_1}$	2	1.985(-0.008)	1.986(-0.007)	1.986(-0.007)
			Count process		
Intercept	ξ_0	2	1.230(-0.385)	2.051(0.025)	0.083(-0.958)
Age	ξ_1	0.02	0.020(0.020)	0.020(0.015)	0.195(8.740)
Std. dev. random effect	$\sqrt{d_2}$	1.5	1.502(0.001)	1.424(-0.050)	1.219(-0.187)
Negative-binomial parameter	α	0	—	—	9.9406
Inflation intercept	γ_0	2	1.884(-0.058)	2.089(0.0445)	—
Inflation Age	γ_1	−0.2	−0.204(-0.019)	−0.196(-0.023)	—
			Common parameter		
Corr. random effect	ρ	−0.5	−0.495(-0.009)	−0.489(0.022)	−0.613(-0.225)
Frequency of convergence			241	250	250

to be drawn from this is that the user should carefully assess whether one or the other correction, both of them, or perhaps none of the two is necessary. Note also that joint modeling may be of interest in its own right. For example, one may be interested in measuring the strength of the association between both processes (estimating one or more correlation parameters) or in assessing its significance. Also, it is possible to derive prediction equations for an outcome or set of outcomes in one sequence, based on the outcomes in the other sequence and/or earlier measurements of the same sequence.

Table 7.10: *Simulation study under scenario S_3. Mean and Relative bias (RB) of the parameter estimates in the H(NN-)(PNG), (NN-)(PNG), and (NN-)(PN-).*

Effect	Parameter	True	H(NN-)(PNG)	(NN-)(PNG)	(NN-)(PN-)
			Mean (RB)	Mean (RB)	Mean (RB)
			Continuous process		
Intercept	β_0	3.3	3.294(-0.002)	3.295(-0.002)	3.295(-0.002)
Age	β_1	0.77	0.772(0.003)	0.772(0.003)	0.772(0.003)
Age×Age	β_2	−0.03	−0.030(0.007)	−0.030(0.007)	−0.030(0.007)
Std. dev. error	σ	0.6	0.599(-0.001)	0.599(-0.001)	0.599(-0.001)
Std. dev. random effect	$\sqrt{d_1}$	2	1.987(-0.006)	1.988(-0.006)	1.988(-0.006)
			Count process		
Intercept	ξ_0	2	2.063(0.031)	1.995(-0.002)	1.954(-0.023)
Age	ξ_1	0.02	0.019(-0.055)	0.019(-0.030)	0.018(-0.125)
Std. dev. random effect	$\sqrt{d_2}$	1.5	1.414(-0.057)	1.485(-0.009)	1.5223(0.015)
Negative-binomial parameter	α	1	1.008(0.008)	1.0012(0.0012)	—
Inflation intercept	γ_0	0	−1.448	—	—
Inflation Age	γ_1	0	−0.016	—	—
			Common parameter		
Corr. random effect	ρ	−0.5	−0.496(-0.007)	−0.499(-0.003)	−0.486(-0.027)
Frequency of convergence			250	250	249

Chapter 8

General Conclusions

In a lot of applied research, binary and count outcome frequently appear, next to continuous data. Statistical modeling of such data lies within the framework of exponential family distributions (McCullagh and Nelder, 1989; Agresti, 2002; Molenberghs and Verbeke, 2005). The resulting generalized linear models (GLMs) contain three components: a random component that identifies a vector of observations of the outcome and its probability distribution; a systematic component, i.e., a specification for the mean vector in terms of a vector of fixed unknown parameters and known covariate values; and a link function which specifies the function of expectation that the model equates to the systematic component with known link functions, such as the logit and log functions for binary and count data, respectively.

Generally, exponential family distributions are well known for the restrictive assumption that the mean and variance are related. For example, in the case of count data, the Poisson distribution assumes that the mean and the variance are equal. Similarly, for binomial data, the variance and the mean are functions of a single parameter. However, in most practical settings, these mean-variance relationships do not hold for a sample data, say, the variance of a count outcome exceeds its mean, leading to the so-called overdispersion (McCullagh and Nelder, 1989; Agresti, 2002; Molenberghs and Verbeke, 2005). When data are hierarchically organized, such as in clustered or longitudinal settings, correlation will be induced in the data which comes from the repeated measures nature for a given subject (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). Very often, count data are also characterized by the presence of excessive zero observations beyond what can be allowed for by a standard count distribution, such as Poisson distribution (Mullahy, 1986; Lambert, 1992; Greene, 1994). Many applications in public health, medical and biomedical or other studies demand modeling of two or more longitudinal outcomes or time-to-event data jointly to get better insight about their joint evolution. In this regard, a joint modeling of a longitudinal outcome and a hierarchical, overdispersed and zero-inflated count outcome can be an objective of a statistical investigation.

One possible route to deal with overdispersion is to introduce an overdispersion parameter and only specify a relationship between the mean and the variance, and then apply quasi-likelihood, whereby the extra variability in the data will be captured by the dispersion parameter (Wedderburn, 1974). In clustered binary and binomial data, an elegant way to account for overdispersion is through inclusion of beta random-effects, leading to the so-called beta-binomial model, in which the Bernoulli model is combined with a beta distribution (Molenberghs and Verbeke, 2005; Skellam, 1948; Hinde and Demétrio, 1998a; Hinde and Demétrio, 1998b; Kleinman, 1973). Turning to count data, it is common to combine Poisson distribution with a gamma distributed random effect, so that the unconditional distribution of the outcome turns out to be a negative binomial distribution (Breslow, 1984; Hinde and Demétrio, 1998a; Hinde and Demétrio, 1998b). On the other hand, focusing on hierarchical data, the GLM is usually extended to generalized linear mixed models (GLMMs), with a subject-specific random effect, usually a Gaussian type, added in the linear predictor to capture a hierarchy-induced association or to account for overdispersion (Engel and Keen, 1992; Molenberghs and Verbeke, 2005; Pinheiro and Bates, 2000). Molenberghs *et al.* (2010) proposed a flexible and unified modeling framework, termed the *combined model*, to simultaneously capture overdispersion and correlation for a wide range of clustered data, including count, binary and time-to-event. These authors brought together two sets of random effects. The normally distributed subject specific-random effects capture the correlation, while a conjugate measurement-specific random effect on the natural parameter, is used to accommodate overdispersion.

Tsiatis and Davidian (2004) studied joint modeling of longitudinal outcome and time-to-event data. Horrocks and van den Heuvel (2009) consider a joint model, consisting of a linear mixed-effects submodel for the longitudinal outcome and a generalized linear submodel for the primary binary endpoint. Molenberghs and Verbeke (2005) discuss a number of techniques that jointly model continuous and discrete outcomes. A joint model for a longitudinal continuous and a count sequence, the latter possibly overdispersed and zero-inflated, requires to assemble aspects coming from each one of them in one single model. On the one hand, a subject specific random-effect is included to account for the correlation in the continuous component. For the count outcome, on the other hand, clustering and overdispersion are accommodated through two separate sets of random effects in a generalized linear model as defined by Molenberghs *et al.* (2010). The association among the two sequences could be captured through subject-specific random-effects which are allowed to correlate. An excessive number of zero counts is often accounted for using the so called zero-inflated or hurdle model.

In this thesis, we studied and proposed different statistical modeling strategies of hierarchical data allowing for overdispersion for binary data (Chapter 4), and both overdispersion and zero-inflation, for count data (Chapters 5 and 6). In addition, a joint modeling framework for longitudinal continuous outcome and a hierarchical, overdispersed and zero-inflated count data was proposed (Chapter 7). A brief overview of the resulting conclusions for the pertinent chapters is now presented.

In Chapter 4, we have shown modeling of overdispersed hierarchical binary data. We considered both the likelihood, similar to the combined model of Molenberghs *et al.* (2010), and proposed its implementation in the Bayesian paradigm. The Beta-binomial distribution, which is a compound distribution of the binomial and its conjugate beta, was employed to capture overdispersion, and Gaussian random effects were included in the linear predictor, to capture correlation due to the data hierarchy. In the Bayesian approach, the ability to specify prior distribution helped to incorporate more information in inference, especially for complex models, like the combined model, that attempt to capture overdispersion and clustering using two separate sets of random effects (Spiegelhalter *et al.*, 2002). Beta-binomial approximates the binomial distribution arbitrarily well when its two non-negative parameters, α and β, determining its shape, are sufficiently larger (Gelman *et al.*, 2004). If one or both of these parameters are less than 1, then the probability mass function will go to infinity near its boundaries, 0 and 1, and hence not concave. As a result, the mode does not exist, leading to computational problems in MCMC. For this reason, we used the restriction $\alpha > 1,\ \beta > 1$, such that the density is always concave and unimodal whereby it is always finite over the support $[0, 1]$. We considered two real world data sets and analyzed, first in the likelihood context, and then in the Bayesian, which could also be considered as sensitivity analysis (Kassahun *et al.*, 2012).

Two longitudinal binary data sets, collected in south western Ethiopia: the Jimma infant growth study, where the child's early growth is studied, and the Jimma longitudinal family survey of youth where the adolescent's school attendance is studied over time, were considered. One of the key indicators of infant growth is Body Mass Index (BMI). Many studies suggest that Breastfeeding status, and socio-economic condition of the parents, among others, are potential risk factors of BMI (Macro., 2008; WHO, 2009). School attendance among adolescents varies among gender groups in a way that girls are at higher risk of school absentism as compared to boys. Moreover, adolescents living in urban areas have have a better school attendance rate, unlike those in the rural setting similar to Freedman *et al.* (1999). The analysis showed that the combined model results in model improvement in fit, and hence the preferred one, based on likelihood comparison, and DIC criterion. This implies that instead of accounting for overdispersion, and correlation separately, both can be accommodated simultaneously, by allowing two separate sets of the beta, and the normal random effects at once. Further, the two estimation approaches result in fairly similar parameter estimates and inferences in both of our case studies. Our data analysis showed that early initiation of breastfeeding has a protective effect against the risk of overweight in late infancy, while proportion of overweight seems to be invariant among males and females overtime. Gender is significantly associated with school attendance, where girls have a lower rate of attendance as compared to boys.

In Chapter 5, we extended the combined modeling idea of Molenberghs *et al.* (2010) for hierarchical count data, who brought together normal random effects to induce association between repeated Poisson data, and a gamma distributed random factor in the log-linear predictor to account for the overdispersion, i.e., (PNG) to further deal with an excess of zero observations. A zero-inflation extension of such model, ZI(PNG), assumes that there are two processes as sources of zeros: zeros may come from the point-mass or from the poisson-normal-gamma process as a mixture. Two real world count data sets characterized by correlation, overdispersion as well as excessive zeros were considered: the Jimma Infant growth study, where the number of days of diarrheal illness were studied, and the Epilepsy study, where the number of epileptic seizures that patients experience were the focus of investigation. Further, a simulation study was conducted based on the Jimma Infants data in three scenarios: without excessive zeros, moderate excessive zeros and higher proportion of excessive zeros. Both the real data sets and the simulated data were analyzed with (PNG) and ZI(PNG) and their special cases. We found that, when correlation, overdispersion as well as excessive zeros are appearing at once, the ZI(PNG) is the most preferred one. Any failure to account for excess zeros, overdispersion, and/or correlation has a substantial impact on bias and predicted probabilities. This was clearly shown on such key model parameters as the intercept term, the overdispersion parameter, and the variance of the random effects. In the simulated study, all scenarios suggest that the ZI(PNG) is the preferred one in terms of relative bias and predicted probabilities of zeros.

Chapter 6 was devoted to marginalized modeling strategies for hierarchical count data, characterized by correlation, overdispersion, and excess zeros: a marginalized hurdle Poisson-normal-gamma model and a marginalized zero-inflated Poisson-normal model. In these framework, the marginalization concept that Heagerty (1999) applied to multilevel models, adjustment for an excess of zero counts based on the hurdle model or the zero-inflated model (Mullahy, 1986; Lambert, 1992), and the combined modeling framework for overdispersion and correlation (Molenberghs *et al.*, 2010) are merged together in one single model. Two Gaussian distributed random-effects vectors, possibly correlated, were included such that one of them captures the correlation in the positive counts profile, while the other does the same in the excess zero model component. The correlation between these random effects can be interpreted as the correlation between the data generation processes of the zero state and the positive counts.

In terms of link function choices for the zero-inflation, we considered logit link for the marginal model, and a probit one for the conditional model, thus retaining the odds-ratio interpretation of the covariate effects, while taking computational advantage of the probit link. Marginal interpretation is possible not only for the count part, but also for the ZI component. Two real data sets on An. mosquito, collected through two techniques: IRC and CDC near to a hydroelectric dam (at risk) and away from the dam (control) were studied. We found that MH(PNG) model and MZI(PNG) model worked well. We considered AIC and deviance statistics for model comparison. As a result, the hurdle models showed better fit to the data, and lead to parameter estimates relatively superior in precision. Furthermore, when the percentage of zeros is subject to change with a given covariate, MH(PNG) tends to give a better prediction. In both case studies, the covariates village and season are affecting the positive counts part and the ZI part oppositely, implying the two processes are operating inversely. This was observed further from the negative sign of the correlation parameter of the two random effects. On the other hand, in the presence of modest overdispersion, as observed in the Jimma Longitudinal Family Survey of Youth, it appears that overdispersion and the excess zero aspects may not be well separated, and hence the Poisson-normal GLMM with only zero-inflation adjustment was sufficient. In a univariate setting, one might employ a likelihood based approach as proposed by Vuong (1989) for comparison of non-nested models, such as H(P--) and ZI(P--). Min and Agresti (2005) suggested a number of advantages of the hurdle model, including its flexibility to work well both in zero-inflation and zero-deflation situations and the possibility to test for evidence of zero-inflation. On the other hand, Todem *et al.* (2012) suggest an extended ZI model that can handle both zero-inflation and zero-deflation, though the proposed link function does not have the computational flexibility of the probit link. Furthermore, this model can be used to assess assumption of excessive zeros in a given data. One might expect a marked difference among the hurdle and the ZI models when the percentage of excess zeros is moderate or low, whereby the former is likely to perform better.

Chapter 7 proposes a general joint modeling framework for a longitudinal continuous sequence and an overdispersed, zero-inflated hierarchical count data. For the continuous end point, a Gaussian distributed subject specific random-effect is included to account for the correlation in the continuous component. For the count outcome, on the other hand, clustering and overdispersion are accommodated through inclusion of a Gaussian and gamma distributed random effects in a generalized linear model as proposed by Molenberghs *et al.* (2010). An excessive number of zero counts is accounted for using the so called zero-inflated or hurdle model adjustments. The association among the two sequences is captured through subject-specific random-effects which are allowed to correlate.

We analysed the Jimma Infant data, where body weight as well as number of days of diarrhoeal illnesses were measured repeatedly for each infant. The two outcomes were modelled jointly to capture association between them. The two end points show a strong inverse relationship as evidenced by the correlation of the random effects in the (NN-)(PNG). Furthermore, model fit was improved when random effects are allowed to correlate. Comparing the separate and joint models, while parameter estimates for the continuous outcome remain the same, small changes are observed in the count part. Turning to adjustments for zero-inflation either through hurdle or ZI approaches, the linear mixed model (NN-) is left unaltered because the excess zero adjustment aspect applies to the count process only. However, estimates of the count process change with their corresponding standard errors getting relatively smaller. Fitting a ZI(NN-)(PNG) model, even when correctly specified, is relatively more complex than a H(NN-)(PNG). The latter has several additional advantages, in particular the possibility to test for zero inflation. In the real data analysis, there were no model convergence issues, which is reassuring. Of course, as we learned from analysing these data, both overdispersion as well as additional zero inflation were convincingly present. Furthermore, the set of data was very large. These are comfortable conditions to reach convergence. In the simulation study, the hurdle model was at an advantage when it came to model fit. Practically, one may consider both approaches, hurdle and ZI, and perhaps use the relevant parameters from the hurdle model as starting values for the ZI fit (Kassahun *et al.*, 2013).

In conclusion, we note that our approach corrects for overdispersion and/or allows for joint modelling. In our example, both phenomena were present, although overdispersion results in a larger deviance reduction than joint modelling. One lesson to be drawn from this is that the user should carefully assess whether one or the other correction, both of them, or perhaps none of the two is necessary. Note also that joint modelling may be of interest in its own right. For example, one may be interested in measuring the strength of the association between both processes (estimating one or more correlation parameters) or in assessing its significance. Also, it is possible to derive prediction equations for an outcome or set of outcomes in one sequence, based on the outcomes in the other sequence and/or earlier measurements of the same sequence.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. New York: John Wiley & Sons, second edition.

Andy, H., Jane, A., Kelvin, K., and Geoffery, J. (2006). Multi-level zero-inflated modeling of correlated count data with excess zeros. *Epidemiology*, **15**:47–61.

Asefa, M. and Tessema, F. (2002). Infant survivorship and occurrence of multiple-births: A longitudinal community-based study south west ethiopia. *Ethiopian Journal of Health Development*, **16**:5–11.

Belachew, T., Haley, C., Lindtsrom, D., Gebremariam, A., Getachew, Y., Lachat, C., and Kolsteren, P. (2011). Gender differences in food insecurity and morbidity among adolescents in southwest ethiopia. *Pediatrics*, **127**:e397–e404.

Bergmann, K. E., Bergmann, R. L., Von Kries, R., Bohm, O., Richter, R., Dudenhausen, J. W., and Wahn, W. (2003). Early determinants of childhood overweight and adiposity in a birth cohort study: role of breast-feeding. *International Journal of Obesity*, **27**:162–172.

Booth, J., Casella, G., Friedl, H., and Hobert, J. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, **3**:179–191.

Breslow, N. (1984). Extra-poisson variation in log-linear models. *Applied Statistics*, **33**:38–44.

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

Breslow, N. E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**:81–91.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Computing Science and Statistics*, **7**:434–455.

Engel, B. and Keen, A. (1992). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**:1–22.

Faught, E., Wilder, B. J., Ramsay, R. E., Reife, R. A., Kramer, L. D., Pledger, G. W., and Karim, R. M. (1996). Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages. *Neurology*, **46**:1684–1690.

Freedman, D. S., Dietz, W. H., Srinivasan, S. R., and Berenson, G. S. (1999). The relation of overwight to cardiovascular risk factors among children and adolescents: The bogalusa heart study. *Pediatrics*, **103**:1175–1182.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis, Second Edition*. Boca Raton: Chapman & Hall/CRC, second edition.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Boca Raton: Chapman & Hall/CRC.

Gillman, M. W., Rifas-Shiman, S. L., Berkey, C. S., Frazier, A. L., Rockett, H. R., Camargo Jr, C. A., Field, A. E., and Colditz, G. A. (2006). Breast-feeding and overweight in adolescence: Within-family analysis. *Epidemiology*, **17**:112–114.

Goldstein, H. (2002). *Multilevel Statistical Models*. Oxford: Oxford University Press, third edition.

Greene, W. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models. working paper ec- 94-10, department of economics, new york university. *Working Paper*, pages 9–10.

Griswold, M. E. and Zeger, S. L. (2004). On marginalized multilevel models and their computation (november 2004). *Johns Hopkins University, Department of Biostatistics Working Paper ♯99*.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**:688–698.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, **15**:1–26.

Hinde, J. and Demétrio, C. G. B. (1998a). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, **27**:151–170.

Hinde, J. and Demétrio, C. G. B. (1998b). Overdispersion: Models and estimation. *São Paulo: XIII Sinape.*

Horrocks, J. and van den Heuvel, M. J. (2009). Prediction of pregnancy: A joint model for longitudinal and binary data. *Bayesian Analysis*, **4**:523–538.

Iddi, S. and Molenberghs, G. (2012). A combined overdispersed and marginalized multilevel model. *Computational Statistics and Data Analysis*, **56**:1944–1951.

Kassahun, W., Neyens, T., Faes, C., Molenberghs, G., and Verbeke, G. (2014a). A zero-inflated overdispersed and hirarchical poisson model. *Statistical Modelling, Accepted for Publication.*

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. Archives of Public Health*, http://dx.doi.10.1186/0778-7367-70-7.*

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2013). A joint model for hierarchical continuous and zero-inflated overdispersed count data. Journal of Statistical Computation and Simulation*, http://dx.doi.org/10.1080/00949655.2013.829058.*

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2014b). Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. Submitted for publication.

Kleinman, J. (1973). Proportions with extraneous variance: single and independent samples. Journal of the American Statistical Association*, **68**:46–54.*

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

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics*, **34**:1–14.*

Lawless, J. (1987). Negative binomial and mixed poisson regression. The Canadian Journal of Statistics*, **15**:209–225.*

Lee, A. H., Wang, K., Scott, J., Yau, K. K. W., and McLachlan, G. J. (2006). Multilevel zero-inflated poisson regression modelling of correlated count data with excess zeros. Statistical Methods in Medical Research*, **15**:47–61.*

Lee, K., Joo, Y., Song, J. J., and Harper, D. W. (2011). Analysis of zero-inflated clustered count data: a marginalized model approach. Computational Statistics and Data Analysis*, **55**:824–837.*

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

Macro., I. (2008). Nutrition of Young Children and Women Ethiopia 2005*. Maryland: Macro International.*

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models*. London: Chapman & Hall/CRC.*

Mei, Z., Grummer, L. M., Pietrobelli, A., Goulding, A., Goran, M. I., and Dietz, H. (2002). Validity of body mass index as compared with other body-composition screening indexes for the assessement of body fatness in children and adolescents. The American Journal of Clinical Nutrition*, **75**:978–985.*

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. Statistical Modelling*, **5**:1–19.*

Molenberghs, G., Kenward, M., Verbeke, G., Iddi, S., and Efendi, A. (2013). On the connections between bridge distributions, marginalized multilevel models, and generalized linear mixed models. International Journal of Statistics and Probability*, **2**.*

Molenberghs, G. and Verbeke, G. (2005). Models for Discrete Longitudinal Data*. New York: Springer.*

Molenberghs, G., Verbeke, G., and Demétrio, C. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime Data Analysis*, **13**:513–531.*

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.*

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics*, **33**:341–365.*

Neelon, B. H., Malley, A. J., and Normand, S. T. (2010). A bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. Statistical Modelling*, **4**:421–439.*

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society, Series B*, **135**:370–384.*

Owen, C. G., Martin, R. M., Whincup, P. H., Smith, G. D., and Cook, D. G. (2005). Effect of infant feeding on the risk of obesity across the life course: A quantitative review of published evidence. Official Journal of The Americal Academy of Pediatrics*, **115**:1367–1377.*

Pinheiro, J. C. and Bates, D. M. (2000). Mixed effects models in S and S-Plus. *New-York: Springer-Verlag.*

Skellam, J. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. Journal of the Royal Statistical Society, Series B*, **10**:257–261.*

Skrondal, A. and Rabe-Hesketh, S. (2004). Generalized Latent Variable Modeling. *London: Chapman & Hall/CRC.*

Spiegelhalter, D., Best, N., Carlin, B., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B*, **64**:583–639.*

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). WinBugs User Manual. *Second edition.*

Todem, D., Hsu, W. W., and Kim, K. M. (2012). On the efficiency of score tests for homogeneity in two-component parametric models for discrete data. Biometrics*, **68**:975–982.*

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. Statistica Sinica*, **14**:809–834.*

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. *New York: Springer, fourth edition.*

Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. *New York: Springer.*

Verbeke, G. and Molenberghs, G. (2010). Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. Statistical Modelling, ***10**:391–419.*

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, ***57**:307–333.*

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

WHO, O. (2009). World Health Statistics. *Switzerland: WHO Press.*

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. Journal of Statistical Computation and Simulation, ***48**:233–243.*

World, B. (2005). Education in Ethiopia: Strengthening the Foundation for Sustainable Progress. *Washington D. C.: The World Bank.*

Yewhalaw, D., Kassahun, W., Woldemichael, K., Tushune, K., Sudaker, M., Kaba, D., Duchateau, L., Wim Van Bortel, W., and Speybroeck, N. (2010). The influence of the gilgel-gibe hydroelectric dam in ethiopia on caregivers' knowledge, perceptions and health-seeking behaviour towards childhood malaria. Malaria Journal, ***9**:47.*

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. Biometrics, ***44**:1049–1060.*

Appendix A

SAS and WinBugs Codes for Overdispersed Hierarchical Binary Data

A.1 A SAS Program

A SAS NLMIXED program, for the Logist-normal model and combined model:

A.1.1 Jimma infants growth study

A.1.1.1 Combined model

```
proc nlmixed data=infant noad qpoints=10;
parms Beta_0 =-3.23 Beta_1=0.0602 beta_2=0.0402  Beta_3=-0.8369
Beta_4 =-0.552 Beta_5 =1.7266 Beta_6=-0.003 Beta_7=-0.0262
Beta_8=-0.0184 Beta_9=-0.1584 sd1=1.3662   sd2=0.2576  const=0.0944;
eta = Beta_0+b1+ (Beta_1+b2)*time + Beta_2*sex + Beta_3*(place=1)
+Beta_4*(place=2) + Beta_5*(Bf)+ Beta_6*(sex)*time+
Beta_7*time*(place=1)+ Beta_8*time*(place=2)+ Beta_9*time*(BF);
expeta = exp(eta);
ll = -log(1+const) + BMIBIN*eta - BMIBIN*log(1+expeta)
+ (1-BMIBIN)*log((1-expeta/(1+expeta)) + const);
model BMIBIN ~ general(ll);
```

```
random b1 b2 ~normal([0,0],[sd1**2,0,sd2**2]) subject=id;
run;
```

A.1.1.2 Logistic-normal model

```
proc nlmixed data=infant noad qpoints=10;
parms Beta_0 =-3.23 Beta_1=0.0602 beta_2=0.0402  Beta_3=-0.8369
Beta_4 =-0.552 Beta_5 =1.7266 Beta_6=-0.003 Beta_7=-0.0262
Beta_8=-0.0184 Beta_9=-0.1584 sd1=1.3662  sd2=0.2576;
eta = Beta_0+b1+ (Beta_1+b2)*time + Beta_2*sex + Beta_3*(place=1)
+Beta_4*(place=2)+Beta_5*(Bf)+ Beta_6*(sex)*time +
Beta_7*time*(place=1) + Beta_8*time*(place=2)+ Beta_9*time*(BF);
expeta = exp(eta);
p=expeta/(1+expeta) ;
model BMIBIN ~ binary(p);
random b1 b2 ~normal([0,0],[sd1**2,0,sd2**2]) subject=id;
run;
```

A.1.2 The Jimma Longitudinal Family Survey of Youth

A.1.2.1 Combined model

```
proc nlmixed data=ado noad qpoints=10 ;
parms Beta_0 =1.1652 Beta_1=0.04351 Beta_2=1.0911 Beta_3=1.1051
Beta_5=-1.2249 Beta_6=0.1471 Beta_7=0.3903 const=0.05 sd=0.5;
eta = Beta_0  +Beta_1*age+ Beta_2*(typplace=1)
+ Beta_3*(typplace=2) + Beta_5*currwork + Beta_6*sex
+Beta_7*round + b1;

expeta = exp(eta);
ll = -log(1+const) + currscho*eta - currscho*log(1+expeta)
+ (1-currscho)*log((1-expeta/(1+expeta)) + const);
model currscho ~ general(ll);
random b1~normal(0,sd*sd) subject=id ;
run;
```

A.1.2.2 Logistic-normal model

```
proc nlmixed data=ado noad qpoints=10 ;
```

```
parms Beta_0 =1.1652 Beta_1=0.04351 Beta_2=1.0911 Beta_3=1.1051
Beta_5=-1.2249 Beta_6=0.1471 Beta_7=0.3903 sd=0.5;
eta = Beta_0  +Beta_1*age+ Beta_2*(typplace=1)
+ Beta_3*(typplace=2) + Beta_5*currwork + Beta_6*sex
+Beta_7*round + b1;

expeta = exp(eta);
p=expeta/(1+expeta) ;
model currscho ~ binary(p);
random b1~normal(0,sd*sd) subject=id ;
run;
```

A.2 WinBugs Implementation

A WinBugs program, for the Logist-normal model and combined model:

A.2.1 Jimma infants growth study

A.2.1.1 Combined model

```
model {

for (i in 1:49112) {

BMIBIN[i]~dbern(p[i])
p[i]<-kappa[i]*theta[i]
theta[i]~dbeta(a,b)
logit(kappa[i]) <- alpha0 + (s[ID[i]]+alpha1)*TIME[i]
+alpha2*SEX[i]+alpha3*RUR[i]+alpha4*URB[i]+alpha5*BF[i]
+alpha6 * SEX[i]*TIME[i]+ alpha7 * RUR[i] *TIME[i]
+ alpha8*URB[i]*TIME[i]+alpha9*BF[i]*TIME[i]
+ u[ID[i]]

}
 for (j in 1:7969) {
 u[j] ~ dnorm(0.0,tau1)
 s[j]~ dnorm(0.0,tau2)
 }
a~dunif(3,5)
b~dunif(1.1,1.5)
c<-b/a
alpha0 ~ dnorm(0.0,1.0E-6)
alpha1 ~ dnorm(0.0,1.0E-6)
alpha2 ~ dnorm(0.0,1.0E-6)
alpha3 ~ dnorm(0.0,1.0E-6)
alpha4 ~ dnorm(0.0,1.0E-6)
alpha5 ~ dnorm(0.0,1.0E-6)
alpha6 ~ dnorm(0.0,1.0E-6)
alpha7 ~ dnorm(0.0,1.0E-6)
alpha8 ~ dnorm(0.0,1.0E-6)
alpha9~ dnorm(0.0,1.0E-6)
tau1~ dgamma(0.001,0.001)
tau2~ dgamma(0.001,0.001)
sd1<-sqrt(1/tau1)
sd2<-sqrt(1/tau2)
}
```

A.2.1.2 Logistic-normal model

```
model {

for (i in 1:49112) {

BMIBIN[i]~dbern(p[i])
p[i]<-kappa[i]
logit(kappa[i]) <- alpha0 + (s[ID[i]]+alpha1)*TIME[i]
+alpha2*SEX[i]+alpha3*RUR[i]+alpha4*URB[i]+alpha5*BF[i]
+alpha6 * SEX[i]*TIME[i]+ alpha7 * RUR[i] *TIME[i]
+ alpha8*URB[i]*TIME[i]+alpha9*BF[i]*TIME[i]
+ u[ID[i]]


}
```

```
 for (j in 1:7969) {
 u[j] ~ dnorm(0.0,tau1)
 s[j]~ dnorm(0.0,tau2)
 }

alpha0 ~ dnorm(0.0,1.0E-6)
alpha1 ~ dnorm(0.0,1.0E-6)
alpha2 ~ dnorm(0.0,1.0E-6)
alpha3 ~ dnorm(0.0,1.0E-6)
alpha4 ~ dnorm(0.0,1.0E-6)
alpha5 ~ dnorm(0.0,1.0E-6)
alpha6 ~ dnorm(0.0,1.0E-6)
alpha7 ~ dnorm(0.0,1.0E-6)
alpha8 ~ dnorm(0.0,1.0E-6)
alpha9~ dnorm(0.0,1.0E-6)
tau1~ dgamma(0.001,0.001)
tau2~ dgamma(0.001,0.001)
sd1<-sqrt(1/tau1)
sd2<-sqrt(1/tau2)
}
```

A.2.2 The Jimma Longitudinal Family Survey of Youth

A.2.2.1 Combined model

```
Model {

for (i in 1:3815) {
SCHO[i] ~ dbern(p[i])
p[i]<-theta[i]*kappa[i]
theta[i]~dbeta(a,b)
logit(kappa[i]) <- alpha0 + alpha1*AGE[i]+alpha2*URB[i]
+alpha3*SURB[i]+alpha4*WORK[i]+alpha5 * SEX[i]
+ alpha6 * ROUND[i]  + u[ID[i]]
}

 for (j in 1:1956)  {
```

```
 u[j] ~ dnorm(0,tau)

 }

a~dunif(110,210)
b~dunif(1.1,2.2)
c<-b/a
alpha0 ~ dnorm(0.0,1.0E-6)
alpha1 ~ dnorm(0.0,1.0E-6)
alpha2 ~ dnorm(0.0,1.0E-6)
alpha3 ~ dnorm(0.0,1.0E-6)
alpha4 ~ dnorm(0.0,1.0E-6)
alpha5 ~ dnorm(0.0,1.0E-6)
alpha6 ~ dnorm(0.0,1.0E-6)
 tau ~ dgamma(0.001,0.001)
sd<-1/sqrt(tau)

}
```

A.2.2.2 Logistic-normal model

```
Model {

for (i in 1:3815) {
SCHO[i] ~ dbern(p[i])
p[i]<-kappa[i]
logit(kappa[i]) <- alpha0 + alpha1*AGE[i]+alpha2*URB[i]
+alpha3*SURB[i]+alpha4*WORK[i]+alpha5 * SEX[i]
+ alpha6 * ROUND[i]  + u[ID[i]]

}

 for (j in 1:1956) {
 u[j] ~ dnorm(0,tau)
 }

alpha0 ~ dnorm(0.0,1.0E-6)
alpha1 ~ dnorm(0.0,1.0E-6)
alpha2 ~ dnorm(0.0,1.0E-6)
alpha3 ~ dnorm(0.0,1.0E-6)
alpha4 ~ dnorm(0.0,1.0E-6)
alpha5 ~ dnorm(0.0,1.0E-6)
alpha6 ~ dnorm(0.0,1.0E-6)
tau ~ dgamma(0.001,0.001)
sd<-1/sqrt(tau)
}
```

Appendix B

A SAS Program for the Zero-Inflated Models (Epilepsy Study)

/* y is the response variable (number of epileptic seizures) in the Epilepsy Data*/

B.1 (P--), ZI(P--)

B.1.1 (P--)

```
proc nlmixed data=epilepsy qpoints=20;
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1;
if (trt = 0) then eta = int0 + slope0*time;
else if (trt = 1) then eta = int1  + slope1*time;
lambda = exp(eta);
loglik=-lambda+y*eta-log(fact(y));
model y~ general(loglik);
estimate "difference in slope" slope1-slope0;
estimate "ratio of slopes" slope1/slope0;
run;
```

B.1.2 ZI(P--)

```
proc nlmixed data=epilepsy qpoints=20;
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 a0=0 a1=0;
```

```
eta_prob = a0+ a1*time ;
p_0 = exp(eta_prob) / (1 + exp(eta_prob));
if (trt = 0) then eta = int0 + slope0*time;
else if (trt = 1) then eta = int1  + slope1*time;
lambda = exp(eta);
if y = 0 then loglik = log(p_0 + (1 - p_0) * exp(-lambda));
else loglik = log(1 - p_0) + y * log(lambda)- lambda - lgamma(y+1);
model y~ general(loglik);
estimate "difference in slope" slope1-slope0;
estimate "ratio of slopes" slope1/slope0;
run;
```

B.2 (PN-), ZI(PN-)

B.2.1 (PN-)

```
proc nlmixed data=epilepsy qpoints=20;
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 sigma=1;
if (trt = 0) then eta = int0 + b + slope0*time;
else if (trt = 1) then eta = int1 + b + slope1*time;
lambda = exp(eta);
loglik=-lambda+y*eta-log(fact(y));
model nseizw ~ general(loglik);
random b ~ normal(0,sigma**2) subject = id;
estimate "difference in slope" slope1-slope0;
estimate "ratio of slopes" slope1/slope0;
estimate "variance RIs" sigma**2;
run;
```

B.2.2 ZI(PN-)

```
proc nlmixed data=epilepsy  qpoints=20;
parms int0=0.8179 slope0=-0.014 int1=0.647 slope1=-0.012
d11=0.98 rho=0 d22=1.10 a0=-3 a1=0.1;
eta_prob = a0+ a1*time+b2 ;
p_0 = exp(eta_prob) / (1 + exp(eta_prob));
if (trt = 0) then eta = int0 + b1 + slope0*time;
else if (trt = 1) then eta = int1 + b1 + slope1*time;
lambda = exp(eta);
if y = 0 then loglik = log(p_0 + (1 - p_0) * exp(-lambda));
else loglik = log(1 - p_0) + y * log(lambda) - lambda
- log(fact(y));
random b1 b2~normal([0,0],[d11**2,rho*d11*d22,d22**2]) subject = id;
model y ~ general(loglik);
estimate "difference in slope" slope1-slope0;
estimate "ratio of slopes" slope1/slope0;
estimate "variance d11" d11**2;
estimate "variance d22" d22**2;
run;
```

B.3 (P-G), ZI(P-G)

B.3.1 (P-G)

```
proc nlmixed data=epilepsy qpoints=20;
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 alpha=2;
if (trt = 0) then eta = int0 + slope0*time;
else if (trt = 1) then eta = int1 + slope1*time;
lambda = exp(eta);
beta=1/alpha;
loglik=lgamma(alpha+y)-lgamma(alpha)+y*log(beta)-
(y+alpha)*log(1+beta*lambda)+y*eta-lgamma(y+1);
model y ~ general(loglik);
estimate "difference in slope" slope1-slope0;
estimate "ratio of slopes" slope1/slope0;
estimate "beta=1/alpha" 1/alpha;
run;
```

B.3.2 ZI(P-G)

```
proc nlmixed data=epilepsy qpoints=20;
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 alpha=0.05 a0=-1 a1=0.1;
if (trt = 0) then eta = int0 + slope0*time;
else if (trt = 1) then eta = int1 + slope1*time;
lambda = exp(eta);
eta_prob=a0+a1*time;
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*lambda);
if y=0 then
ll = log(p_0+ (1-p_0)*(p**m));
else ll = log(1-p_0) + log(gamma(m + y)) - log(gamma(y + 1))
     - log(gamma(m)) + m*log(p) + y*log(1-p);
model y ~ general(ll);
estimate "difference in slope" slope1-slope0;
estimate "ratio of slopes" slope1/slope0;
estimate "beta=1/alpha" 1/alpha;
run;
```

B.4 (PNG), ZI(PNG)

B.4.1 (PNG)

```
proc nlmixed data=epilepsy  qpoints=20;
bounds alpha>0,sigma>0;
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 sigma=1 alpha=1 ;
f (trt = 0) then eta = int0 + b + slope0*time;
else if (trt = 1) then eta = int1 + b + slope1*time;
lambda = exp(eta);
beta=1/alpha;
loglik=lgamma(alpha+y)-lgamma(alpha)+y*log(beta)
-(y+alpha)*log(1+beta*lambda)+y*eta-lgamma(y+1);
random b ˜ normal(0,sigma**2) subject = id ;
model y˜ general(loglik);
predict lambda out=lamczc;
estimate "difference in slope" slope1-slope0;
estimate "ratio of slopes" slope1/slope0;
estimate "variance RIs" sigma**2;
estimate "beta=1/alpha" 1/alpha;
run;
```

B.4.2 ZI(PNG)

```
proc nlmixed data=epilepsy  qpoints=20;
parms int0= 0.8511 slope0=-0.01048 int1=0.8165 slope1=-0.008 alpha=0.2937
d11=1.0810 rho=0 d22=3.19 a0=-1.78 a1=0.052;
if (trt = 0) then eta = int0 + b1 + slope0*time;
else if (trt = 1) then eta = int1 + b1 + slope1*time;
lambda = exp(eta);
eta_prob = a0+a1*time+b2 ;
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*lambda);

if y=0 then
ll = log(p_0 + (1-p_0)*(p**m));
```

```
else ll = log(1-p_0) + log(gamma(m + y)) - log(gamma(y + 1))
     - log(gamma(m)) + m*log(p) + y*log(1-p);
model y ~ general(ll);
random b1 b2 ~ normal([0,0],[d11**2,rho*d11*d22,d22**2]) subject = id;
estimate "difference in slope" slope1-slope0;
estimate "ratio of slopes" slope1/slope0;
estimate "variance d11" d11**2;
estimate "variance d22" d22**2;
estimate "beta=1/alpha" 1/alpha;
run;
```

Appendix C

A SAS Program for Marginalized Models (IRC Dataset)

C.1 ZI(PNG) and MZI(PNG) Implementaion

C.1.1 ZI(PNG)$_\ell$

```
proc nlmixed data=IRC ;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma1=0.5 sigma2=0.5 alpha=1
      a0=0 a1=0 a2=0 a3=0 tau=0;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time+b1;
lambda_c =exp(eta);
eta_prob=a0+a1*time+a2*village+a3*season+b2;
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*lambda_c);
if   gamb=0 then
ll =  log(p_0 + (1-p_0)*(p**m));
else
ll = log(1-p_0)+lgamma(gamb+m)-lgamma(gamb+1)-lgamma(m) +
        gamb*log(alpha*lambda_c)-(gamb+m)*log(1+alpha*lambda_c);
model gamb ~ general(ll);
random b1 b2 ~normal([0,0],[sigma1**2,tau*sigma1*sigma2,sigma2**2])
       subject=id;
```

```
run;
```

Note: Special cases easily follow, for example, ZI(PN-)$_\ell$ can be fitted by replacing the above ZI(PNG)$_\ell$ likelihood by:

```
if   gamb=0 then
ll =  log(p_0 + (1-p_0)*exp(-lambda_c));
else
ll = log(1 - p_0) + gamb * log(lambda_c) - lambda_c - log(fact(gamb));
```

C.1.2 MZI(PNG)$_\ell$

```
proc nlmixed data=IRC;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma1=0.5 sigma2=0.5 alpha=1
      a0=0 a1=0 a2=0 a3=0 tau=0;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time;
eta_prob=a0+a1*time+a2*village+a3*season;
p_eta_prob=exp(eta_prob)/(1+exp(eta_prob));
delta1=eta-sigma1*sigma1/2;
delta2=sqrt(1+(sigma2*sigma2)) * probit(p_eta_prob);
lambda_c =exp(delta1+b1);
p_0=probnorm(delta2+b2);
m = 1/alpha;
p = 1/(1+alpha*lambda_c);
if   gamb=0 then
ll =  log(p_0 + (1-p_0)*(p**m));
else
ll = log(1-p_0)+lgamma(gamb+m)-lgamma(gamb+1)-lgamma(m) +
        gamb*log(alpha*lambda_c)-(gamb+m)*log(1+alpha*lambda_c);
model gamb ~ general(ll);
random b1 b2 ~normal([0,0],[sigma1**2,tau*sigma1*sigma2,sigma2**2])
             subject=id;
run;
```

In fitting MZI(PNG)$_\ell$, logit link for the marginal, and probit link for the conditional model were used in the zero-inflation part, by making use of their connection, as discussed in Section 6.1.

C.1.3 ZI(PNG)$_p$

```
proc nlmixed data=IRC  ;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma1=0.5 sigma2=0.5 alpha=1
      a0=0 a1=0 a2=0 a3=0 tau=0;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time;
lambda_c =exp(eta+b1);
eta_prob=a0+a1*time+a2*village+a3*season;
p_0=probnorm(eta_prob+b2);
m = 1/alpha;
p = 1/(1+alpha*lambda_c);
if   gamb=0 then
ll =  log(p_0 + (1-p_0)*(p**m));
else
ll = log(1-p_0)+lgamma(gamb+m)-lgamma(gamb+1)-lgamma(m) +
        gamb*log(alpha*lambda_c)-(gamb+m)*log(1+alpha*lambda_c);
model gamb ~ general(ll);
random b1 b2 ~normal([0,0],[sigma1**2,tau*sigma1*sigma2,sigma2**2])
         subject=id;
run;
```

C.1.4 MZI(PNG)$_p$

```
proc nlmixed data=IRC;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma1=0.5 sigma2=0.5 alpha=1
      a0=0 a1=0 a2=0 a3=0 tau=0;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time;
eta_prob=a0+a1*time+a2*village+a3*season;
p_eta_prob=probnorm(eta_prob);
delta1=eta-sigma1*sigma1/2;
delta2=sqrt(1+(sigma2*sigma2)) * probit(p_eta_prob);
lambda_c =exp(delta1+b1);
p_0=probnorm(delta2+b2);
m = 1/alpha;
p = 1/(1+alpha*lambda_c);
if   gamb=0 then
ll =  log(p_0 + (1-p_0)*(p**m));
else
ll = log(1-p_0)+lgamma(gamb+m)-lgamma(gamb+1)-lgamma(m) +
        gamb*log(alpha*lambda_c)-(gamb+m)*log(1+alpha*lambda_c);
model gamb ~ general(ll);
random b1 b2 ~normal([0,0],[sigma1**2,tau*sigma1*sigma2,sigma2**2])
       subject=id;
run;
```

C.2 H(PNG) and MH(PNG) Implementaion

C.2.1 H(PNG)$_\ell$

```
proc nlmixed data=IRC ;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma1=0.5 sigma2=0.5 alpha=1
a0=0  a1=0 a2=0 a3=0 tau=0;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time;
lambda_c =exp(eta+b1);
eta_prob=a0+a1*time+a2*village+a3*season;
p_0=exp(eta_prob+b2)/(1+exp(eta_prob+b2));
m = 1/alpha;
p = 1/(1+alpha*lambda_c);
if gamb=0 then ll = log(p_0);
else ll = log(1-p_0) + log(gamma(m + gamb)) - log(gamma(gamb + 1))
     - log(gamma(m)) + gamb*log(alpha*lambda_c)-
             (gamb+m)*log(1/p)-log(1-(1/p)**(-m));
model gamb ~ general(ll);
random b1 b2 ~normal([0,0],[sigma1**2,tau*sigma1*sigma2,sigma2**2])
        subject=id;
run;
```

C.2.2 MH(PNG)$_\ell$

```
proc nlmixed data=IRC ;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma1=1 sigma2=1 alpha=1
      a0=0 a1=0 a2=0 a3=0 tau=0;
bounds sigma1>0,sigma2>0,alpha>0;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time;
eta_prob=a0+a1*time+a2*village+a3*season;
p_eta_prob=exp(eta_prob)/(1+exp(eta_prob));
delta1=eta-sigma1*sigma1/2;
delta2=sqrt(1+(sigma2*sigma2)) * probit(p_eta_prob);
lambda_c =exp(delta1+b1);
p_0=probnorm(delta2+b2);
m = 1/alpha;
p = 1/(1+alpha*lambda_c);
if gamb=0 then ll = log(p_0);
else ll = log(1-p_0) + log(gamma(m + gamb)) - log(gamma(gamb + 1))
     - log(gamma(m)) + gamb*log(alpha*lambda_c)-(gamb+m)*log(1/p)
        -log(1-(1/p)**(-m));
model gamb ~ general(ll);
random b1 b2 ~normal([0,0],[sigma1**2,tau*sigma1*sigma2,sigma2**2])
             subject=id;
run;
```

C.2.3 H(PNG)$_p$

```
proc nlmixed data=IRC ;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma1=0.5 sigma2=0.5 alpha=1
      a0=0 a1=0 a2=0 a3=0 tau=0;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time;
lambda_c =exp(eta+b1);
eta_prob=a0+a1*time+a2*village+a3*season;
p_0=probnorm(eta_prob+b2);
m = 1/alpha;
p = 1/(1+alpha*lambda_c);
if gamb=0 then ll = log(p_0);
else ll = log(1-p_0) + log(gamma(m + gamb)) - log(gamma(gamb + 1))
     - log(gamma(m)) + gamb*log(alpha*lambda_c)-(gamb+m)*log(1/p)
             -log(1-(1/p)**(-m));
model gamb ~ general(ll);
random b1 b2 ~normal([0,0],[sigma1**2,tau*sigma1*sigma2,sigma2**2])
             subject=id;
run;
```

C.2.4 MH(PNG)$_p$

```
proc nlmixed data=IRC6 ;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma1=0.5 sigma2=0.5 alpha=1
      a0=0 a1=0 a2=0 a3=0 tau=0;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time;
eta_prob=a0+a1*time+a2*village+a3*season;
p_eta_prob=probnorm(eta_prob);
delta1=eta-sigma1*sigma1/2;
delta2=sqrt(1+(sigma2*sigma2)) * probit(p_eta_prob);
lambda_c =exp(delta1+b1);
p_0=probnorm(delta2+b2);
m = 1/alpha;
p = 1/(1+alpha*lambda_c);
if gamb=0 then ll = log(p_0);
else ll = log(1-p_0) + log(gamma(m + gamb)) - log(gamma(gamb + 1))
    - log(gamma(m)) + gamb*log(alpha*lambda_c)-(gamb+m)*log(1/p)
        -log(1-(1/p)**(-m));
model gamb ~ general(ll);
random b1 b2 ~normal([0,0],[sigma1**2,tau*sigma1*sigma2,sigma2**2])
             subject=id;
```

C.3 M(PNG) and M(PN-) Implementaion

C.3.1 M(PNG)

```
proc nlmixed data=IRC  ;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma=1 alpha=1;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time;
delta=eta-sigma*sigma/2;
lambda_c =exp(delta+b);
m = 1/alpha;
p = 1/(1+alpha*lambda_c);
ll = lgamma(gamb+m)-lgamma(gamb+1)-lgamma(m) +
        gamb*log(alpha*lambda_c)-(gamb+m)*log(1+alpha*lambda_c);
model gamb ~ general(ll);
random b ~normal(0,sigma**2) subject=id;
run;
```

C.3.2 M(PN-)

```
proc nlmixed data=IRC  ;
parms b_0=0  b_1=1 b_2=0 b_3=2 b_4=0 sigma=1;
eta = b_0 + b_1*village +b_2*time+b_3*season+b_4*Village*time;
delta=eta-sigma*sigma/2;
lambda_c =exp(delta+b);
ll=-lambda_c+gamb*log(lambda_c)-  log(fact(gamb));
model gamb ~ general(ll);
random b ~normal(0,sigma**2) subject=id;
run;
```

Appendix D

A SAS Program for the Joint Model

/* resp is the response variable*/
/*name='1' is an indicator for the count sequence
and name='2' is an indicator for the continuous sequence*/

D.1 ZI(NN-)(PNG), H(NN-)(PNG) and special cases

D.1.1 (NN-) & (PNG)

```
proc nlmixed data=joint   qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.03359 sigma=0.7202
beta21=-1.1599 beta22=0.2250  tau1=0.6845 tau2=0.3699 alpha=17;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1 ;
dens = -0.5*log(3.14159265358) - log(sigma)
    -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age+b2 ;
expeta=exp(eta);
m = 1/alpha;
p = 1/(1+alpha*expeta);
```

```
ll = lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
        resp*log(alpha*expeta)-(resp+m)*log(1/p);
end;
model resp ~ general(ll);
random b1 b2~normal([0,0],[tau1**2,0,tau2**2]) subject=id;

run;
```

D.1.2 (NN-)(PNG)

```
proc nlmixed data=joint  qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.03359  sigma=0.7202
beta21=-1.1599 beta22=0.2250  tau1=0.6845 tau2=0.3699 rho=-0.1
alpha=17;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age+b1 ;
dens = -0.5*log(3.14159265358) - log(sigma)
    -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age+b2 ;
expeta=exp(eta);
m = 1/alpha;
p = 1/(1+alpha*expeta);
ll = lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
        resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta);
end;
model resp ~ general(ll);
random b1 b2~normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2])
            subject=id;
run;
```

D.1.3 (NN-) & H(PNG)

```
proc nlmixed data=joint  qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033  sigma=0.76
beta21=2.0484 beta22=0.01788   tau1=0.688 tau2=0.48 alpha=0.333
a0=2.04 a1=-0.037;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
    -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age + b2;
expeta=exp(eta);
eta_prob=a0+a1*age;
expeta_prob=exp(eta_prob);
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if   resp=0 then do;
ll =  eta_prob-log(1+expeta_prob);
end;
else do;
ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
        resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta)
        - log(1 -( 1 + alpha*expeta)**(-m));
end;
end;
model resp ~ general(ll);
random b1 b2~normal([0,0],[tau1**2,0,tau2**2]) subject=id;
run;
```

D.1.4 H(NN-)(PNG)

```
proc nlmixed data=joint  qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033  sigma=0.76
beta21=2.0484 beta22=0.01788   tau1=0.688 tau2=0.48 rho=-0.1
alpha=0.333 a0=2.04 a1=-0.037;

if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
    -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;

if name = "1" then do;
eta = beta21 + beta22*age + b2;

expeta=exp(eta);
eta_prob=a0+a1*age;
expeta_prob=exp(eta_prob);
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if   resp=0 then do;
ll =  eta_prob-log(1+expeta_prob);end;
else do;

ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
        resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta)
        - log(1 -( 1 + alpha*expeta)**(-m));
end;
end;

model resp ˜ general(ll);
random b1 b2 ˜normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2])
           subject=id;
run
```

D.1.5 (NN-) & ZI(PNG)

```
 proc nlmixed data=joint  qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033  sigma=0.76
beta21=2.0484 beta22=0.01788   tau1=0.688 tau2=0.48 alpha=0.333
a0=2.04 a1=-0.037;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age+ b1;
dens = -0.5*log(3.14159265358) - log(sigma)
    -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age+ b2;
expeta=exp(eta);
eta_prob=a0+a1*age;
expeta_prob=exp(eta_prob);
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if   resp=0 then do;
ll =  log(p_0 + (1-p_0)*(p**m));end;
else do;
ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
     resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta);
end;
end;
model resp ~ general(ll);
random b1 b2 ~normal([0,0],[tau1**2,0,tau2**2]) subject=id;
run;
```

D.1.6 ZI(NN-)(PNG)

```
proc nlmixed data=joint   qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033  sigma=0.76
beta21=2.0484 beta22=0.01788 tau1=0.688 tau2=0.48 rho=-0.1 alpha=0.333
a0=2.04 a1=-0.037;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
    -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age + b2;
expeta=exp(eta);
eta_prob=a0+a1*age;
expeta_prob=exp(eta_prob);
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if   resp=0 then do;
ll =  log(p_0 + (1-p_0)*(p**m));
end;
else do;
ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
      resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta);
end;
end;
model resp ~ general(ll);
random b1 b2~normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2])
            subject=id;
run;
```

D.2 Test of Zero-inflation

/* Comparing H(NN-)(PNG) with a1 versus H(NN-)(PNG) without a1*/

D.2.1 H(NN-)(PNG) without a_1

```
 roc nlmixed data=joint   qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033  sigma=0.76
beta21=2.0484 beta22=0.01788 tau1=0.688 tau2=0.48 rho=-0.1 alpha=0.333;

if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
    -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;

if name = "1" then do;
eta = beta21 + beta22*age + b2;

expeta=exp(eta);
eta_prob=beta21+beta22*age+b2;
expeta_prob=exp(eta_prob);

p_0=exp(-exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if   resp=0 then do;
ll =  eta_prob-log(1+expeta_prob);end;
else do;

ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
        resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta)
        - log(1 -( 1 + alpha*expeta)**(-m));
end;
end;
```

```
model resp ~ general(ll);
random b1 b2~normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2])
           subject=id;

run;
```

D.2.2 H(NN-)(PNG) with a_1

```
 roc nlmixed data=joint   qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033  sigma=0.76
beta21=2.0484 beta22=0.01788 tau1=0.688 tau2=0.48 rho=-0.1 alpha=0.333
a1=0;

if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
    -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;

if name = "1" then do;
eta = beta21 + beta22*age + b2;

expeta=exp(eta);
eta_prob=a1+beta21+beta22*age+b2;
expeta_prob=exp(eta_prob);
p_0=1-exp(-exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if   resp=0 then do;
ll =  eta_prob-log(1+expeta_prob);end;
else do;

ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
        resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta)
        - log(1 -( 1 + alpha*expeta)**(-m));
end;
end;

model resp ~ general(ll);
random b1 b2~normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2])
            subject=id;
```

```
run
```

D.3 Simulation Study

D.3.1 Scenario One

```
data jointsim1;
call streaminit(1234);
do ss=1 to 250 ;
mean1=0; /*mean for b1*/
mean2=0; /*mean for b2*/
sig1=2; /*SD for b1*/
sig2=1.5; /*SD for b2*/
rho=-0.5; /*Correlation between b1 and b2*/
do kk=1 to 200;
r1 = rannor(1245);
r2 = rannor(2923);
b1 = mean1 + sig1*r1;
      /*b1 random effect for continuous part*/
b2 = mean2 + rho*sig2*r1+
     sqrt(sig2**2-sig2**2*rho**2)*r2;
    /*b2 random effect for count part*/
do TT=1 to 10; /*10 time points*/
sim=ss;
e=rand(normal,0,0.6);
id=kk;
age=TT;
mu=3.3+0.77*age-0.03*age*age+b1+e; /* continuous part*/
kappa = exp(2 + 0.02*age+b2);    /* count part*/

theta = 1;

parm1 = 1/(1+kappa/theta);
yneg = rand(NEGB,parm1,theta);
p1=2-0.2*age;   /* zero-inflation part*/
```

```
p2=exp(p1);
p=p2/(1+p2);
inf=rand(bern,p);
if inf=1 then do;
ynegzim=0;
end;
else do;
ynegzim=yneg;
end;
numdays=ynegzim;
output ;
end;
end ;
end;
data jointsim1;
set jointsim1;
run;
```

D.3.2 Scenario Two

```
data jointsim2;
call streaminit(1234);
do ss=1 to 250 ;
mean1=0; *mean for b1;
mean2=0; *mean for b2;
sig1=2; *SD for b1;
sig2=1.5; *SD for b2;
rho=-0.5; *Correlation between b1 and b2;
do kk=1 to 200;
r1 = rannor(1245);
r2 = rannor(2923);
b1 = mean1 + sig1*r1;
b2 = mean2 + rho*sig2*r1+sqrt(sig2**2-sig2**2*rho**2)*r2;
do TT=1 to 10; /*10 time points*/
sim=ss;
e=rand(normal,0,0.6);
id=kk;
```

```
age=TT;
mu=3.3+0.77*age-0.03*age*age+b1+e;
kappa = exp(2 + 0.02*age+b2);
ypois = rand(POISSON,kappa);
p1=2-0.2*age;
p2=exp(p1);
p=p2/(1+p2);
inf=rand(bern,p);
if inf=1 then do;
ypoiszim=0;
end;
else do;
ypoiszim=ypois;
end;
numdays=ypoiszim;
output ;
end;
end ;
end;
data jointsim2;
set jointsim2;
run;
```

D.3.3 Scenario Three

```
data jointsim3;
call streaminit(1234);
do ss=1 to 250 ;
mean1=0; *mean for b1;
mean2=0; *mean for b2;
sig1=2; *SD for b1;
sig2=1.5; *SD for b2;
rho=-0.5; *Correlation between b1 and b2;
do kk=1 to 200;
r1 = rannor(1245);
r2 = rannor(2923);
```

```
b1 = mean1 + sig1*r1;
b2 = mean2 + rho*sig2*r1+sqrt(sig2**2-sig2**2*rho**2)*r2;
do TT=1 to 10; /*10 time points*/
sim=ss;
e=rand(normal,0,0.6);
id=kk;
age=TT;
mu=3.3+0.77*age-0.03*age*age+b1+e;
kappa = exp(2 + 0.02*age+b2);

theta = 1;

parm1 = 1/(1+kappa/theta);
yneg = rand(NEGB,parm1,theta);

numdays=yneg;
output ;
end;
end ;
end;
data jointsim3;
set jointsim3;
run;
```

Appendix E

Posterior Densities

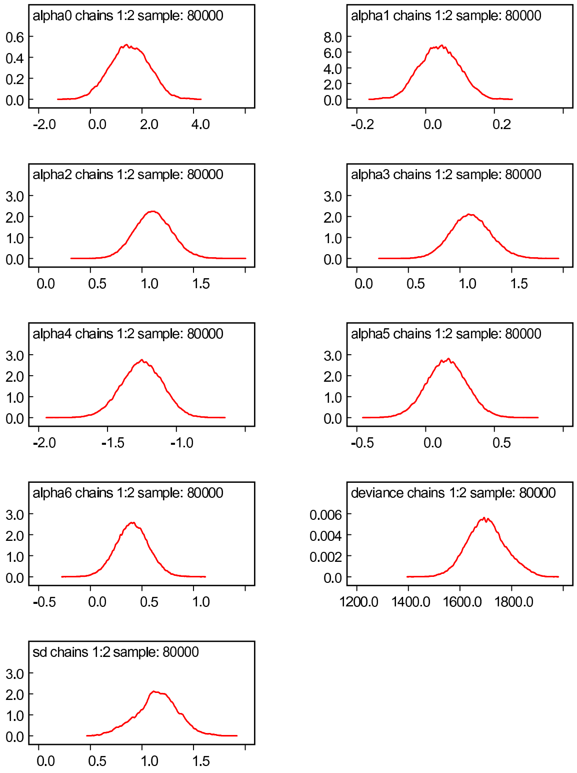

Figure E.1: *Jimma Infants Growth Study, posterior density for the logistic-normal model (pD= 5400.7)*

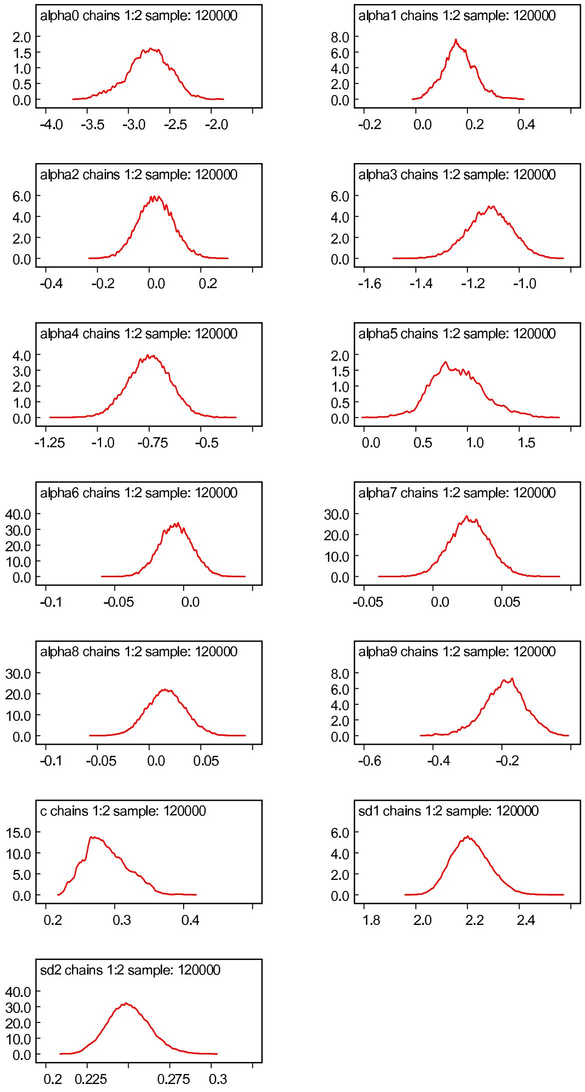

Figure E.2: *Jimma Infants Growth Study, posterior density for the combined model (pD= 6218.3)*

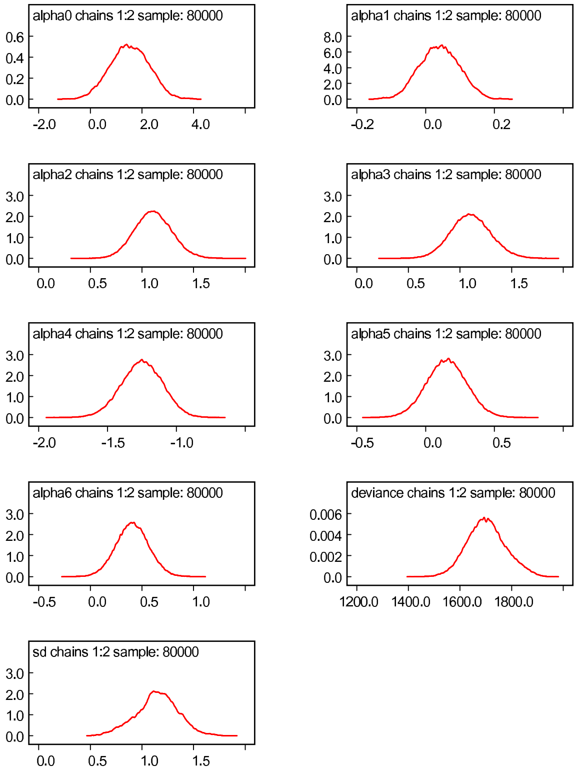

Figure E.3: *Jimma Longitudinal Family Survey, posterior density for the logistic-normal model (pD= 211.9)*

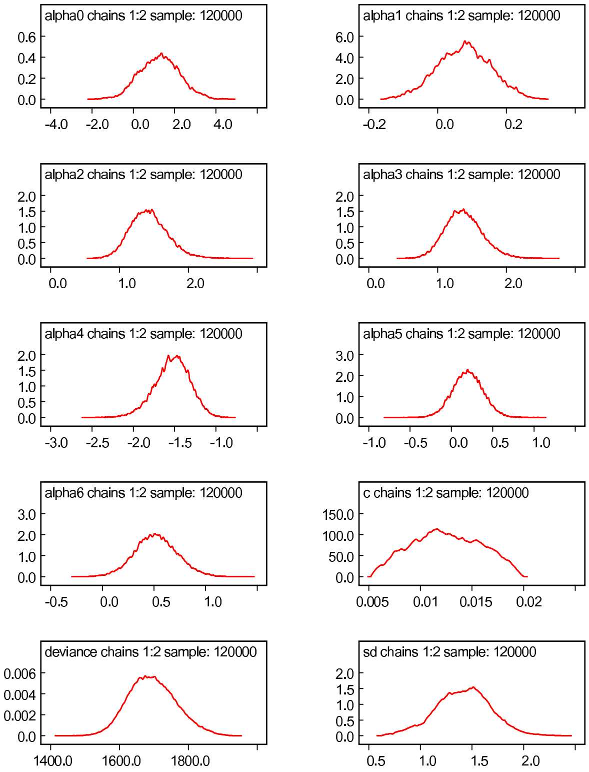

Figure E.4: *Jimma Longitudinal Family Survey, posterior density for the combined model (pD= 241.5)*