ABSTRACT

Title of Dissertation: FIRST ORDER AUTOREGRESSIVE MIXED EFFECTS ZERO INFLATED POISSON MODEL FOR LONGITUDINAL DATA – A BAYESIAN APPROACH

Chin-Fang Weng, Doctor of Philosophy, 2014

Dissertation Directed by: Professor Robert J. Mislevy
Department of Human Development and Quantitative Methodology

The First Order Autoregressive (AR(1)) Mixed Effects Zero Inflated Poisson (ZIP) Model was developed to analyze longitudinal zero inflated Poisson data through the Bayesian Approach. The model describes the effect of covariates via regression and time varying correlations within subject. Subjects are classified into a "perfect" state with response equal to zero and a Poisson state with response following a Poisson regression model. The probability of belonging to the perfect state or Poisson state is governed by a logistic regression model. Both models include autocorrelated random effects, and there is correlation between random effects in the logistic and Poisson regressions.

Parameter estimation is investigated using simulation studies and analyses (both frequentist and Bayesian) of simpler mixed effect models. In the large sample setting we investigate the Fisher information of the model. The Fisher information matrix is then used to determine an adequate sample size for the AR(1) ZIP model. Simulation studies demonstrate the capability of Bayesian methods to estimate the parameters of the AR(1)

ZIP model for longitudinal zero inflated Poisson data. However, a tremendous computation time and a huge sample size are required by the full AR(1) ZIP model.

Simpler models were fitted to simulated AR(1) ZIP data to investigate whether simplifying the assumed random structure could permit accurate estimates of fixed effect parameters. However, simulations showed that the bias of two nested models, ZIP model and mixed effects ZIP model, are too large to be acceptable. The AR(1) ZIP model was fitted to data on numbers of cigarettes smoked, collected in the National Longitudinal Study of Youth. It was found that decisions on whether to smoke and on the number of cigarettes to smoke were significantly related to age, sex, race and smoking behavior by peers. The random effect variances, autocorrelation coefficients and correlation between logistic and Poisson random effect were all significant.

FIRST ORDER AUTOREGRESSIVE MIXED EFFECTS ZERO INFLATED POISSON MODEL ON LONGITUDINAL DATA – A BAYESIAN APPROACH

by

Chin-Fang Weng

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory committee:

Dr. Robert J. Mislevy, Chair
Dr. Eric V. Slud
Dr. George Macready
Dr. Hong Jiao
Dr. Jeffrey R. Harring

DEDICATION

To Paul

ACKNOWLEDGMENTS

First and foremost, I must thank my advisor, Professor Robert Mislevy, for all his help and kindness while I was working on my dissertation. Dr. Mislevy awakened my interest in Bayesian statistics, allowed me to choose my own research path, let me be an independent researcher, and stood by me throughout my career as a doctoral student. I am grateful to him beyond what words can express.

I owe a debt to Professor Eric Slud, who served on my Ph.D. committee and was my M.A. advisor in the STAT Program. His ideas and suggestions greatly improved my dissertation, and he was always available for advice.

Professors Jeffrey Harring, Hong Jiao and George Macready were kind enough to serve on my doctoral committee. I appreciate their careful reading of this dissertation and their helpful and valuable suggestions.

Professor James Yorke of Mathematics was my Advanced Calculus teacher and a constant source of encouragement. He inspired me not only in my Master's studies in STAT but also during my Ph.D. program in EDMS.

I thank my colleague Mark Bauder for useful discussions that helped me to sharpen my thinking and understand my problem more deeply.

Min Min gave me invaluable support and encouragement. She is the perfect definition of a good friend.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

1.1 Motivation

It is commonly seen that a set of count data contains an excess of zeros relative to standard distributions for such data, so that data analyses encounter an over-dispersion problem. Such zero inflated data appear in many fields, such as social science, medical research, and industrial processes. Many researchers have studied this problem and developed various zero-inflated models in response. This dissertation concerns extensions of the zero-inflated Poisson (ZIP) model.

1.2 Theme of the Dissertation

The ZIP model originally was introduced by Lambert (1992). The ZIP model treated the zero inflated count data as coming from a mixture distribution: one component is a Poisson distribution, where **y** = 0, 1, 2, …, and the other component is a zero state, which generates zero counts only. Note that, when a count is zero, it may have come from either the Poisson state or from the zero state. This is why the ordinary Poisson model has an overdispersion problem when applied to ZIP data. Sometimes, the zeros can be 60%, 70%, or higher in the data set, which introduces bias into an analysis using a standard model for counts, such as the Poisson.

In the ZIP model, there are two Generalized Linear Models (GLMs): a logistic model and a Poisson model. The logistic model yields a binary outcome and filters the count data into two classes, either from the Poisson class or from the zero class. The Poisson model component works as an ordinary Poisson regression; that is, the log of the

Poisson rate parameter can be modeled as a linear function of covariates. Note that these two models work simultaneously.

Later the ZIP model was extended to the mixed effect ZIP model (Neelon O'Malley, & Normand, 2010) by introducing random effects in to the model; that is, a random intercept logistic model and a random intercept Poisson model (a General Linear Mixed Model, or GLMM). With the random effects, the mixed effect ZIP model can address the correlation within subjects encountered in analyses of repeated measurements data, clustered data, or meta-analysis.

In this study, the mixed effect ZIP model is extended to a first-order regressive or AR(1) mixed effect ZIP model. That is, the ZIP model additionally has an AR(1) random intercept logistic model and an AR(1) random intercept Poisson model with respect to the time-ordered aspect of observations. The AR(1) structure of the random effects thus serves to model the correlation arising from time waves in longitudinal data. Moreover, the AR(1) random effects in the logistic and Poisson models can be correlated. The AR(1) mixed effect ZIP model is an extremely general model such that previously published models are special cases of AR(1) mixed effect ZIP model. A thorough literature review has shown no prior publication of work on an AR(1) mixed effect ZIP model.

Our investigations of this new model have shown estimation within a Bayesian framework to be possible but arduous. The focus of the dissertation research therefore is to understand the importance of including all random components, in the sense of robustness. That is, if data have been generated by a process described by the full model, what are the qualities of estimates of fixed-effects parameters in the full model (which are

typically the effects of primary interest in applied work) in comparison with successively simpler sub-models?

Table 1.1 presents a hierarchy of models that we will examine. The Poisson Model is the simplest one, and the model structural complexity is increased from left to the right, to the AR(1) Mixed Effect ZIP model. We want to investigate the robustness of the AR(1) model in terms of regression coefficient estimations.

Table 1.1 Hierarchy of ZIP Models

Components	Poisson Model	ZIP Model	Mixed Effect ZIP Model	AR(1) Mixed Effect ZIP Model
GLM (Poisson)	X	X		
GLM (Logistic)		X		
GLMM(Poisson)			X	X
GLMM(Logistic)			X	X
Corr of GLMMs			X	X
AR(1)				X

1.3 Contributions

This dissertation contains the following contributions:

a) We develop a new model – the AR(1) ZIP model – which can better fit longitudinal count data.

b) The AR(1) model is implemented by a Bayesian estimation method through BUGS software.

c) The AR(1) ZIP model requires large samples. We show how to approximate the Fisher Information of the model and use it to determine the needed sample size.

d) For simulated data generated according to the AR(1) ZIP model, simulation studies show that simpler models (ZIP with no random effects or ZIP with random intercepts) do not yield valid estimates of regression parameters.

Chapter II

Literature Review

2.1 Development of Analyses of Zero Inflated Count Data

Poisson regression models provide a standard framework for the analysis of count data. However, in practice, count data are often overdispersed compared to the Poisson distribution with the same mean. One cause of overdispersion is that the zero counts exceed the expected frequency in the Poisson distribution. Examples of zero-inflated data appear in many fields, such as road safety (Miaou, 1994), sexual behavior (Heilbron, 1994), and reading tests (Jansen, 1986). The excessive zero counts happen when the data contains both structural zeros and sampling zeros (Ridout, Demetrio, & Hide, 1998). Suppose that the underlying population is a mixture of two groups, say perfect and imperfect. Members of the perfect class will produce zero errors only. That is, one mechanism generates only zeros, while the other process generates both zero and nonzero counts. When this situation happens, the standard Poisson model will not describe this data well and will have an overdispersion problem. In other words, there are too many counts at higher values in relation to the Poisson mean, and thus to the best-fitting Poisson. An example comes from a manufacturing production process with the following characteristics: the production process is in a near perfect state so that zero defects are observed with high probability. However when the environment changes to an imperfect state, defective units may be produced. The environmental changes are unobservable and random. In this practical case, the proportion of observed zeros exceeds what a Poisson model would predict.

Several classes of discrete mixture models have been proposed for zero-modified data. These include hurdle models (Mullahy, 1986; Heilbron, 1994) and zero-inflated models (Lambert, 1992). The hurdle model is a two-stage model for count data. One part is a binary model for whether the response outcome is zero or positive. Conditional on a positive outcome, the second part uses a truncated count model that modifies an ordinary distribution by conditioning on a positive outcome. For instance, this might be a truncated Poisson distribution or a truncated negative binomial. The hurdle model can handle zero inflation or deflation. The zero-inflated model (Lambert, 1992) only handles zero inflation. Two types of zeros can occur: one comes from the zero state and the other from the ordinary count distribution state. That is, the distribution is a mixture of an ordinary count model, such as the Poisson or negative binomial, with one that is degenerate at zero. Min and Agresti (2005) argue that zero-inflated count models are more natural than a hurdle model and that one should think of the population as a mixture.

Lambert (1992) proposed the Zero Inflated Poisson (ZIP) model, which contains logistic regression and Poisson regression components, and additional applications appeared. Based on the features of data, extended ZIP models were developed. Ghosh, Mukhopadhyay, & Lu (2006) extend the ZIP models to a broad class of distributions (e.g., power series distributions) to fit zero-inflated data with a relatively small to moderate sample size. Some data observations are correlated, such as repeated measurements, spatially correlated data, or longitudinal data, so random effects were introduced into models. Hall (2000) incorporated one random intercept effect into the Poisson regression only, but not in the logistic regression, to account for correlated

responses in a repeated measures designed experiment. Neelon, O'Malley, & Normand (2010) included correlated random effects in both components of the model.

Many authors took frequentist approaches to estimate the parameters in ZIP models. In the case of a mixed effect ZIP model the computational techniques are complicated and require approximations of high dimensional integrals (Hall, 2000; Min & Agresti, 2005). For applications in meta-analysis (as encountered in the author's work; Weng, 2008), it was not until 2008 that Bates created the logistic random intercept computation function, glmer ( ), provided in R software (R, 2008). Recent versions of SAS software can fit random effect GLM models using PROC GLMMIX. Certain more general mixed effect regression models, but neither the ZIP model nor the mixed effect ZIP model can be estimated using SAS PROC NLMIXED.

2.2 Poisson Model and Zero Inflated Poisson Model

2.2.1 Generalized Linear Models: Poisson Model and Logistic Regression Model

The unity of many statistical methods was demonstrated by Nelder and Wedderburn (1972) using the idea of the Generalized Linear Model (GLM). GLMs extend ordinary regression models to encompass nonnormal response distributions and modeling functions of the mean (Agresti, 2002). That is, response variables may have distributions other than the normal distribution – they may even be categorical rather than continuous. Also, the relationship between the response and explanatory variables need not be of the linear form.

Consider the Poisson regression model: in the Poisson distribution, the probability function for the discrete random variable Y is $f(y,\theta) = \frac{\theta^y e^{-\theta}}{y!}$. where y takes the values

0, 1, 2, …. This can be rewritten as $f(y,\theta) = \exp(y\log\theta - \theta - \log(y!))$. In a sample $(\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, N$, the Y_i are independent Poisson random variables with expected value $E[Y_i] = \mu_i = \exp(\boldsymbol{X}_i'\boldsymbol{\beta})$, where $\boldsymbol{X}_i$ is a covariate vector and $\boldsymbol{\beta}$ is an unknown parameter vector. The link function, $\eta_i = g(\mu_i) = \boldsymbol{X}_i'\boldsymbol{\beta}$, relates the mean response μ_i to $\boldsymbol{X}_i$. Consider the logistic regression model: the Y_i are independent Bernoulli random variables with expected values $E[Y_i] = \pi_i = \frac{\exp(\boldsymbol{X}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i'\boldsymbol{\beta})}$. There, the logit transformation, $\eta = \log\left(\frac{\pi}{1-\pi}\right)$, serves to link the linear predictor, $\eta_i = \boldsymbol{X}_i'\boldsymbol{\beta}$, to the mean response: $g(\mu_i) = g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{X}_i'\boldsymbol{\beta}$. Note that π_i lies in the interval [0, 1] and η_i lies in $(-\infty, +\infty)$.

2.2.2 Zero Inflated Poisson Model

One often assumes that a set of count data has a Poisson distribution. In many practical cases, however, the proportion of observed zeros exceeds what a Poisson model would predict. The zero inflated Poisson (ZIP) model accounts for the excess zeros by assuming the data are drawn from a mixture of a Poisson population and a population producing zeros with probability one.

In a ZIP model, there are two generalized linear models (GLM): a Poisson regression model and a logistic regression model. The ZIP model accounts for the excess zeros by assuming the data are drawn from a mixture of a Poisson population and a degenerate distribution at zero. That is, the random variable equals zero with probability one; there is no variation. In ZIP regression, the responses $\boldsymbol{Y}$ are independent and can be written as:

$$\Pr(Y = y) = Pf_1(y) + (1 - P)f_2(y;\lambda), \quad (2.1)$$

where $f_1(y) = \begin{cases} 1 \; if \; y = 0, \\ 0 \; if \; y \neq 0, \end{cases}$

$$f_2(y;\lambda) = e^{-\lambda}\lambda^y / y!,$$

$$\log(\lambda) = \beta \boldsymbol{X},$$

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \alpha \boldsymbol{Z}.$$

The function $f_2(y;\lambda)$ is a standard Poisson distribution for modeling counts with mean λ and support {0, 1, 2,…}, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are covariate matrices, and P is a mixture proportion with $0 \leq P \leq 1$. Note that the covariate matrix $\boldsymbol{X}$, that affects the Poisson mean, may or may not be the same as the covariate matrix $\boldsymbol{Z}$, that affects the probability of perfect state membership.

2.3. Mixed Effect Zero Inflated Poisson Model

Fixed Effect vs. Random Effect

In a study, variable effects are either fixed or random depending on how the levels of the variables that appear in the study are determined. If one imagines that an experiment is repeated and the levels of an effect are identical, that effect would be considered as a fixed effect. If the levels of an effect change in an uncontrollable way when the experiment is repeated, the effect would be considered as a random effect. The levels of a random effect are regarded as a sample from some distribution.

An example of fixed effect occurs in a clinical trial analyzing the effectiveness of three drugs. Assume that the three drugs are the only candidates for the clinical trial, that the subjects are a representative sample from a population of interest and that the conclusions of the clinical trial are restricted to just those three drugs. Then the effect of

the variable drug is a fixed effect, since the drug factor in the study represents all levels in which inferences are to be made.

Suppose instead that the clinical trial was performed in ten clinics, that the subjects in a clinic are drawn from the population served by that clinic and that the ten clinics are a sample from a larger population of clinics. A replication of the trial would use a different random sample of clinics. The factor of interest is still drug, and clinic is a nuisance factor. The conclusions of the clinical trials are not restricted only to the patients served by the ten clinics but rather to the population of patients served by all clinics. The ten clinics are a random sample of the clinic population. The variable clinic can be treated as a random effect variable and inferences are valid for the population of patients served by all clinics.

In real world problems, there may be random subject-to-subject variation not explained by covariates. So a random subject effect is introduced into models. In a case of clinical meta-analysis data, the individual subject parameter is not of research interest and there may be many subjects; to eliminate nuisance subject parameters, a random subject effect is introduced into models. In a clustered or repeated measures design, observations nested within subjects are correlated with each other while observations on different subjects are independent of each other. So again a random subject effect is introduced into such models. The resulting model is a mixed effect model; that is, the model includes both fixed effects for the covariates and the random effect for subjects typically with the assumption that the conditional responses are normally distributed.

2.3.1 Generalized Linear Mixed Model

A random effect vector $\boldsymbol{u}$ is incorporated into the GLM as follows: Assume $Y_{ij}|\boldsymbol{u} \sim f_{y_{ij}|\boldsymbol{u}}(y_{ij}\,|\boldsymbol{u}_i)$ are conditionally independent, given $\boldsymbol{u}$,

$$E[Y_{ij}|\boldsymbol{u}] = \mu_{ij},$$

$$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{X}'_{ij}\beta + \boldsymbol{W}'_{ij}u_i,$$

$$\boldsymbol{u} \sim f_U(\boldsymbol{u}),$$

where $\boldsymbol{W}_{ij}$ represents a vector of covariates. The so-called dispersion parameter τ is often known to be 1, as in binomial or Poisson models. The resulting model is a generalized linear mixed model (GLMM).

Note that the random effect in a GLMM can be a random intercept effect, a random slope effect, or both. A growth model, which is often used in agricultural research studies, is an example of a GLMM having random intercept and random slope effects. An example of a GLMM having random intercept effect only is the Rasch model in item response theory (IRT) (Rasch, 1960). In the Rasch model, item responses $(j = 1,2,\dots,J)$ are nested within subject$(i = 1,2,\dots,n)$. The probability of a correct response to the dichotomous item j $(Y_{ij} = 1)$ conditional on the random effect or "ability" of subject i (θ_i) in terms of the logistic cdf is $P(Y_{ij} = 1|\theta_i) = \Omega(\theta_i - b_j)$, where $\Omega(\cdot)$ is the cumulative logistic function, and b_j is the item parameter. In the Rasch model, subjects are modeled as a random effect and items are included as covariates. Note that in IRT literature the subject ability usually is denoted as θ instead of u (Hedeker & Gibbons, 2006). A special case of GLMM is no random effect at all. That is, one drops the $\boldsymbol{u}$ term, so that the model reduces to a GLM.

The following is an example of a random intercept effect GLMM in a repeated measurement data setting. Assume there are $i = 1, \dots, n$ subjects and $j = 1, \dots, J$ repeated observations nested within each subject. A random-intercept model augments the linear predictor with a single random effect for subject *i*,

$$\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i$$

where u_i is the random effect (one for each subject). These random effects represent the influence of subject i on his/her repeated observations that is not captured by the covariates. These are treated as random effects because the sampled subjects are thought to represent a population, and their corresponding effect terms are assumed to be distributed as $N(0, \sigma_u^2)$. The parameter σ_u^2 indicates the variance in the population distribution, and therefore the degree of heterogeneity of subjects. The conditional mean of Y_{ij}, denoted as μ_{ij}, is specified as $E\left[Y_{ij} \middle| u_i, \boldsymbol{x}_{ij}\right]$. Note that the assumption of normal distributions for the random effects is not necessary; one can have some distribution other than normal. However some prefer to use the normal distribution for ease of interpretation (Weng, 2008).

2.3.2 Random Intercept Zero Inflated Poisson Model

In the previous section we presented a basic fixed-effects ZIP model. Now we expand the basic ZIP model by incorporating subject random effects. For a random effect ZIP model, the two generalized linear models are extended to two generalized linear mixed models (GLMM): mixed effect Poisson regression model and mixed effect logistic regression model:

$$logit(p_{ij}) = \log\left(\frac{p_{\text{ij}}}{1- p_{ij}}\right) = \boldsymbol{Z}_{ij}\alpha + u_i \qquad (2.2)$$

$$\log(\lambda_{ij}) = \boldsymbol{X}_{ij}\beta + v_i \quad (2.3)$$

where α and β are vectors of fixed effect coefficients and u_i and v_i denote random effects with

$$u_i \sim N(0, \sigma_u^2),$$

$$v_i \sim N(0, \sigma_v^2).$$

The normal random intercepts (u_i and v_i) capture the heterogeneity of subjects. The variances σ_u^2 and σ_v^2 measure the differences among subjects. Note that here the subject effects do not change over time. Neelon, O'Malley, & Mormand (2010) allowed u_i and v_i to be correlated.

2.4 AR(1) Mixed Effects Model

There are designs in which data are collected in such a way that data points are correlated, such as cluster sampling data, matched pair data, repeated measures data, time series data, and longitudinal data. For repeated measures data, observations taken on the same subject tend to be more similar than observations taken on different subjects. The assumptions of the repeated measures models used in psychological research are that the measurements have equal variances at all time points and that the correlation between measurements on the same subject is the same regardless of the time lag between measurements. For time series data, observations taken close in time tend to be more similar than measurements taken far apart in time, so correlations decrease as the time lag increases. Longitudinal data typically exhibits the features of time series data.

In mixed models, the variance-covariance matrix of the observations involves the covariance structure of the random effects and the covariance structure of the random

errors. We need to select an appropriate covariance structure for the random errors. If one chooses a structure that is too simple, one risks increasing the Type I error rate; if too complex, one sacrifices power and efficiency.

The first-order autoregressive covariance structure takes into account a common trend in longitudinal data. The covariance pattern model with first order autoregressive [AR(1)] structure is commonly used in time series analysis to describe the correlation structure. The AR(1) process only depends on two parameters. The covariance for time points j and j' is

$$\sigma_{jj'} = = \sigma^2 \rho^{|j-j'|},$$

where ρ is the AR(1) parameter and σ^2 is the error variance. The basic idea is that random effects are autocorrelated over time and the correlations decrease as observations are further apart in time. The assumption in the AR(1) model is that the longitudinal data are equally spaced. This means that the distance between time 1 and 2 is the same as time 2 and 3, time 3 and 4, and so on.

In the previous section, a GLMM model was described to handle repeated measures data. For longitudinal data, AR(1) random effects are incorporated into GLMM to yield models that allow for subject heterogeneity and some form of time dependence of the errors (Sun, Speckman, & Tsutakawa , 2000). Other correlation structures for time series data might have been chosen (e.g., AR(p) with $p > 1$ or moving average, MA(q), $q > 0$), but AR(1) is simple, is easy to simulate and captures the time series correlation structure. See Chatfield (2009) for more details on time series models.

2.5 Bayesian Data Analyses

2.5.1 Maximum Likelihood Estimation

For a parametric statistical model $\mathcal{H}$, a common estimator is the maximum likelihood estimator. Let $X_1, \cdots, X_n$ be i.i.d. with PDF $f(x;\theta)$ where $\theta \in \Theta$, and Θ is the parameter space. The likelihood function is

$$L_n(\theta) = f(X_1, \cdots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta).$$

The log likelihood function is $l_n(\theta) = \log[L_n(\theta)]$. The maximum likelihood estimator (MLE) $\hat{\theta}_{ML}$ is a point in the parameter space such that $l_n(\hat{\theta}_{ML}) = \sup_{\theta\in\Theta} l_n(\theta)$.

The MLE has several nice properties (Bickel & Doksum, 2007, pp. 331-332). They are stated here for a scalar parameter, but similar results are true in the multiparameter case.

a) Consistency – When $\mathcal{H}$ is identifiable, under general conditions, the MLE $\hat{\theta}_{ML} \xrightarrow{p} \theta^*$, where θ^* is the true value of the parameter θ.

b) Asymptotic Normality – Given *n i.i.d.* observations, the Fisher information (Bickel & Doksum, 2007, p. 180) is defined as $I_n(\theta) = nE_\theta\left[\left(\frac{\partial}{\partial\theta} log f(X;\theta)\right)^2\right]$, and under regularity conditions this is also equal to $-nE_\theta\left[\frac{\partial^2}{\partial\theta^2} log f(X;\theta)\right]$.

Let $se_n = \sqrt{Var_\theta(\hat{\theta}_{ML})}$. Under appropriate regularity conditions, $se_n = \sqrt{1/I_n(\theta)}$, and $\frac{\hat{\theta}_n-\theta}{se_n} \xrightarrow{d} N(0,1)$, as $n \to \infty$. Furthermore, let $\widehat{se}_n = \sqrt{1/I_n(\hat{\theta}_{ML})}$. Then $\frac{\hat{\theta}_n-\theta}{\widehat{se}_n} \xrightarrow{d} N(0,1)$.

c) Cramer-Rao Lower Bound – Let $\hat{\theta}_n$ be any unbiased estimator of θ. Its variance is bounded below by the inverse Fisher information as $n \to \infty$: $Var_\theta(\hat{\theta}_n) \geq \frac{1}{I_n(\theta)}$.

The limiting variance of the maximum likelihood estimator is equal to this lower bound under suitable regularity conditions (Bickel & Doksum, 2007, p. 331).

The Fisher information plays a central role in maximum likelihood estimation. Let's take a closer look at the Information Matrix. The asymptotic variance-covariance matrix of the ML estimator of a vector valued parameter $\boldsymbol{\theta}$, is calculated as the inverse of the Fisher information matrix, $[\mathbf{I}(\boldsymbol{\theta})]^{-1}$ (Bickel & Doksum, 2007, pp. 331-332, pp. 386-387; Shao, 2003, pp. 290-292) where

$$\mathbf{I}_N(\boldsymbol{\theta}) = E[\{(\frac{\partial}{\partial \boldsymbol{\theta}}) \log L(\boldsymbol{\theta})\}\{(\frac{\partial}{\partial \boldsymbol{\theta}}) \log L(\boldsymbol{\theta})\}']$$

$$= -E[\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) \log L(\boldsymbol{\theta})]. \qquad (2.4)$$

Thus, the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\theta}}_{ML}$ is:

$$\text{a-var-cov}(\widehat{\boldsymbol{\theta}}_{ML}) = [\mathbf{I}_N(\boldsymbol{\theta})]^{-1}$$

$$= \left(-E\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]\right)^{-1} \qquad (2.5)$$

From above, we see that the inverse of the information matrix is exactly the same as the Cramer-Rao lower bound.

Asymptotic Efficiency

Suppose that $\hat{\theta}$ is a consistent, asymptotically unbiased estimator of θ. Under regularity conditions the asymptotic variance of a suitably normalized version of $\hat{\theta}$ is bounded below by the inverse Fisher information. A multiparameter version of this bound is also available. (This is an asymptotic version of the Cramer-Rao bound.) An

estimator whose asymptotic variance achieves this lower bound is said to be asymptotically efficient.

Eigenvalues and Eigenvectors of the Fisher Information Matrix

Suppose that one has a sample of independent and identical observations $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)$ whose distribution depends on a multidimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\mathcal{R}^p}$.

Let $\boldsymbol{I}_1$= Fisher information about $\boldsymbol{\theta}$ in one observation, and let

$\boldsymbol{I}_N(\theta) = \boldsymbol{I}_N = N\boldsymbol{I}_1 = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} logL(\boldsymbol{\theta})\right]$ = Fisher information in *N* observations.

Under regularity conditions, asymptotically

$$\sqrt{N}\left[\widehat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}\right] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}_1^{-1}),$$

$$\therefore \widehat{\boldsymbol{\theta}}_{ML} \cong N(\theta, N^{-1}\boldsymbol{I}_1^{-1}).$$

where $N^{-1}\boldsymbol{I}_1^{-1}$ is the Cramer-Rao lower bound for an unbiased estimator.

Furthermore, $\boldsymbol{I}_1$ can be written

$$\boldsymbol{I}_1 = \boldsymbol{QDQ}'$$

where $\boldsymbol{D} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are eigenvalues, the matrix $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_p]$, and its columns are eigenvectors. Furthermore, $\boldsymbol{QQ}' = \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$, so that $\boldsymbol{q}_j'\boldsymbol{q}_k = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$. Since $var(\boldsymbol{q}_k'\widehat{\boldsymbol{\theta}}) \cong \frac{1}{N}\left(\frac{1}{\lambda_k}\right)$, for all *k* =1,… ,*p*, the "hardest" estimation problem in terms of variance is to estimate $\boldsymbol{q}_p'\widehat{\boldsymbol{\theta}}$ (Shao, 2003, pp. 290-292).

2.5.2 Bayesian Inference

The frequentist statistical methods have the following features:

a) Probability refers to limiting relative frequency.

b) Data are random.

c) Estimators are functions of data, so they are random.

d) Parameters are fixed, unknown constants. They are not subject to probabilistic interpretation.

e) Procedures are subject to probabilistic statements. For example, a 95% confidence interval traps the true parameter value in 95% of all possible samples.

An alternative is the Bayesian approach:

a) Probability refers to degree of belief. Therefore parameters are random and described probabilistically.

b) Inference about a parameter θ is carried out by computing its probability distribution given the observed data. One starts with a prior distribution $p(\theta)$ and also chooses a likelihood function $p(x|\theta)$ – note this is a function of θ, not x. After observing data x, one applies Bayes' Theorem to obtain the posterior distribution $p(\theta|x)$:

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\int p(\theta')p(x|\theta')d\theta'} \propto p(\theta)p(x|\theta),$$

where $C \equiv \int p(\theta')p(x|\theta')d\theta'$ is known as the normalizing constant.

c) For large *n* and under regularity conditions, posterior distributions are approximately normal. Specifically $p(\theta|\boldsymbol{x}_n) \approx N\left[\tilde{\theta}_n, \left(-\frac{\partial^2}{\partial\theta^2}\log \boldsymbol{L}_n(\tilde{\theta}_n)\right)^{-1}\right]$ (Bickel & Doksum, 2007, p. 339) where $\tilde{\theta}_n$ is the Bayesian estimate. That is, the posterior distribution converges to this limiting normal distribution. This is a consequence of the Bernstein-von Mises theorem stated in the next section.

Asymptotic Equivalence of Bayes and M.L. Estimators

A precise statement of the large sample behavior of the posterior distribution is given in the Bernstein-von Mises theorem.

Bernstein – von Mises (BvM) Theorem: Under suitable regularity conditions, the posterior distribution is asymptotically normal with asymptotic mean equal to the maximum likelihood estimator (MLE) and asymptotic variance equal to the inverse of the total Fisher information matrix, as for the MLE. See Bickel & Doksum (2007, p. 339).

This theorem has the following important consequences. Let $\hat{\theta}_n$ be the MLE based on *n* observations, let $\tilde{\theta}_n$ be the posterior mean based on *n* observations, and let θ^* be the true value of the parameter θ. Under general conditions that apply in exponential families:

a) $\left|\hat{\theta}_n - \theta^*\right|$ decreases to 0 in probability at rate $\frac{1}{\sqrt{n}}$, and $\left|\tilde{\theta}_n - \theta^*\right|$ decreases to 0 in probability at rate $\frac{1}{\sqrt{n}}$. Thus, the two estimators converge to the truth at the same rate.

b) $\left|\hat{\theta}_n - \tilde{\theta}_n\right|$ decreases to 0 in probability at rate $\frac{1}{n}$. Therefore, for large *n*, the difference between the posterior mean and the MLE is of smaller order of magnitude than the error of estimation due to sampling as *n* goes to infinity.

c) The asymptotic variance for the MLE is the inverse of the expected information: $I(\theta)^{-1} = \left\{-E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log L_n(\boldsymbol{x}_n|\theta)\right]\right\}^{-1}$. This is the same as the variance of the asymptotic posterior distribution. The variance of the approximate posterior distribution is the observed information:

$\tilde{I}_n(\tilde{\theta}) = -\frac{\partial^2}{\partial\theta^2}\ln L_n(\boldsymbol{x}_n|\theta)\,|_{\theta=\tilde{\theta}}$.

2.5.2.1 Prior Distribution

A standard maximum likelihood inferential model is often formally identical to a Bayesian model in which the prior probability distribution is an appropriate uniform distribution function, although often the prior distribution is improper: its integral is infinite rather than equal to one.

Under suitable regularity conditions, Bayes estimates and maximum likelihood estimates are asymptotically identical for any proper prior distribution (Chao, 1970) with support in a neighborhood of the true parameter values. Therefore with sufficiently large sample size the choice of the prior has a negligible effect on the inference.

A common frequentist criticism of the Bayesian approach is that "subjective" priors have great impact on the posterior distribution for problems with small sample sizes. There is a developing literature on robust Bayesian inference that seeks to mitigate this problem by developing estimators that are relatively insensitive to a wide range of prior distributions (Berger, 1984). The strongest argument for inclusion of priors is that there often exists scientific evidence at hand before the statistical model is developed and it would be foolish to ignore such previous knowledge. Furthermore, a formal statement of the prior distribution is an overt, nonambiguous assertion within the model specification that the reader can accept or dismiss (Gelman et al. 2003, p. 14).

2.5.2.2 MCMC Estimation

The integral in the denominator of the formula for the posterior distribution

$$p(\theta|Y) = \frac{p(\theta)L(\theta|Y)}{\int_{\Theta} p(\theta)L(\theta|Y)d\theta}$$

is often impossible to evaluate analytically even if the integrand can be expressed in a simple form. In such a case, one can not calculate the posterior distribution explicitly and therefore one can not calculate expressions such as the posterior mean in closed form. Moreover, in many cases, one can not draw samples directly from the posterior to approximate such quantities by simulation. In such problems, Markov chain Monte Carlo (MCMC) enables one to draw samples which are approximately distributed according to $p(\theta|Y)$ and thereby one can approximate quantities related to $p(\theta|Y)$ (Gelman et al., 2003, pp. 285-287).

Standard Monte Carlo simulation would generate independent random variables $\theta_1, \dots, \theta_N$ distributed exactly according to $p(\theta|Y)$. These values would be used to approximate the posterior expectation of some function $h(\theta)$, $E[h(\theta)|Y]$, as an average:

$$E[h(\theta)|Y] = \int h(\theta)p(\theta|Y)d\theta \cong \frac{1}{N}\sum_{i=1}^{N} h(\theta_i).$$

$E[h(\theta)|Y]$ might be a posterior mean or higher order moment of the parameter θ.

MCMC methods approximate integral quantities by drawing a very long sequence of values $\theta_1, \dots, \theta_N$ from a Markov chain whose limiting probability distribution exists and is equal to $p(\theta|Y)$. The primary distinction between standard Monte Carlo simulation methods and MCMC methods is the dependence structure of consecutive simulated values. Standard Monte Carlo methods produce a set of independent simulated values according to some desired probability distribution. MCMC methods produce chains in which each of the simulated values is dependent on the preceding value. The basic principle is that if a properly constructed chain has run long enough, it will produce an almost stationary time sequence of approximately identically distributed values whose

marginal distributions will be extremely close to the posterior distribution of interest. One can summarize this distribution by letting the chain wander around to generate a large approximately representative sample from the limiting distribution and then producing summary statistics from the simulated values (Gelman et al., 2003, pp. 285-287).

The Gibbs Sampler

The Gibbs sampler is a widely used MCMC technique for producing useful chain values. It requires specific knowledge about the nature of the conditional relationships among the variables of interest. The basic idea is to generate a new realization of each variable from its so-called full conditional, which is defined as its conditional distribution, given the current values of all the other variables and the observed data. Then by cycling repeatedly through these conditional random draws, one parameter at a time, we can eventually approach the joint distribution of interest (Gelman et al., 2003, pp. 287-289).

The Gibbs sampler generates a sequence of *k* dimensional random vectors $\theta_t = [\theta_{t1} \dots, \theta_{tk}], t = 1,2, \dots$ using a transition mechanism based on sampling from the conditional distribution$\pi(\theta_{tj}|\theta_{t1} \dots, \theta_{t,j-1}, \theta_{t-1,j+1} \dots, \theta_{t-1,k,} X)$. This is a Markovian updating scheme based on sampling from the conditional probability distributions. The limiting distribution of interest is the posterior distribution, $p(\theta|Y)$. Regularity conditions are needed to assure the convergence of the Gibbs sampler to the desired limiting distribution (Shao, 2003, pp. 247-248).

There should be an analytically definable full conditional distribution for each component of the θ vector and these probability distributions need to be written explicitly so that it is possible to draw samples from the described distribution. However, Gibbs samplers can work also when the conditionals themselves are obtained by some simulation technique like a rejection method. This requirement facilitates the iterative nature of the Gibbs sampling algorithm, which cycles through these full conditionals drawing parameter values based on the most recent version of all the previous parameters in the list. The order does not matter, but it is essential that the most recent draws from the other samples be used. The following is the procedure:

a) Choose a starting value $\boldsymbol{\theta}_0 = [\theta_{01,} \ldots, \theta_{0k}]$.

b) At time t, starting at $j = 1$, complete the single cycle by drawing values from the k conditional distributions given by:

$$\theta_{t1} \sim \pi(\theta_1 | \theta_{t-1,2}, \ldots, \theta_{t-1,k}, X)$$

$$\theta_{t2} \sim \pi(\theta_2 | \theta_{t1}, \theta_{t-1,3}, \ldots, \theta_{t-1,k}, X)$$

$$\vdots$$

$$\theta_{t,k-1} \sim \pi(\theta_{k-1} | \theta_{t1}, \ldots, \theta_{t,k-2}, \theta_{t-1,k}, X)$$

$$\theta_{tk} \sim \pi\left(\theta_k \middle| \theta_{t1}, \ldots, \theta_{t,k-1}, X\right).$$

c) Increment t and repeat until a convergence criterion is satisfied (noting that convergence is to stationarity rather than to a point, as it would be for iteratively calculated randomization-based estimators).

The Metropolis Algorithm

Another Markov chain Monte Carlo tool is the Metropolis algorithm (Gelman et al., 2003, pp. 289-292; Shao, 2003, pp. 249-250 and references therein). It works with

the joint distribution rather than the conditional distributions for the parameters in the model. The idea behind the Metropolis algorithm is that, when we cannot easily generate values from the joint (posterior) distribution of interest, we can often find a "similar" distribution that is easy to sample from. We need to make sure that this alternate distribution is defined over the same support as the target distribution and that it does not favor areas of low density of the target. Once a candidate point in multivariate space has been produced by this candidate generating distribution we will accept or reject it based on the target distribution. The resulting Markov chain wanders around and favors higher density regions. It will also explore other lower density regions, but with lower probability as we would want. Note that unlike the Gibbs sampler, the Metropolis algorithm does not necessarily have to move to a new position at each iteration; a candidate sample from the proposal distribution is accepted as the next value in the chain as a function of its relative density in the proposal distribution and the target distribution.

It was noted in the preceding discussion of the Gibbs sampler that when it is not possible to draw directly from a parameter's full conditional, a simulation method can be employed. The family of related MCMC programs that are collectively called BUGS (Ntzoufras, 2009; Spiegelhalter, et al., 2003) uses Metropolis sampling to approximate draws for full conditionals when it is not possible to draw directly from full-conditionals. This is called Metropolis-within-Gibbs sampling.

It is important, for the Gibbs samplers and the Metropolis algorithm, in real applications to run the Markov chain for some initial period to let it settle into the distribution of interest before recording values. This is called "burn in."

Assessing the Convergence of Markov Chains

The empirical results from a given MCMC analysis are not reliable until the chain has come close to its stationary distribution and has sufficiently mixed throughout. Some convergence problems may come directly from badly chosen priors, which present difficult algorithmic challenges especially in fully-developed specifications with large numbers of parameters, near-collinearity, or parameters for which little information is available in the data. One common alternative is to use highly diffuse but proper priors; however, this alternative can sometimes lead to slow convergence of the Markov chain.

There are basically three approaches to determining convergence for Markov chains: (1) assessing the theoretical and mathematical properties of particular chains, (2) diagnosing summary statistics from in-progress models, and (3) perfect sampling (Gill, 2008). In this chapter, only the second approach is discussed. It is essential to remember that the convergence diagnostics described below are actually indicators of "non-convergence." That is, failing to find evidence of non-convergence with these procedures does not imply convergence. Also the information each diagnostic provides has limitations; thus, multiple diagnostics should be used.

A simple way to monitor convergence is monitor the Monte Carlo error (MC error), which measures the variability of each estimate due to the simulation. The MC error must be low in order to calculate the parameter of interest with increased precision (Ntzoufras, 2009). Monitoring autocorrelations is useful and low values indicate fast convergence (although parameters that are not well determined from the data can have high autocorrelations even when the chains have converged to stationarity). Another way is to monitor the trace plots, the plots of the generated values versus iteration number. If

all values are within a zone without strong periodicities and trends, then convergence is more likely. Another way is to run multiple chains with different starting points (Brooks & Gelman, 1998). This strategy supports both visual checks and statistical checks. When we observe that the different chains mix, the convergence is very likely. Finally, several statistical tests have been developed and used as convergence diagnostics. CODA (Best et al., 1996) and BOA (Smith, 2005) software have been developed to implement diagnostics to the output of BUGS and WinBUGS software. We may improve the mixing of the chain, reducing the time to reach convergence, by implementing transformation of the parameters of the parameters of interest (Gilk & Roberts, 1996).

2.5.2.3 Program Language of Bayesian Inference Using Gibbs Sampling (BUGS) Modeling for Nonstandard Distributions

The zero inflated models introduce an extra probability parameter to capture an excess of zero values. The zero inflated version of A random variable $Y \sim ZID(\pi_0, \theta)$, (where ZID denotes zero inflated distribution), has a probability function of the form

$$f_{ZID}(y) = \pi_0 I(y = 0) + (1 - \pi_0) f_D(y; \theta)$$

where $f_D(y; \theta)$ is the probability function of distribution D with parameters θ. The distribution D could be Poisson, gamma, binomial, negative binomial, generalized Poisson, bivariate Poisson, or multivariate Poisson. From the equation above, we can see that ZID is a nonstandard distribution which is not listed among BUGS' prespecified distributions. There is an approach to specify nonstandard prior and likelihood in BUGS: the zeros-ones trick (Spiegelhalter et al., 2003; Ntzoufras, 2009). Thus the programs for

our three models all involve the zero-ones trick. This trick allows arbitrary sampling distributions to be used. However it has been shown that this method can be very inefficient and give a very high MC error. (Spiegelhalter et al., 2003).

The idea of the zeros-ones trick is to use the Poisson distribution to indirectly specify an arbitrary model likelihood (Ntzoufras, 2009). Assume a model with log-likelihood $\sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \log f(y_i|\theta)$. Then the log-likelihood can be re-written as

$$f(y|\theta) = \prod_{i=1}^{n} e^{l_i} = \prod_{i=1}^{n} \frac{e^{-(-l_i)}(-l_i)^0}{0!} = \prod_{i=1}^{n} f_p(0; -l_i)$$

where $f_p(k;\mu)$ denotes a Poisson probability function with mean μ. Hence, $f_p(0; -l_i) = P[\Xi_i = 0]$, where the new random variables $\Xi_i, i = 1, \dots, n$ follow the Poisson distribution with means equal to $-l_i$, and all observed values are set equal to zero. To ensure the positivity of the mean of each new random variable, a positive constant term *C* is added to the mean. This is equivalent to multiplying each likelihood factor by e^{-C}. This action does not affect the likelihood since it is equivalent to multiplying the resulting (unnormalized) posterior distribution by a constant term, e^{-nC}. The likelihood is equal to

$$f(y|\theta) = \frac{e^{-(-l_i + C)}(-l_i + C)^0}{0!} = \prod_{i=1}^{n} f_p(0; -l_i + C)$$

The constant *C* must be selected in such a way that $-l_i + C > 0$ for all *i = 1, 2, …, n.* For example, the normal model can be fitted in WinBUGS using the following code (Ntzoufras, 2009, p. 276):

```
C <- 10000
for (i in 1:n) {
    zeros[i] <- 0
```

zeros[i] ~ dpois(zeros.mean[i])

zeros.mean[i] <- -l[i] + C

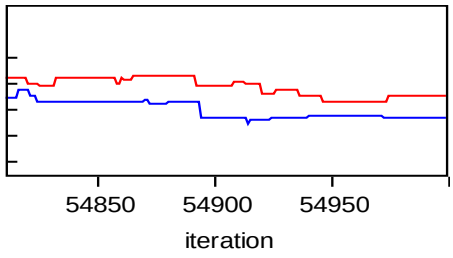
l[i] <- -0.5*log(2*3.14) – 0.5log(s2) – 0.5*pow(y[i] – mu[i], 2)/(s2) }

Convergence Checking By Using Bugs Sample Monitor Tool

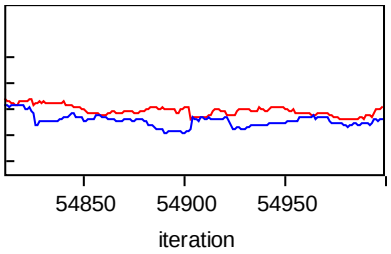
To check convergence when running several MCMC chains, several built-in diagnostic / monitor tools are used while running BUGS. As noted in the previous section on diagnosing convergence, these are diagnostic tools rather than certain evidence of convergence. They are only capable of identifying nonconvergence.

a) **Trace plot**: this is a plot of generated values against each iteration number. Shown below are examples with two independent chains.

   a. This is an example where possible non-convergence evidence was found: The chains have not mixed well. (The chains may be in fact moving around the same stationary distribution – i.e., may have converged – but are mixing very slowly. Looking at a longer interval of the chain would be the next step to see if mixing occurs, but at a slower rate than could be discerned in this interval of about 200 cycles.)

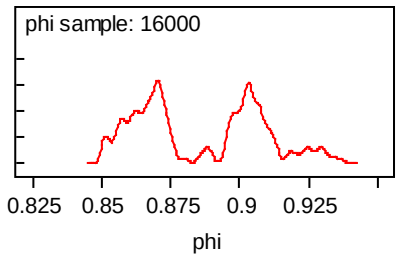


   b. This is an example where “evidence of non-convergence” was not found: the chains have mixed well.

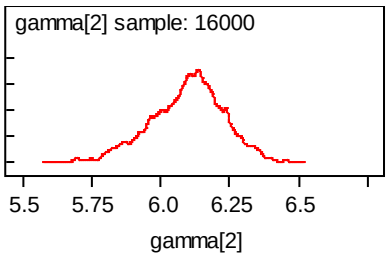


b) **Posterior density**: According to the Bernstein – von Mises theorem, the posterior distribution for continuous parameters is asymptotically normal under broadly satisfied conditions (although problems can be constructed in which the true posterior for a parameter is in fact bimodal). BUGS produces an approximate visual kernel estimate of the posterior density or probability function.

a. This is an example where possible "non-convergence evidence" was found.

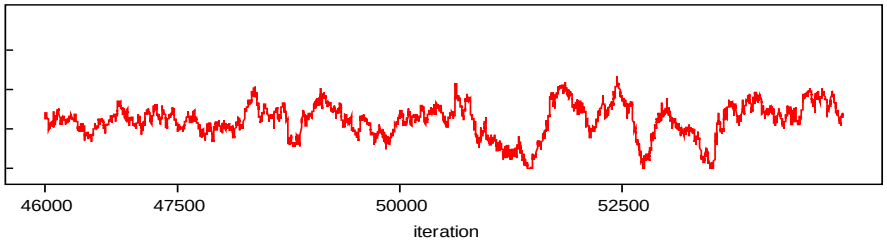


b. This is an example where evidence of "non-convergence evidence" was not found.

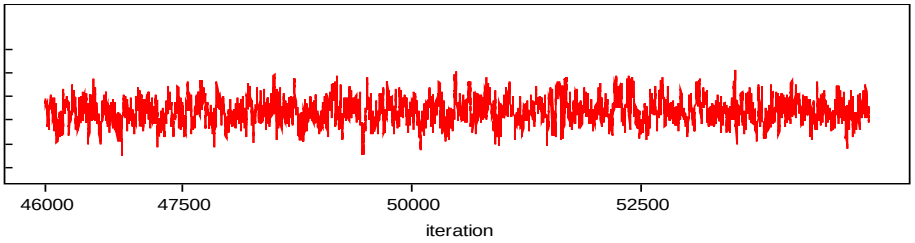


c) **History**: This is a full trace plot of all stored values. Values within a zone without strong trends or periodicities are suggestive of convergence of the chain.

a. This is an Example where non-convergence evidence was found. (Again, examining a longer run is important, since slow mixing can also produce the graph shown immediately below. Viewed from a further distance in this sense the chain shown below might look essentially like the graph below it.)

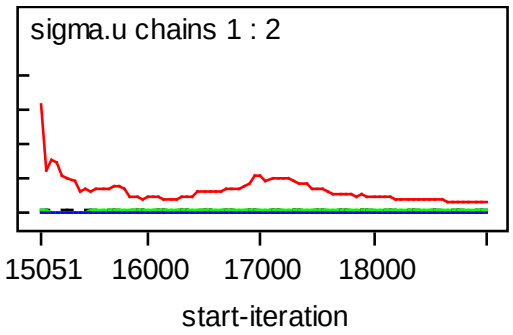


b. This is an example where "non-convergence evidence" was not found

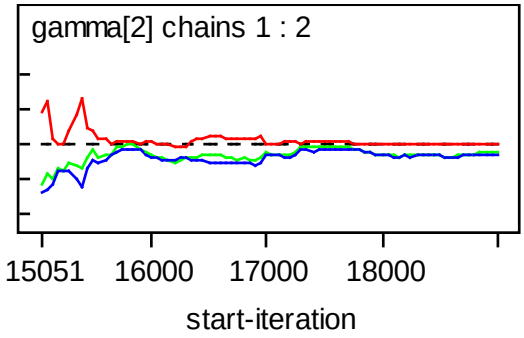


d) **The plot of Gelman-Rubin diagnostic**: This is an ANOVA type diagnostic (Gelman & Rubin, 1992); multiple chains converging to the same stationary distribution should each show similar within-chain variation. This common within-chain variance should be equal to the pooled variance. Values of R close to one (shown as the red line in the plots below) suggest convergence.

a. This is an example where evidence of non-convergence was found.



b. This is an example where evidence of non-convergence was not found.



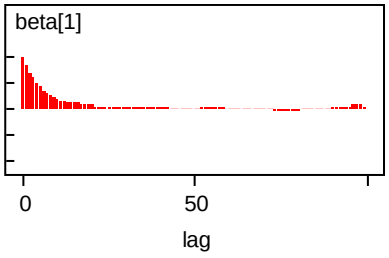
Monte Carlo Error and the Posterior Standard Deviation

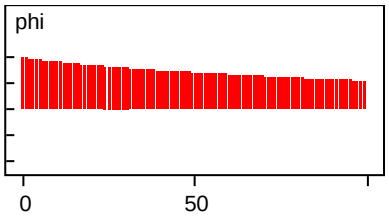
The BUGS output produces two error measures: the Monte Carlo (MC) error and the posterior standard deviation. The MC error measures the variability of each variable's average across the Monte Carlo chain. This variability is due to running the chain for a finite number of cycles and can be made as small as desired by increasing the number of draws. On the other hand, the posterior standard deviation, the analog of the standard error in conventional statistical inference, represents genuine uncertainty due to having a finite set of data and cannot be reduced other than by obtaining additional real data. Ideally, we would like to run the program until there is both no evidence of non-convergence and negligible MC error, and then calculate the posterior means. In BUGS, the batch mean method (Roberts, 1996, p.50) is used to estimate the MC error. The

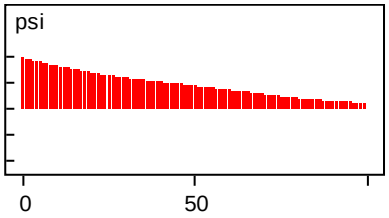
formula for the posterior standard deviation can be found in Carlin and Louis (2000, p. 172) and in Ntzoufras (2009, pp. 39-40).

Autocorrelation

This is a plot of all autocorrelations using lags from 1 to 50. If autocorrelations are low, then convergence tends to be obtained in a relatively lower number of iterations. Thus autocorrelation indicates how fast the convergence will be. If autocorrelations are large, the sampler will need many iterations to get close to the target distribution. Once it gets there, we will need to draw a very large number of additional iterations in order to get accurate estimation of features of the posterior. The Monte Carlo sampling error will be high. It might appear at first blush that there is a way to reduce the autocorrelation among repeated draws by using the "thin" tool. For example, when thin = 10, then BUGS will generate 10 iterations but will store only the last one in every sequence of 10 generated values. The autocorrelation among the saved values will in fact be lower. However, the chain lengths needed to reach stationarity or to achieve a satisfactory approximation of posteriors are the same. There is no real advantage in thinning except to reduce storage requirements and the cost of handling the simulations when very long runs are being carried out and subsequently analyzed. Examples of autocorrelation plots are presented below:




phi
0
50
lag


psi
0
50
lag

Chapter III

AR(1) Mixed Effects ZIP Model

3.1 Data Example

Before discussing the model, we would like to describe a data set which motivated the new model development. The National Longitudinal Survey of Youth (NLSY97, 1997) data set consists of a national sample of youths who were 12 to 17 years old as of December 31, 1996. Round 1 of the survey took place in 1997 and subsequent surveys were conducted annually. Cigarette use is our target to measure. Three questions were asked:

a) Have you ever smoked a cigarette?

b) During the past 30 days, on how many days did you smoke a cigarette?

c) When you smoked a cigarette during the past 30 days, how many cigarettes did you usually smoke each day?

In the NLSY97 dataset, there are approximately 9,000 youths observed at 13 time points. The observation "number of cigarettes smoked per day in the last 30 days" is denoted as Y_{ij}, for subject $i = 1, \dots, n$, at time point $j = 1, \dots, J$. Assume the responses are independent distributed across subjects. That is, Subject 1's response doesn't affect Subject 2's response. Since the individual subject parameters are not of research interest, the subject is treated as a random effect. That is, subjects with same covariates may have different responses because of the randomness. Within each individual, there is a random fluctuation at the 13 time points too. It is very likely that a subject's smoking status given the covariates will not remain a constant over the 13 time points. So we can say that subject is a multidimensional random effect. The fluctuations of each individual are

independent of the others. The distributions of the randomness will be described in mathematical expressions in the following section.

There are 13 time points in the data set, and a first-order autoregressive model is used to describe the autocorrelations across the time points (note that these autocorrelations are unrelated to the autocorrelations of draws in MCMC chains discussed earlier). There are four covariates: sex (two levels), race (four levels), age (six levels), and peers (what percentage of your friends are smokers?; five levels), the effects associated with which are also treated as fixed effects. Although there are many parameters in the model, the data set is big enough to estimate them. In the population, we expect that there is an increasing tendency to smoke, which is a function of time point or age; that is, a positive Poisson regression coefficient on the age variable is expected.

The survey question (a) "Have you ever smoked a cigarette?" can be a covariate too. Question (a) presumably provides very good information about the latent zero-class indicator that the ZIP model must posit.

The survey question (c) "When you smoked a cigarette during the past 30 days, how many cigarettes did you usually smoke each day?" yields the response variable, Y_{ij}. The Y_{ij}, the number of cigarettes a subject smokes, is a non-negative integer. We will model Y_{ij} with various forms of Poisson distributions. The average number of cigarettes a subject smokes at time j is a quantity of interest.

We looked the descriptive statistics of this data set before attempting any serious model building or analysis. Take the responses at Time 1 as an example: This data set shows excessive numbers of zeros so that a ZIP model is proposed instead of Poisson model. From Table 3.1.1 the total number of data points is 5436 (non-smokers) + 3517

(smokers) = 8953. We can think of the 146 smokers (smoking not observed) as sampling zeros; and think of the 5436 (non-smokers) as structural zeros. If the whole data set is treated as a sample from a Poisson distribution without covariates, the mean of sample is 0.9 [8106 (total number of cigarettes smoked) / 8953 = 0.9], we would expect $8953e^{-0.9}$ = 3640 zeros. However there are 5582 (146 + 5436 = 5582) zeros (note that we did not count zeros from missing values) in the data set. A ZIP model would better fit the data by having 5436 zeros as structural zeros or 5436 observations in the perfect class.

Table 3.1 Description of Smoking Data in the NLSY97 at Time 1

<table>
<tr><td colspan="3">Smokers: 3517</td><td>Non-smokers: 5436</td></tr>
<tr><td colspan="2">Observed smokers: 1618</td><td rowspan="2">Missing values: 1899</td><td rowspan="2">Non-smokers:5436</td></tr>
<tr><td>Zero cigarette smokers: 146</td><td>Non-zero cigarette smokers: 1472</td></tr>
</table>

3.2 AR(1) Mixed Effects ZIP Model

In this section the proposed model, AR(1) mixed effect ZIP model, is presented in three steps. In Subsection 3.2.1, a basic fixed-effects ZIP model is described with a rectangular data structure. In Subsection 3.2.2, the basic ZIP model is extended to incorporate two random subject effects: one in the logistic model and the other in the Poisson model. Also, in this mixed effect ZIP model correlation between the two random subject effects is incorporated. In Subsection 3.2.3, the AR(1) mixed effect ZIP model, in which two time-varying random subject effects are incorporated in the ZIP model, is described.

3.2.1 The Fixed Effect ZIP Model

In a rectangular data set, let the discrete random variable Y_{ij} be the *j*-th observed count on the *i*-th subject, $i = 1, \dots, n,\ j = 1, \dots, J$. Assume that Y_{ij} is distributed as a

mixture of two components: (1) responses are zero with probability one (perfect state); (2) responses follow a Poisson distribution (Poisson state). Assume that an unobserved random variable, z_{ij}, indicates the state membership of the subject, either the perfect state or the Poisson state. Note that if $y_{ij} > 0$, the subject was in the Poisson state, but if $y_{ij} = 0$, the subject may have been in either of the two states. This key feature makes the ZIP model different from the Poisson model. The z_{ij} are assumed to be from a Bernoulli distribution with parameter p_{ij}, such that $P(z_{ij} = 1) = p_{ij}$ and $P(z_{ij} = 0) = 1 - p_{ij}$. If $z_{ij} = 1$ then $Y_{ij} = 0$, coming from the perfect state and if $z_{ij} = 0$ then $Y_{ij} = y$, $y = 0, 1, 2, \ldots$, coming from a Poisson distribution. Therefore, Y_{ij} has the zero-inflated Poisson (ZIP) distribution:

$$Y_{ij} \sim \begin{cases} 0, & \text{with probability } p_{ij} \\ \text{Poisson}(\lambda_{ij}), & \text{with probability } (1 - p_{ij}), \end{cases} \tag{3.1}$$

where Poisson (λ_{ij}) is defined as $\mathrm{P}(Y_{ij} = y_{ij}) = \exp(-\lambda_{ij})\lambda_{ij}^{y_{ij}}/y_{ij}!$.

Covariates can enter into the model in two places: in the logistic regression model and in the Poisson regression model.

a) The logistic regression model is for predicting the status, perfect state or Poisson state. The probability p_{ij} is expressed as

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{\text{ij}}}{1 - p_{ij}}\right) = \boldsymbol{Z}_{ij}'\alpha = \alpha_0 + \alpha_1 Z_{ij}^1 + \cdots + \alpha_h Z_{ij}^h. \tag{3.2}$$

b) The log-linear regression for the Poisson mean is expressed as

$$\log(\lambda_{ij}) = \boldsymbol{X}_{ij}'\beta = \beta_0 + \beta_1 X_{ij}^1 + \cdots + \beta_k X_{ij}^k. \tag{3.3}$$

where $\boldsymbol{Z}_{ij}$ and X_{ij} are covariate vectors and α and β are vectors of regression coefficients for the logistic regression model and Poisson regression model, respectively. The

components of $\boldsymbol{Z}_{ij}$ and $\boldsymbol{X}_{ij}$ could be the same or different from each other. This model was essentially proposed by Lambert (1992).

3.2.2 The Mixed Effect ZIP Model

In the previous paragraphs a basic ZIP model was presented. In a repeated measures study, observations within a subject are correlated, and one might model this situation by adding a subject parameter to a regression model. Now we extend the ZIP model by incorporating a subject random effect (or random intercept effect).

For the mixed effect ZIP model, two random subject effects are introduced in the model so that

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{\text{ij}}}{1 - p_{ij}}\right) = \boldsymbol{Z}_{ij}'\alpha + u_i \qquad (3.4)$$

$$\log(\lambda_{ij}) = \boldsymbol{X}_{ij}'\beta + v_i \qquad (3.5)$$

where α and β are vectors of fixed effect coefficients and u_i and v_i denote random random effects with

$$u_i \sim N(0, \sigma_u^2),$$

$$v_i \sim N(0, \sigma_v^2).$$

The normal random intercepts (u_i and v_i) capture the heterogeneity of subjects. The variances σ_u^2 and σ_v^2 indicate the differences among subjects. Note that here the subject effects do not change over time. Time-varying random effects will be described later.

Furthermore, assume that u_i and v_i are correlated because the same subject is modeled in the two regressions. The random effects follow a bivariate normal distribution with zero means and covariance matrix

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} \tag{3.6}$$

where ρ denotes corr(u_i , v_i). This is a mixed effect ZIP model for repeated measures data without involving time. Finally, assume that $y_{ij}|\{u_i, v_i : i = 1, \dots, n\}$ are conditionally independent and conditionally follow a ZIP distribution. This model is presented by Neelon et al. (2010).

3.2.3 AR(1) Mixed Effects ZIP Model

It is assumed that a subject can move from the perfect state to the Poisson state and back again across time points *j*. For example, a non-smoker at Time 1 might begin smoking at Time 2; and a smoker at Time 1, Time 2, and Time 3 might decide to quit at Time 4. This is saying that any one cell in the person-by-occasions matrix can be a structural zero. This is different from some latent class models, where a person who begins as a zero-class member necessarily remains there at all time future points. To capture the variation across time points within a subject, first-order autoregressive (AR(1)) processes u_{ij}, v_{ij} are substituted for u_i, v_i in equations (3.4) and (3.5). This is our proposed model.

Time-varying random effects are incorporated into the model as follows. Let u_{ij} and v_{ij} denote the random effects, where *i* indexes subject and *j* indexes time. The other covariates remain as fixed effects. Then write

$$u_{ij} = \varphi u_{i,j-1} + e_{ij}, \tag{3.7}$$

where $j = 2, \dots, J$, and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, and e_{ij} are independent of $u_{ij} \sim N(0, \sigma_u^2)$. The parameter φ is a first order autocorrelation coefficient with $-1 < \varphi < 1$.

Furthermore,

$$Var(u_{ij}) = \varphi^2 Var(u_{i,j-1}) + \sigma_e^2. \tag{3.8}$$

To make $Var(u_{ij})$ independent of the time index *j*, assume

$$\sigma_u^2 = \sigma_u^2\varphi^2 + \sigma_e^2 \quad \rightarrow \quad \sigma_e^2 = \sigma_u^2(1 - \varphi^2)$$

Then the variance-covariance matrix of u_{ij} is

$$Var\ Cov\begin{bmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{bmatrix} = \sigma_u^2 \begin{bmatrix} 1 & \varphi & \varphi^2 & \dots & \varphi^{T-1} \\ & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \varphi^2 \\ & & & & \varphi \\ \varphi^{T-1} & & \dots & & 1 \end{bmatrix}, \tag{3.9}$$

$$Cov(u_{i,j-1}, u_{ij}) = E[u_{i,j-1}(\varphi u_{i,j-1} + e_{ij})] = E[\varphi u_{ij}^2] + 0 = \varphi\sigma_u^2, \tag{3.10}$$

$$Corr(u_{ij}, u_{i,j-1}) = \frac{\varphi\sigma_u^2}{\sigma_u^2} = \varphi = \text{first order autocorrelation.} \tag{3.11}$$

Similarly, an AR(1) model is also assumed for the v_{ij} with variance σ_v^2 and first order autocorrelation ψ. We also assume that $\rho = corr(u_{ij}, v_{ij}) \neq 0$. Let Φ be the autoregressive coefficient matrix such that $\Phi = \begin{bmatrix} \varphi & 0 \\ 0 & \psi \end{bmatrix}$ where φ is an autocorrelation coefficient for u_{ij} and ψ is an autocorrelation coefficient for v_{ij}, that is,

$v_{i,j} = \psi v_{i,j-1} + e'_{ij}$.

The autoregressive model is thus

$$\begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix} = \Phi \begin{bmatrix} u_{i,j-1} \\ v_{i,j-1} \end{bmatrix} + \begin{bmatrix} e_{ij} \\ e'_{ij} \end{bmatrix} \tag{3.12}$$

where the terms on the right hand side are independent. Then

$$V = var\ cov\begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_u^2 & \rho_{uv}\sigma_u\sigma_v \\ \rho_{uv}\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix}$$

$$= \Phi V \Phi + \begin{bmatrix} \sigma_e^2 & \rho_{ee'}\sigma_e\sigma_{e'} \\ \rho_{ee'}\sigma_e\sigma_{e'} & \sigma_{e'}^2 \end{bmatrix} \qquad (3.13)$$

where $\rho_{uv}\sigma_u\sigma_v = \varphi\psi\rho_{uv}\sigma_u\sigma_v + \rho_{ee'}\sigma_e\sigma_{e'}$

$$= \varphi\psi\rho_{uv}\sigma_u\sigma_v + \rho_{ee'}\sigma_u\sigma_v\sqrt{(1-\varphi^2)(1-\psi^2)} \qquad (3.14)$$

$$\rho_{ee'} = (1-\varphi\psi)\rho_{uv}/\sqrt{(1-\varphi^2)(1-\psi^2)} \qquad (3.15)$$

$$\sigma_e^2 = \sigma_u^2(1-\varphi^2) \qquad (3.16)$$

$$\sigma_{e'}^2 = \sigma_v^2(1-\psi^2) \qquad (3.17)$$

Interpretation of Mixed Effect Logistic Regression

As in section 3.2.2, the y_{ij} are conditionally independent ZIP variables, given $\{u_{ij}, v_{ij}: i = 1, \dots, n; j = 1, \dots, J\}$. Because the AR(1) model is conditioned on the random effects, the interpretation of results is not immediate. The following is an example (Agresti, 2002; pp. 496-500) for a single covariate X_1.

In fixed effect logistic regression:

$$P[\text{nonsmoker}|\mathrm{X}_1] = \frac{\exp[\beta_0 + \beta_1 X_1]}{1 + \exp[\beta_0 + \beta_1 X_1]}.$$

In mixed effect logistic regression the subject specific probability is:

$$P[\text{nonsmoker}|\mathrm{U}, \mathrm{X}_1] = \frac{\exp[\beta_0 + \beta_1 X_1 + U]}{1 + \exp[\beta_0 + \beta_1 X_1 + U]}.$$

The population average probability is:

$$P[\text{nonsmoker}|\mathrm{X}_1] = \int_{-\infty}^{\infty} \frac{1}{\sigma_U\sqrt{2\pi}} e^{-u^2/(2\sigma_U^2)} \frac{\exp[\beta_0 + \beta_1 X_1 + u]}{1 + \exp[\beta_0 + \beta_1 X_1 + u]} du$$

where $U \sim N(0, \sigma_U^2)$.

In this random effect model the link function is no longer the logit. If no random

Effects are present, the odds in favor of success increase by a factor $\exp(\beta_1)$ when X_1 in creases by one unit. This statement is true for all subjects, regardless of the values of X_1, and it is true when averaging over the population. In the presence of random effects, the same statement is true of the conditional odds of success, given the *u* and X_1. However, the average change in odds is no longer $\exp(\beta_1)$, but instead it is a nonconstant function of X_1. The link function no longer has the logistic shape and the effect of X_1 is reduced (Agresti 2002, p.500). Note also that the integral above can not be expressed analytically. Analysis of the model will therefore be more computational.

Chapter IV

Research Questions and Pilot Study

4.1 Research Questions

Sample Size For AR(1) Mixed Effect Zip Model

In general, the more parameters a model has, the more data points are needed. Also, some parameters are harder than the other parameters to estimate; that is, some parameters need more data points to provide enough information for estimation. The question arises: How big a sample size is needed to estimate the AR(1) Mixed effect ZIP model?

Importance of Random Effect Parameters

The AR (1) model contains fixed effect parameters and extra five random effect parameters to describe a zero inflated Poisson longitudinal data set. The five random effect parameters greatly increase the complexity of the model structure and hence the estimation difficulty (due in large part because the five parameters of the random effects and autoregression structure require in MCMC estimation the generation of correlated effect values for each sampled individual at each time point). A practical concern arises: when the fixed effects parameters are the main interest, how do simpler submodels of the AR(1) model perform, in terms of magnitude of bias and estimation error? In other words, when the full model is misspecified by ignoring some random effects, is estimation of fixed effects robust against this misspecification?

Research Questions

The following research questions were studied in the pilot study in connection with the AR(1) mixed effect zip model.

a) Can we estimate this model using MCMC methods in a computationally efficient way?
b) How do we determine the sample size to obtain sufficiently accurate estimates?
c) Can a simpler model be fitted to actual AR(1) zip data with only minor bias and large computational savings?
d) Which parameters in the AR(1) zip model can be estimated most accurately? Which are estimated least accurately?

4.2 Pilot Study

4.2.1 ZIP, MIXED, and AR(1) Model Features

Full Model and Nested Models

The ZIP model and MIXED model are nested in the AR(1) model, i.e., the full model. We can interpret these two nested models in terms of the AR(1) model structure.

Table 4.1 ZIP, MIXED, and AR(1) Model Structures

Random Effect Parameter	ZIP Model	Mixed Model	AR(1) Model
φ	0	1	$-1 < \varphi < 1$
ψ	0	1	$-1 < \psi < 1$
σ_U	0 [implies $U_{ij} \equiv 0$]	> 0 [and $U_{ij} \equiv U_i$]	> 0
σ_V	0 [implies $V_{ij} \equiv 0$]	> 0 [and $V_{ij} \equiv V_i$]	> 0
ρ_{UV}	---	$-1 \leq \rho_{UV} \leq 1$	$-1 < \rho_{UV} < 1$

Recall that the AR(1) model states $U_{ij} = \varphi U_{i,j-1} + \sqrt{1-\varphi^2} Z_{ij}$, where the random fluctuation term Z_{ij} is independent of $U_{i,j-1}$ and satisfies $E(Z_{ij}) = 0$ and $Var(Z_{ij}) = \sigma_U^2$. If $\varphi = 1$ then $U_{ij} = U_{i,j-1}$ for all j. If the AR(1) model is the true model, fitting the Mixed Model to a data set could cause biased estimates of fixed effect parameters

because of the misspecification of φ and ψ. The ZIP model is only correct for independent data, and could cause bias when analyzing either longitudinal data or cluster data.

Computation Requirements

The complexity of these three models is different, so the requirements for their computation are different too. The numbers of random quantities which must be generated at each MCMC iteration for the case of five time points are tabulated below.

Table 4.2 ZIP, MIXED, AR(1) Model Computation Requirement at Each MCMC Iteration

Data	$y_{ij,}\ j = 1, \cdots, 5.\ i = 1, \cdots, N.$	
Model	Fixed effect parameters, random effect parameters, random effects in a model.	Total computation requirement per MCMC iteration
ZIP	5 Fixed effect parameters: $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1$.	5
Mixed	5 Fixed effect parameters: $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1$. 3 Random effect parameters: $\sigma_u, \sigma_v, \rho_{uv}$. 2 Random effects per subject: u_i, v_i.	5 + 3 + 2*N
AR(1)	5 Fixed effect parameters: $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1$. 5 Random effect parameters: $\sigma_u, \sigma_v, \rho_{uv}, \varphi, \psi$. 10 Random effects per subject: $u_{i1}, \cdots, u_{i5}.\quad v_{i1}, \cdots, v_{i5}$.	5 + 5 + 10*N

From the table we see the AR(1) model requires much more computational work than ZIP and the Mixed model. Thus if interest lies mainly in the fixed effects, it might be a better strategy to fit a simpler model that is a "correct model" or a "nearly correct model" for the data at the cost of some bias.

Slowness and High Autocorrelation

The ZIP, Mixed, and AR(1) models all suffer from the slowness of WinBUGS/OpenBUGS, especially the AR(1) model. It is not surprising for it to take 60

hours to for the AR(1) model to converge with sample size N = 2000 and one predictor variable only. Here are possible reasons:

a) Model Complexity

As shown in the previous paragraph, the AR(1) model requires computational work in proportion to "5 + 5 + 10*N" for each MCMC iteration.

b) Non-standard distributions

All three models involve a ZIP distribution, which is a mixture of a Poisson distribution and a degenerate distribution at zero. There is an approach to specify nonstandard prior and likelihoods in BUGS: the zeros-ones trick (Spiegelhalter et al., 2003d; Ntzoufras, 2009). Thus our three models programs all employ the zero-ones trick. This trick allows arbitrary sampling distributions to be used. However, it has been shown that this method can be very inefficient and give a very high MC error (Spiegelhalter et al., 2003).

4.2.2 Generating Simulated Longitudinal Zero Inflated Poisson Data

There are three subsections presented here. First, we talk about the structure of NLSY97 data; second, the generation of simulated data motivated by the NLSY97 is presented; third, we discuss the sensitivity of parameter choices for the data.

4.2.2.1 Data Structure of NLSY97

In NLSY97, there are 8,953 subjects, who are aged from 12 to 16 in year 1997. The total number of subjects decreases every year, for example, from 8,953 (in year 1997) to 7,851 (in year 2001). Each AGE group is about 20% of the data. For the SEX

variable, the males and females are equal proportions of the data. For the RACE variable, there are four levels: black (26%), Hispanic (21%), mixed race (1%), and non black/Hispanic (non_b_h) (52%). The PEER variable (What percent of your friends are smokers?) is in ordinal scale. The levels are: "1" ="LE 10%", "2" ="ABT 25%", "3" ="ABT 50%", "4" ="ABT 75%", and "5" ="GT 90%". Each subject reports whether he or she ever smoked. Those who respond "yes" are the smoker or Poisson class. Those who respond "no" are the nonsmoker or perfect class. This is a longitudinal survey with 14 annual waves starting in the year 1997. The number of cigarettes smoked per day is the research interest in this dissertation.

The following descriptive statistics were taken from smoker class (Poisson class) members only to present the important features of the NLSY97 data. The descriptive statistics of the number of cigarettes smoked per day from year 1997 to year 2010 are presented in Table 4.3. There is an increasing trend of number of cigarettes smoked per day, but it is not linear.

Table 4.3 Number of Cigarette Smoked Per Day from Year 1997 to Year 2010

Year	Number of Cigarettes Smoked Per Day	Year	Number of Cigarettes Smoked Per Day
1997	5.01	2004	9.05
1998	6.24	2005	9.48
1999	7.28	2006	9.16
2000	8.01	2007	9.16
2001	8.47	2008	9.35
2002	8.63	2009	9.31
2003	8.79	2010	9.36

We will look more closely at data from Year 1997 to Year 2001 in Table 4.4. In the NLSY97 data, the survey question "Have you ever smoked cigarettes before?" classified subjects into "Self-Indicated Smoker" and "Self-Indicated Non-Smoker"

groups. We use this survey question as a proxy variable for the latent class-membership indicator variable, smoker (Poisson class) or non-smoker (perfect class). Although they are not really Poisson class and perfect class, this breakdown is useful in designing our simulation study. Based on this definition, we see the Poisson Class (the smokers) slightly increased from 39% up to 42%. The perfect class (the nonsmokers, the structural zeros) remained at around 60% of the total population. Thus, despite the binning, the NLSY97 can plausibly be modeled as zero inflated Poisson data. The Poisson zeros (the sampling zeros) are 4.1% of the Poisson class in 1997 and 3.2% of the Poisson class in 2001. The observed numbers of (indicated smoker) Poisson zeros from 1997 to 2001 are: 146, 144, 103, 111, and 104. However, the expected numbers of Poisson zeros from Year 1997 to Year 2001 are: 23, 6, 2, 1, and 0.7. These expected numbers of Poisson zeros are obtained by using the observed means from Table 4.3 as λ in this equation:

Expected Poisson Zeros = (size of Poisson class) * (probability of Poisson zero)

$$= n\left(\frac{e^{-\lambda}\lambda^{0}}{0!}\right) = ne^{-\lambda}.$$

As we can see, there are big discrepancies between the numbers of observed Poisson zeros and expected Poisson zeros. This is evidence of an over dispersion problem. Therefore, we will incorporate this real data feature into our simulated data.

Table 4.4 Descriptive Statistics of NLSY97 from Year 1997 to Year 2001

Year	1997	1998	1999	2000	2001
Total subjects	8,953	8,358	8,148	8,027	7,851
Mean of Poisson class Cigarettes Smoked Per Day	5.01	6.24	7.28	8.01	8.47
Perfect class (Self-Indicated Non-Smoker)	5,436	5,317	5,082	4,825	4,591
Percentage of perfect class	62%	65%	64%	61%	60%
Poisson class (Self-Indicated Smoker)	3,517	3,041	3,066	3,202	3,260
Observed Zeros from Poisson Class	146	144	103	111	104
Expected Zeros from Poisson Class	23	6	2	1	0.7

Furthermore, the distribution of the number of cigarettes smoked per day is not quite Poisson either. In Table 4.5 we see some spikes at cigarette numbers 10, 15, 20, and 30. We suspect that there are two latent classes of subjects: light smokers and heavy smokers. The light smokers answered this question in units of numbers of cigarettes, while some heavy smokers answered this question in units of packages of cigarettes.

Table 4.5 Number of Cigarettes Smoked Per Day in Year 1997

Numbers of Cig. Smoked Per Day	Numbers of smokers	Numbers of Cig. Smoked Per Day	Numbers of smokers	Numbers of Cig. Smoked Per Day	Numbers of smokers
0	29	9	2	18	4
1	94	10	50	19	1
2	57	11	3	20	32
3	50	12	4	22	0
4	32	13	3	23	1
5	36	14	0	24	2
6	12	15	29	25	3
7	16	16	5	26	1
8	12	17	5	30	7

As stated above, the distribution of observed numbers of counts does not seem to be Poisson and there is evidence of over dispersion. Random effect models can deal with unmeasured predictors; that is, the random effects part would be treated as a fixed effects part if those explanatory variables had been observable. Random effects also sometimes represent random measurement error in the explanatory variables (Agresti, 2002). Thus, the random effect parameters, σ_u and σ_v, are probably important for analyzing NLSY97 data.

Next, the correlation matrix of smoking status, non-smoker or smoker, is presented in Table 4.6. These correlations provide a good idea what magnitude the correlation should be in logistic regression part when we generate simulated data. Based

on Table 4.6, the first order autocorrelations are around 0.6. We found by simulation experiments that such values corresponded to an AR(1) parameter ϕ of approximately 0.85 to 0.90 in the logistic regression.

Table 4.6 Correlations of Smoking Status, Non-smoker or Smoker, from 1997 to 2001

Year	1997	1998	1999	2000	2001
1997	1.00	.5	.43	.39	.35
1998	.50	1.00	.61	.52	.47
1999	.43	.61	1.00	.64	.55
2000	.39	.52	.64	1.00	.65
2001	.35	.47	.55	.65	1.00

Next, the correlations of numbers of cigarettes smoked are presented in Table 4.7. The observed first order autocorrelation is about 0.5. We found by simulation experiments that such values corresponded to an AR(1) parameter ψ of approximately 0.85 or 0.90 in the Poisson regression.

Table 4.7 Correlations of Number of Cigarettes Smoked, from 1997 to 2001

Year	1997	1998	1999	2000	2001
1997	1.00	.45	.34	.35	.30
1998	.45	1.00	.51	.44	.35
1999	.34	.51	1.00	.55	.40
2000	.35	.44	.55	1.00	.51
2001	.3	.35	.40	.51	1.00

The two autocorrelation matrices above showed that the autocorrelations (ϕ and ψ) are important in the structure of the NLSY97 data. Analyzing the NLSY97 data but ignoring the AR(1) correlations and autocorrelation would be a mistake. The practical question, and a key question addressed in this dissertation, is what impact omitting the autocorrelations (ϕ and ψ) has on inferences about fixed effect parameters.

4.2.2.2 Generating Simulated Data

This simulated data were generated based on a random effect AR(1) ZIP model. An R program for generating the data was written according the formulas in Chapter III:

$$p[y_{ij} = 0] = \pi_{ij} + (1 - \pi_{ij})(1 - \exp(-\mu_{ij}))$$

$$p[y_{ij} = y: y > 0] = (1 - \pi_{ij}) exp(-\mu_{ij}) \mu_{ij}^{y} / y!$$

where

$$\text{logit } \pi_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_{ij}$$

$$\log \mu_{ij} = \gamma_0 + \gamma_1 X_2 + v_{ij}$$

$$u_{ij} \sim N(0, \sigma_u^2)$$

$$v_{ij} \sim N(0, \sigma_v^2)$$

$$Cov(u_{ij}, v_{ij}) = \rho_{uv} \sigma_u \sigma_v$$

$$u_{i,j+1} = \phi u_{i,j} + z_{i,j+1}$$

$$v_{i,j+1} = \psi v_{i,j} + z'_{i,j+1}$$

$$z_{ij} \sim N(0, (1 - \phi^2)\sigma_u^2)$$

$$z'_{ij} \sim N(0, (1 - \psi^2)\sigma_v^2)$$

$$Corr(z_{ij}, z'_{ij}) = \rho_{uv}(1 - \phi\psi)/\sqrt{(1 - \phi^2)(1 - \psi^2)}$$

The subscript $i, i = 1, \cdots, N,$ indexes subjects. The subscript $j, j = 1, \cdots, T = 5,$ indexes time points, which are equally spaced in this study. There are two types of response variables: one is Bernoulli (in the logistic regression for class membership), and the other is counts (in the Poisson regression for members of the Poisson class only). A binary covariate (X_1) with equiprobable values 1 and 0 was created to represent SEX. A quadratic function of time was the second covariate (X_2). The quadratic term is meant to

capture the nonlinear trend in the mean of number cigarettes smoked per day. The two components in the simulated data are:

$$\text{Logistic regression: } \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_{ij}.$$

$$\text{Poisson regression: } \gamma_0 + \gamma_1 X_2 + v_{ij}.$$

The parameter values are:

Fixed Effect	True value	Random Effect	True value
β_0	.00	σ_u	3.50
β_1	.50	σ_v	.80
β_2	-3.00	φ	.85
γ_0	.50	ψ	.85
γ_1	6.00	ρ_{uv}	.10

By manipulating the ten parameter values, the simulated data presents all the features of NLSY97 data such as excessive zeros, excessive Poisson zeros, AR(1) correlations, and auto correlations. Only the bumps in the histograms of the numbers of cigarettes smoked are not simulated.

There are two sample sizes: N = 2000 and N = 5000. The N = 2000 and N = 5000 are used in the section dealing with sample size determination. In the section dealing with model comparison (ZIP, MIXED, and AR(1) models), N = 2000 is used.

Table 4.8 Descriptive Statistics of Simulated Data

Year	1	2	3	4	5
Total subjects	5000	5000	5000	5000	5000
Mean of Poisson class Cigarettes Smoked Per Day	1.845	2.879	4.053	4.989	5.217
Perfect class (Self-Indicated Non-Smoker)	2502	2414	2324	2278	2270
Percentage of perfect class	50%	48%	46%	45%	45%
Poisson class (Self-Indicated Smoker)	2498	2586	2676	2722	2730
Observed Poisson Zeros	314	179	103	70	65

Table 4.9 Correlations of Zero Indicator, from Year 1 to Year 5

Year	1	2	3	4	5
1	1.00	.47	.39	.32	.27
2	.47	1.00	.47	.39	.33
3	.39	.47	1.00	.47	.40
4	.32	.39	.47	1.00	.49
5	.27	.33	.40	.49	1.00

Table 4.10 Correlations of Poisson Counts, from Year 1 to Year 5

Year	1	2	3	4	5
1	1.00	.65	.56	.49	.41
2	.65	1.00	.70	.58	.48
3	.56	.70	1.00	.70	.59
4	.49	.58	.70	1.00	.73
5	.41	.48	.59	.73	1.00

Choices of Parameter Values for Generating Simulated Data

Recall that there is big discrepancy in the NLSY97 data set between the observed Poisson zeros and expected Poisson zeros – there are far more observed Poisson zeros in the data set than theoretical expected Poisson zeros. From the first set of parameter values (see Table 4.11), we generated very few Poisson zeros. This essentially means the logistic regression part is not necessary for ZIP model because there is no need to distinguish Poisson zeros and perfect zeros. In other words, with such a large separation between the data for the two classes, we have nearly distinct data subsets for the two classes. This is less interesting as a mixture problem. By adjusting the parameter values we were able to generate data with nontrivial zeros for the Poisson class, and hence a nontrivial mixture problem.

Table 4.11 Parameter Values of Simulated Data

Fixed Effect	1st set value	Final set value	Random Effect	1st set value	Final set value
β_0	1.0	0	σ_u	3.5	3.5
β_1	0.5	0.5	σ_v	0.4	0.8
β_2	- 4.0	-3.0	φ	0.95	0.85
γ_0	2.0	0.5	ψ	0.95	0.85
γ_1	1.0	6.0	$\rho_{u,v}$	0.7	0.1

4.2.3 Program Language for Bayesian Inference Using Gibbs Sampling (BUGS) Estimation

With the Bayesian approach, the distinction between fixed and random effects is less stark, as every effect has a probability distribution. From a Bayesian perspective all parameters are seen as random quantities arising from proper probability distributions, thus all effects are random. The key concept to distinguish between random and fixed terms is the implicit versus explicit prior distribution.

Consider the OLS regression model $Y = X\beta + e$. From a classical statistical point of view, X is considered a fixed variable matrix, and β is considered as a fixed regression parameter vector, that is, a vector of fixed effects. The only random quantity in this model is the vector e, which is the source of the randomness of Y. However, from a Bayesian view, β is considered as a vector of random variables, because it has an explicit prior in the formulation. From a Bayesian point of view, the classical approach implicitly assumes an improper uniform prior distribution on β. For more examples, see Lynch (2007). For a Bayesian a variable with hyperpriors may be regarded as a random effect; that is, the effects are modeled as exchangeable before data are observed; others are treated as fixed effects.

Regarding the values of hyperpriors, there are Empirical Bayes (EB) and Full Bayes (FB) approaches. EB estimates the highest-level hyperpriors, and then treats the resulting point estimates as known true values to estimate lower-level parameters. This process controverts Bayesian philosophy by "using the data twice" (Lindley, 1969, p. 421).

If EB approach is used to estimate the hyperparameters, this can lead to overoptimistic estimates of precision of the estimated parameters of interest. For example, in an analysis of traffic safety data, Carriquiry and Pawlovich (2004) concluded that "often EB analysts obtain unrealistically low standard errors." The Full Bayes (FB) approach is used in this pilot study.

Choices of Priors and Initial Values

In the simulation study we attempted to use reasonably informative priors for the unknown parameters, because we found that the BUGS program is very sensitive to the choice of the priors and initial values in the ZIP model, the MIXED model, especially in the AR(1) model. For example in a Normal distribution, replacing $\tau = \frac{1}{variance} = 0.1$ with $\tau = 0.01$ will cause the failure of the BUGS program. We believe that if the prior is too vague, extremely large parameter values may be drawn by the MCMC process and that these extreme values lead to numerical instability which cause BUGS to crash. Unfortunately, BUGS does not provide useful diagnostics to identify the cause of a crash.

Since we have sample sizes of 2000 and 5000, we expect the prior effect to wash out because of the large sample size. However, the BUGS program does not always behave in this way unless we use highly informative priors. In simpler models, such as the GLMM logistic model, we arbitrarily chose prior values and the estimates were very

good. For details, readers are referred to Section 4.2.4 dealing with Fisher information and sample size. We present the parameter values, prior choices and initial values in Table 4.12. Initial values of the random effects were selected by bugs from the assumed random effect distributions, given the initial values of the parameters listed in Table 4.12.

Table 4.12 Parameter, Prior, and Initial Values of Simulated Data

Fixed Effect	True value	Prior choices	initial values(1)	initial values(2)
β_0	.00	~dnorm(0, 0.1)	.00	0.25
β_1	.50	~dnorm(0.5, 0.1)	0.50	0.25
β_2	-3.00	~dnorm(-3, 0.1)	-3.00	-2.00
γ_0	.50	~dnorm(0.5, 0.1)	0.50	0.20
γ_1	6.00	~dnorm(6, 0.1)	6.00	5.00
Random Effect				
σ_U	3.50	Tau.u ~gamma(0.1, 1.2)	Tau.u =0.082	Tau.u =0.07
σ_V	.80	Tau.v ~gamma(0.1, 0.006)	Tau.v =1.6	Tau.v =1
φ	.85	~dbeta(2.4, 0.6)	.85	.80
ψ	.85	~dbeta(2.4, 0.6)	.85	.80
$\rho_{u,v}$	.10	Rho.z~ dbeta(.3, 2.7)		

A Constraint In Bugs Program For AR(1) Mixed Zip Model

There is a parameter constraint in the AR(1) Mixed ZIP model. It is not obvious to see this constraint when writing the model and corresponding BUGS program, so we mention it here:

$$\rho_{uv} = \rho_{ee'} * c$$

where $\rho_{ee'} = (1 - \varphi\psi)\rho_{uv}/\sqrt{(1-\varphi^2)(1-\psi^2)}$, and

$$c = \frac{\sqrt{(1-\varphi^2)(1-\psi^2)}}{1-\varphi\psi}.$$

The term *c* has a maximum of 1 when $\varphi = \psi$, and is smaller than 1 otherwise. It is never negative if φ and ψ are in (-1,1). Since $\rho_{ee'}$ has to be less than 1, then given particular values of φ and ψ, ρ_{uv} will be bounded usually by something less than 1.

Thus, if ρ_{uv} gets too big, the $\rho_{ee'}$ is forced to be bigger than 1, which is mathematically invalid and will cause problems for the algorithm. It is better to put the prior on $\rho_{ee'}$ rather than ρ_{uv}.

4.2.4 Use of Fisher Information for Determining Sample Size

The AR(1) Mixed effect ZIP model was developed in an attempt to incorporate all sources of variability in a longitudinal ZIP data set described in Chapter III; hence the model structure is quite complicated. In general, the more parameters a model has, the more data points are needed. Also, some parameters are harder than others to estimate; that is, some parameters need more data points to provide enough information for estimation.

We are concerned about how many observations are needed to ensure the data contains enough information to carry out the estimation with desirable standard errors of estimates. We approach this problem by generating empirical approximations of Fisher information for submodels of the AR(1) model in order to obtain lower bounds for sample sizes for the parameters of primary interest, namely the fixed-effect parameters in the regression models. We then extend these results to obtain the full model sample size.

In section 4.2.4.1, we discuss the ideas of using Fisher Information for determining sample size and the approach to this problem. We start from GLMM: random intercept logistic regression and random intercept Poisson regression in Section 4.2.4.2. After that, we investigate the AR(1) mixed effect ZIP model in Section 4.2.4.3. The most difficult parameters to estimate in the AR(1) Model will be discussed in Section 4.2.4.4.

4.2.4.1 Ideas and Approach for Determining Sample Size

Ideas

When observations are i.i.d., the diagonal elements of the inverse of the Fisher information matrix of the parameters in a model, divided by the sample size, are asymptotic sampling variances of the maximum likelihood estimators of the respective parameters. Therefore they provide a gauge for determining how a large a sample is needed for the distribution of estimates of a given parameter to be concentrated near the true value. A smaller sample would produce estimates with large dispersion and which would be indistinguishable from false values. Thus correct parameter values would be difficult to identify.

Recall that the expected negative second derivative of the log likelihood is the Fisher information matrix (Bickel & Doksum, 2007, pp. 179-188) in *N* observations:

$$E\left[\left[-\frac{\partial^2}{\partial\theta_r\partial\theta_s}logL(\boldsymbol{\theta})\right]\right] = \boldsymbol{\mathcal{I}}_N = \text{Fisher information in } N \text{ observations.}$$

According to large sample theory, if the observations are *i.i.d.*, then $\boldsymbol{\mathcal{I}}_N \to \infty$ but $\frac{1}{N}\boldsymbol{\mathcal{I}}_N \to \boldsymbol{\mathcal{I}}_1$, a non-random matrix that depends on the parameters but is not dependent on sample size. In large samples the distribution of the normalized maximum likelihood estimator is approximately $\sqrt{N}\,(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \sim N(\mathbf{0}, \boldsymbol{\mathcal{I}}_1^{-1})$ where $\widehat{\boldsymbol{\theta}}_N$ is the MLE of $\boldsymbol{\theta}$. Therefore in large samples the MLE has approximate covariance matrix

$VarCov\,(\,\widehat{\boldsymbol{\theta}}_N) \doteq \frac{1}{N}\,\boldsymbol{\mathcal{I}}_1^{-1}$Note that $\boldsymbol{\mathcal{I}}_N = N\boldsymbol{\mathcal{I}}_1$ and $(\boldsymbol{\mathcal{I}}_N)^{-1} = \boldsymbol{\mathcal{I}}_N^{-1} = \frac{1}{N}\boldsymbol{\mathcal{I}}_1^{-1}$.

In Bayesian estimation, if $\widehat{\boldsymbol{\theta}}_B$ is the Bayes estimator of $\boldsymbol{\theta}$, then

$\sqrt{N}(\widehat{\boldsymbol{\theta}}_B - \widehat{\boldsymbol{\theta}}_N) \xrightarrow{P} 0$ (by the Bernstein – von Mises theorem). Therefore, by combining these facts we can say that $VarCov\,(\,\widehat{\boldsymbol{\theta}}_B) \doteq \frac{1}{N}\,\boldsymbol{\mathcal{I}}_1^{-1}$ when *N*, is large but

finite. Markov Chain Monte Carlo methods produce an estimator of the covariance matrix of $\boldsymbol{\theta}_B$. Therefore we can estimate $\boldsymbol{\mathcal{I}}_1^{-1}$ by $N\left(\widehat{Var Cov}\left(\widehat{\boldsymbol{\theta}}_B\right)\right)$. The approximate relationship $\boldsymbol{\mathcal{I}}_1^{-1} \approx N\left(\widehat{Var Cov}\left(\widehat{\boldsymbol{\theta}}_B\right)\right)$ improves as N increases. Furthermore, if N_1 and $N_2$are both large enough, then

$$N_1\left(\widehat{Var Cov}\left(\widehat{\boldsymbol{\theta}}_B\right)\right) \approx N_2\left(\widehat{Var Cov}\left(\widehat{\boldsymbol{\theta}}_B\right)\right) \approx \boldsymbol{\mathcal{I}}_1^{-1}.$$

In a practical problem, we don't know how big N has to be to get an accurate estimate of $\boldsymbol{\mathcal{I}}_1^{-1}$. In a simulation one can generate and analyze samples of two different (large) sizes. As a measure of the closeness of two estimated inverse Fisher information matrices $\boldsymbol{\mathcal{I}}_{1,N_1}^{-1}$ and $\boldsymbol{\mathcal{I}}_{1,N_2}^{-1}$, estimated from sample sizes N_1 and N_2, one can either look at

$$\left|\lambda_1\left(\boldsymbol{\mathcal{I}}_{1,N_1}^{-1} - \boldsymbol{\mathcal{I}}_{1,N_2}^{-1}\right)\right|,$$

where $\lambda_1(\boldsymbol{A})$ is the maximum eigenvalue of a matrix $\boldsymbol{A}$.

or look at

$$\sqrt{\sum_i \sum_j \left|\hat{\boldsymbol{\mathcal{I}}}_{1,N_1}(i,j) - \hat{\boldsymbol{\mathcal{I}}}_{1,N_2}(i,j)\right|^2}$$

where $\boldsymbol{\mathcal{I}}_{1,N_1}^{-1}(i,j)$ is the (i,j) element of $\boldsymbol{\mathcal{I}}_{1,N_1}^{-1}$. If $\boldsymbol{\mathcal{I}}_{1,N_1}^{-1} \approx \boldsymbol{\mathcal{I}}_{1,N_2}^{-1}$ we can assume that we have estimated $\boldsymbol{\mathcal{I}}_1^{-1}$ with reasonable accuracy. Thus, by monitoring the approximate Fisher information for different sample sizes we can determine the ideal sample size for the particular model and level of accuracy that we are interested in.

Approach

It is known that calculating the observed Fisher information can be computationally challenging, especially in a complicated model setting. At this time there

is no convenient way to get the observed Fisher Information in general for models as complex as the AR(1) Mixed Effect ZIP model. Essentially we will use available empirical methods to approximate Fisher information of relevant submodels of the full model. The GLMM model is simpler than the ZIP model with random effects (Mixed ZIP Model). Examining the GLMM model (either Poisson or logistic) might give some insight about the Mixed ZIP Model. The SAS PROC GLIMMIX software computes MLE's for GLMMs.

First, we use the Hessian matrix (matrix of second derivatives of log likelihood, which estimates the negative Fisher information matrix) in the SAS output to approximate the Fisher information for the GLMM logistic model and GLMM Poisson model, respectively. Secondly, we will get posterior samples from OpenBUGS software using the CODA output. These posterior samples yield estimates of the Fisher Information and its eigenvalues using R software, for the GLMM Logistic model and GLMM Poisson model, respectively.

Although the Fisher information plays a central role in MLE, we don't know of any paper which computes the observed Fisher information in a Bayesian estimation setting. According to theory, under regularity conditions the covariance matrices of the MLE and a Bayesian estimate are asymptotically equivalent. However, we aren't sure how fast the Bayes estimates will converge to the MLE's because we don't know how strongly the prior information affects the Bayes estimates for finite sample sizes.

We expect to see discrepancies between the Bayes estimates and MLE when the sample sizes are small. The discrepancies are due to the assumed prior information, MCMC error, sampling error, plus the differences between the two methods. Thus, in the

following step, we compare the Fisher Information estimates between MLE and Bayesian estimation.

We will start from a smaller sample size and move to a larger sample size. Using SAS helps to check on how close the Bayes computations are to the MLE computations when *N* is large. In this way, we will approximate the theoretical quantity in a numerical example. Also, we can make sure our approach to Fisher Information in Bayesian analysis is on the right track. However, after examining estimates of Fisher information in smaller samples, using both SAS (MLE) and BUGS (MCMC), we will only use SAS in larger sample sizes because of the BUGS computational burden. While SAS might need only a couple of seconds to obtain the Hessian matrix, BUGS might take several hours to reach model convergence with a sample size of 10,000.

After our study of GLMM, we will investigate the sample size for the AR(1) Mixed effect model, which is our main interest. Here we will use BUGS (MCMC) only but not MLE. The reason is that we can't compute the MLE in large samples because the MLE requires integrals which can not be calculated analytically:

$$L = \prod_{i=1}^{N} \int f(\boldsymbol{u}_i, \boldsymbol{v}_i) d\boldsymbol{u}_i d\boldsymbol{v}_i$$

$$\times \left\{ \prod_{j=1}^{T} \left[\pi\left(\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + u_{ij}\right) I\{y_{ij} = 0\} \right.\right.$$

$$\left.\left. + \left(1 - \pi\left(x_{ij} + u_{ij}\right)\right) exp\left[-exp\left(\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + v_{ij}\right)\right] \times \frac{exp\left(\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + v_{ij}\right) y_{ij}}{y_{ij}!} \right] \right\}$$

where

$\pi\left(x_{ij}^T\beta + u_{ij}\right) = \frac{exp\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+u_{ij}\right)}{1+\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+u_{ij}\right)}$ and

f depends on $\sigma_u, \sigma_v, \rho_{uv}, \varphi,$ and ψ.

As in the previous step, we will start from a smaller sample size and move to larger sample sizes until the change of estimated Fisher information per observation becomes small. For such a sample size, we conclude that the covariance matrix of the Bayes estimates are close to those of the MLE. Such an *N* would appear to guarantee that the sample Fisher information and the theoretical information are nearly equal.

4.2.4.2. Analyses of Fisher Information of GLMMs on MLE and MCMC

Generate Simulated Data

We create repeated measure data with three covariates (X_1, X_2, X_3) and t responses ($Y_1 \sim Y_t$), where t is the number of time points. Each subject *i* shares its unique random effect, u_i, with each response and forms a cluster. The response variable for logistic regression is a Bernoulli variable, and the response variable for Poisson regression is a count variable. The two models are:

Logistic regression: $logit[P(Y = 1)] = \beta_1 + \beta_2 X_{\geq 1} + \beta_3 X_2 + \beta_4 X_4 + u_i.$

Poisson regression: $\log[\mathrm{E}(\mathrm{Y})] = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_4 + v_i.$

The parameter values for these regressions are: $\beta_1 = -.5, \quad \beta_2 = .1, \quad \beta_3 = .2,$ $\beta_4 = -.3, \quad \sigma_u = .5, \quad \sigma_v = .5.$

Estimation and Sample Size of Simulated Data

Two estimation methods are used. One is MLE, in which SAS PROC GLIMMIX is used. The Hessian matrix will be obtained from PROC GLIMMIX. Then the matrix will be used to get the Fisher information and eigenvalues through R software. The other

method is MCMC, in which OpenBUGS is used. The coda output in BUGS contains samples that MCMC draws from posterior distribution. The sample covariance matrix of these estimates will be calculated using R software and inverted to approximate the Fisher information. There are 3 sample sizes for both regressions:

(1) $N = 300\,,\ t = 5.$ (2) $N = 1200\,,\ t = 5.$ (3) $N = 10000\,,\ t = 5.$

Results from Logistic Regression

In this subsection three tables are presented with sample sizes 300 by 5, 1200 by 5, and 10000 by 5, respectively. Parameter estimates and their associated standard errors, Fisher Information matrices, and eigenvalues are provided in each table. For sample sizes 300 and 1200, results from MLE and MCMC method are both presented; for sample size 10000, only results from MLE are presented.

From Table 4.13, we see the MLE and MCMC estimations are in the right direction but not very accurate. There are some discrepancies among the estimations from the two methods. The small sample size (N=300) also makes these two methods unstable. From Table 4.14, we see the MLE and MCMC estimations are getting more accurate and the discrepancies between the estimates from the two methods are getting smaller. The bigger sample size (N=1200) makes the estimation better and brings MLE and MCMC results closer together.

Next we compare the Fisher Information matrix per cluster (with fixed t) and eigenvalues from MLE between Tables 4.14 and 4.15. We consider the estimates from Table 4.13 unstable because of the small sample size and exclude them from the comparison. The average Fisher Information matrix seems to tend to a limiting matrix as N gets large, as we move from the second to the third Binomial example (Tables 4.14 and

Table 4.15). The Binomial-case limit is reached around *N*=1200 since we see there is not much change in the average information as *N* increases to *N*=10000.

Table 4.13 Logistic Regression, $N = 300$, $t = 5$

Parameter	True value	Estimate (MLE)	se (MLE)	Estimate(MCMC)	Se(MCMC)
Intercept	β_1 -.5	-.4494	.1107	-.5328	.1185
x1	β_2 .1	.0165	.1259	.0945	.1324
x2	β_3 .2	.2916	.0687	.3587	.0724
x3	β_4 -.3	-.3269	.0761	-.2957	.0801
u	σ_u .5	.4302	.0531	.5422	.0833
	Estimated Fisher information per subject/cluster from MLE				
	β_1	β_2	β_3	β_4	σ_u
$\boldsymbol{\beta_1}$	0.8667	0.3815	0.1392	0.8144	0.8144
$\boldsymbol{\beta_2}$		0.3815	-0.0124	0.3694	0.0561
$\boldsymbol{\beta_3}$			0.7480	0.0915	-0.0252
$\boldsymbol{\beta_4}$				1.3475	0.1588
$\boldsymbol{\sigma_u}$					0.3183
	Eigenvalues from MLE				
$\boldsymbol{\lambda}$	2.1552	0.7431	0.3218	0.2910	0.1508
	Estimated Fisher information per subject/cluster from MCMC				
	β_1	β_2	β_3	β_4	σ_u
$\boldsymbol{\beta_1}$	0.7343	0.3125	0.1072	0.6931	0.1069
$\boldsymbol{\beta_2}$		0.3333	-0.0088	0.2959	0.0264
$\boldsymbol{\beta_3}$			0.6943	0.0548	-0.0734
$\boldsymbol{\beta_4}$				1.1730	0.1366
$\boldsymbol{\sigma_u}$					0.5189
	Eigenvalues from MCMC				
$\boldsymbol{\lambda}$	1.8342	0.7243	0.4663	0.2917	0.1374

Table 4.14: Logistic Regression, $N = 1200$, $t = 5$

Parameter	True value	Estimate (MLE)	S e (MLE)	Estimate(MCMC)	Se(MCMC)
Intercept	β_1 -.5	-.4352	.0593	-.4800	.0631
x1	β_2 .1	.0662	.0642	.0805	.0679
x2	β_3 .2	.1755	.0322	.1884	.0344
x3	β_4 -.3	-.3156	.0417	-.3156	.0433
u	σ_u .5	.5202	.0293	.6061	.0502
	Estimated Fisher information per subject/cluster from MLE				
	β_1	β_2	β_3	β_4	σ_u
β_1	.8092	.3846	.0085	.7620	.1107
β_2		.3846	-.0041	.3624	.0509

β_3			.8049	.0354	-.0227
β_4				1.2003	.1317
σ_u					.2603
		Eigenvalues from MLE			
λ	1.9837	.8063	.2917	.2408	.1367
		Estimated Fisher information per subject/cluster from MCMC			
	β_1	β_2	β_3	β_4	σ_u
β_1	.7247	.3468	.0211	.7006	.1124
β_2		.3463	-.0099	.3360	.0455
β_3			.7034	.0381	-.0108
β_4				1.1230	.1308
σ_u					.3485
		Eigenvalues from MCMC			
λ	1.8315	.7019	.3279	.2633	.1214

Table 4.15 Logistic Regression, $N = 10000$, $t = 5$

Parameter	True value	Estimate (MLE)	Se (MLE)	Estimate(MCMC)	Se(MCMC)
Intercept	β_1 -.5	-.5435	.0207		
x1	β_2 .1	.1407	.0222		
x2	β_3 .2	.2154	.0112		
x3	β_4 -.3	-.2944	.0139		
u	σ_u .5	-.5435	.0207		
		Estimated Fisher information per subject/cluster from MLE			
	β_1	β_2	β_3	β_4	σ_u

β_1	.8061	.4062	.0421	.7691	.1176
β_2		.4062	-.0106	.3872	.0551
β_3			.7965	.0503	-.0259
β_4				1.2510	.1408
σ_u					.2628
		Eigenvalues from MLE			
λ	2.0434	.7955	.3089	.2412	.1335

Results from Poisson Regression

We make comparisons for Poisson regression similar to those for Logistic regression. Again, we compare the Fisher information matrix per cluster (with fixed *t*) and eigenvalues from MLE between Tables 4.17 and 4.18 only. The convergence of the estimated average Fisher information matrices seems slower in the Poisson case. As we move from the second to the third Poisson example (Tables 4.17 and 4.18), we still find noticeable differences between *N*=1200 and *N*=10000.

Table 4.16 Poisson Regression, $N = 300$, $t = 5$

Parameter	True value	Estimate (MLE)	S e (MLE)
Intercept	β_1 -.5	-.4632	.0786
x1	β_2 .1	.1086	.0896
x2	β_3 .2	.0467	.0484
x3	β_4 -.3	-.3148	.0539
V	σ_u .5	.4465	.0236

Estimated Fisher information per subject/cluster from MLE

β_1	β_2	β_3	β_4	σ_v

β_1	1.7019	.7764	.2115	1.4731	.7197
β_2		.7764	.0026	.7058	.3212
β_3			1.4803	.0794	.0559
β_4				2.4284	.6550
σ_v					1.8087
		Eigenvalues from MLE			
λ	4.3453	1.4824	1.4227	.6476	.2978

Table 4.17 Poisson Regression, N = 1200 , t = 5

Parameter	True value	Estimate (MLE)	S e (MLE)	Estimate(MCMC)	Se(MCMC)
Intercept	β_1 *-.5*	-.5508	.0441		
x1	β_2 *.1*	.1353	.0476		
x2	β_3 *.2*	.2171	.0240		
x3	β_4 *-.3*	-.3033	.0307		
V	σ_v *.5*	.5167	.0137		

Estimated Fisher information per subject/cluster from MLE

	β_1	β_2	β_3	β_4	σ_v
β_1	1.4717	.7205	.1323	1.2998	.5818
β_2		.7205	.0488	.6419	.2706
β_3			1.4601	.1612	.0142
β_4				2.0309	.5485
σ_v					1.332
		Eigenvalues from MLE			
λ	3.7325	1.4502	1.0407	.5444	.2474

Table 4.18 Poisson Regression, N = 10000 , t = 5

Parameter	True value	Estimate (MLE)	S e (MLE)
Intercept	β_1 *-.5*	-.5023	.0149
x1	β_2 *.1*	.1182	.0160
x2	β_3 *.2*	.1886	.0080
x3	β_4 *-.3*	-.2977	.0099
V	σ_v *.5*	.4994	.0044

Estimated Fisher information per subject/cluster from MLE

β_1	β_2	β_3	β_4	σ_v

β_1	1.5553	.7962	.1662	1.3838	.6209
β_2		.7962	.0653	.7066	.3139
β_3			1.5551	.1551	.0351
β_4				2.2331	.5832
σ_v					1.5224
		Eigenvalues from MLE			
λ	4.0452	1.5400	1.2081	.6115	.2572

Conclusions

a) In both logistic regression and Poisson regression cases, the PROC GLIMMIX results look reasonably satisfactory and show that the logistic regression parameters can be well estimated once *N* gets as large as 1000 or 2000.

b) We see that Bayesian computations lead to reasonably good approximations of the per observation Fisher Information matrix. Furthermore, the Bayesian computations give good estimates of ML variance. We expect that when we move to the larger samples, the posterior covariance matrix and MLE covariance matrix will be reasonably similar.

4.2.4.3 MCMC Analyses of Fisher Information of AR(1) Model and Determining the Sample Size

Simulated samples of sizes of 2000 and 5000 were analyzed under the AR(1) model for purpose of comparison. The AR(1) model with sample size of 5000 was run with 50000 iterations. A sample of 10000 sets of 10 estimates drawn from the

approximate posterior distribution (there are 10 parameters in the model) was obtained using BUGS's coda function, which provides values drawn in the MCMC chains. These values were used to compute an estimated variance-covariance matrix. The inverse of this estimated covariance matrix is used to estimate the Fisher information, its eigenvalues, and its eigenvectors. Results are shown in Tables 4.19 and 4.20.

Table 4.19 AR(1) Model, $N = 2000$, $t = 5$

Parameter	True value	Estimate	Sd	MC_error	val2.5pc	val97.5pc
β_1	0	-.0693	.1814	.0072	-.4252	.2914
β_2	.50	.3985	.1481	.0061	.1059	.6855
β_3	-3.00	-2.5690	.7912	.0320	-4.1960	-1.0170
γ_1	.50	.4994	.0560	.0034	.3919	.6113
γ_2	6.00	6.0580	.2108	.0089	5.6600	6.4760
φ	.85	.8338	.0169	.0014	.7992	.8664
ψ	.85	.8504	.0086	5.339E-4	.8326	.8664
ρ_{uv}	.10	.0995	.0650	.0058	6.694E-5	.2529
σ_u	3.50	3.2140	.2313	.0204	2.7800	3.6800
σ_v	.80	.8048	.0166	.0012	.7713	.8369

Estimated Fisher information per subject/cluster from MCMC

	β_1	β_2	β_3	γ_1	γ_2	φ	ψ	ρ_{uv}	σ_u	σ_v
β_1	.033	-.01	-.116	.001	-.007	.000	.000	-.001	.000	.000
β_2		.022	-.004	.001	-.002	.000	.000	.001	.003	.000
β_3			.626	-.009	.037	.002	.000	-.001	-.032	.000
γ_1				.003	-.01	.000	.000	.002	.002	.000
γ_2					.044	.000	.000	-.001	-.002	.000
φ						.000	.000	.000	-.002	.000
ψ							.000	.000	.000	.000
ρ_{uv}								.004	.004	.000
σ_u									.054	.000
σ_v										.000

Eigenvalues from MCMC

λ	8.813	2.926	2.073	1.418	.124	.109	.018	.011	.009	.000

Table 4.20 AR(1) Model, $N = 5000$, $t = 5$

Parameter	True value	Estimate	Sd	MC_error	val2.5pc	val97.5pc
β_1	0	-.0377	.1227	.0065	-.2856	.1963
β_2	.50	.5484	.1015	.0053	.3547	.7515
β_3	-3.00	-2.7250	.5148	.0265	-3.7500	-1.7150
γ_1	.50	.4864	.0347	.0023	.4146	.5516
γ_2	6.00	6.0970	.1282	.0061	5.8470	6.3590
φ	.85	.8560	.0101	9.519E-4	.8352	.8746
ψ	.85	.8545	.0054	4.395E-4	.8435	.8643

ρ_{uv}	.10	.1204	.0424	.0040	.0217	.1942
σ_u	3.50	3.5110	.1512	.0141	3.2250	3.7820
σ_v	.80	.8094	.0097	8.988E-4	.7920	.8276

The Fisher information per subject/cluster from MCMC

	β_1	β_2	β_3	γ_1	γ_2	φ	ψ	ρ_{uv}	σ_u	σ_u
β_1	.072	.037	.014	-.055	-.012	.028	-.012	.023	.007	.046
β_2		.040	.007	-.029	-.007	.020	.002	.018	.001	.007
β_3			.004	-.008	-.002	.002	.000	.002	.002	.005
γ_1				1.318	.263	.016	-.184	-.607	-.008	.349
γ_2					.065	-.005	-.032	-.120	-.002	.059
φ						3.498	.022	-.051	.157	-.050
ψ							7.659	.160	-.015	-1.606
ρ_{uv}								.400	-.005	-.256
σ_u									.018	-.007
σ_v										2.658

Eigenvalues from MCMC

λ	8.150	3.508	2.357	1.487	.097	.095	.017	.012	.008	.001

Discussion of Results

a) Although $N = 2000$ is sufficient for GLMM, it is obviously insufficient for the AR(1) model. For example, some diagonal elements in the Fisher information matrix are smaller than 10^{-5} (coded as 0), which means almost no information for that particular parameter. We see that the Fisher Information is much improved when $N = 5000$.

b) For this particular data, the Fisher information indicates some parameters are hard to estimate, in terms of variance. These parameters are $\beta_3, \sigma_u, \beta_2, \gamma_2, \beta_1$(see elements in the Fisher information matrix, *N* = 5000).

4.2.4.4 The Most Difficult Parameters to Estimate in the AR(1) Model

In this section we use information from eigenvalues and eigenvectors of matrices related to the Fisher information matrix to obtain insight into the AR(1) model. In principal component analysis, the eigenvector with the largest eigenvalue is the direction along which the data set has the maximum variance. In our problem we use the eigenvalues and eigenvectors of the estimated Fisher information matrix in a similar way. The largest eigenvalue corresponds to a linear function of the parameters with maximum information, or equivalently the function estimated with the smallest asymptotic variance. Similarly, the smallest eigenvalue corresponds to a linear function estimated with the largest asymptotic variance.

In Table 4.21 the ten eigenvalues of the estimated information matrix (based on N=5000) are presented. Based on Table 4.21 we would say there are four large eigenvalues. The other six eigenvalues are quite small, especially the last two eigenvalues.

Table 4.21 The Ten Eigenvalues of AR(1) Model

8.150428495	3.507783275	2.356704440	1.487195778	0.097264029
0.094725558	0.016639069	0.012097777	0.008334560	0.000724701

The eigenvectors are presented in Table 4.22. This is a 10 by 10 matrix. The rows correspond to the ten parameters and the columns present the ten eigenvectors corresponding to the eigenvalues from the largest to the smallest. Here the smaller eigenvalues and eigenvectors are more important to investigate. The elements/loadings in each eigenvector are normalized such that sum of squares of the loadings in each eigenvector is equal to one. The regression coefficients and predictors in the two regression models are:

Logistic regression: $\beta_1 + \beta_2 * sex + \beta_3 * \left(\frac{time}{5}\right)(1 - \frac{time}{5}) + u_{ij}$

Poisson regression: $\gamma_1 + \gamma_2 * \left(\frac{time}{5}\right)(1 - \frac{time}{5}) + v_{ij}$

First, we highlight in yellow the large loadings in the first four eigenvectors. We interpret these eigenvectors as the linear functions of parameters that are easiest to estimate. These involve the parameters γ_1, γ_2, φ, ψ, ρ_{uv}, σ_u, σ_v. The second eigenvector essentially identifies the parameter φ. Note that none of the three logistic regression coefficients have high loadings in the first four eigenvectors.

Next, we highlight in blue the large loadings in the smaller six eigenvectors. We conclude that linear functions of the three logistic regression coefficients were the hardest parameters to estimate, especially β_3. Note that the tenth eigenvector, which has an eigenvalue near zero, essentially identifies β_3. In general, parameters related to logistic regression are harder to estimate. We also notice that regression coefficients involving the quadratic predictor, β_3 and γ_2, are harder to estimate. These results are consistent with the discussion in Section 4.2.4.3.

Let's take β_3 as an example. In simulation, AR(1) model with *N* = 5000, the estimated s.e. was 0.5148, and λ_{10} was 0.000724701. If we want the estimated s.e. to be less than 0.5, the *N* has to be 5524:

$se\left(q_{10}^T\hat{\theta}\right) = \frac{1}{\sqrt{N\lambda_{10}}} < 0.5$

$N \geq \frac{1}{(0.5)^2(.000724)}$ = 5524.

Table 4.22 Parameters and Eigenvectors of AR(1) Model

$\boldsymbol{\lambda}$	**8.150428495**	**3.5077832750**	**2.3567044400**	**1.487195778**	**0.097264029**
β_1	0.002627000	0.0078315629	0.0042732218	0.053230019	0.066761348

β_2	0.000215518	0.0055340834	0.0034163521	0.025205583	0.027109138
β_3	0.000087200	0.0004505660	0.0002479095	0.006935629	0.020035937
γ_1	0.043540820	0.0088705441	0.3796528605	0.797113660	0.412493910
γ_2	0.007753001	0.0004909037	-0.0726045401	0.163163534	-0.099486900
φ	-0.006987224	-0.9983729333	-0.0179533414	-0.024233796	-0.016183366
ψ	-0.957105300	0.0135031324	-0.2784509125	-0.078807726	0.002222653
ρ_{uv}	-0.032675910	0.0167047758	0.2116909238	-0.372488909	-0.898067529
σ_u	0.001366294	-0.0449497463	0.0038878937	-0.003488855	0.081163939
σ_v	0.284373300	0.0244649188	0.8531570177	-0.434640681	-0.031272926
$\boldsymbol{\lambda}$	**0.0947255580**	**0.0166390690**	**0.0120977770**	**0.008334560**	**0.0007247010**
β_1	0.8158251911	0.4707308382	-0.0025058171	0.268145184	-0.1831846257
β_2	0.5466244745	-0.7718167730	0.0682756898	-0.315034824	-0.0133967069
β_3	0.1643590362	0.0975165962	-0.0570951533	-0.008923131	0.9796363244
γ_1	0.0688304168	0.0100104426	-0.2075381777	0.018100629	-0.0104995302
γ_2	0.0004384083	0.1549497234	0.9486398717	-0.180315659	0.0413194493
φ	-0.0127023489	-0.0153659508	0.0113242856	0.038553616	0.0043682125
ψ	-0.0029583784	0.0004382848	-0.0002940427	-0.002751346	-0.0001719452
ρ_{uv}	0.0218496526	0.0733196724	-0.0333442856	-0.039944026	0.0025231199
σ_u	0.0506938926	0.3787757157	-0.2187146712	-0.890368592	-0.0687715847
σ_v	-0.0257264521	-0.0030187629	0.0025136791	-0.009193641	0.0020120047

Chapter V

Data Analyses

In Chapter V We present the results of two data analyses. In Section 5.1 we compare three models using simulated data generated from the ZIP AR(1) MODEL. In Section 5.2 we use the most promising model chosen from Section 5.1 to analyze the real data – NLSY97.

5.1 Model Comparison

We created simulated data from the AR(1) MODEL (see details in Section 4.2.4) with two sample sizes, $N = 2000$ and $N = 5000$, and fitted the ZIP, MIXED, and AR(1) models. Recall that the ZIP model has no random effects and all observations are assumed to be independent. The MIXED model has three random effect parameters, which describe within cluster random effects. Subjects are independent but observations within subject are dependent. The AR(1) model has five random parameters, which model longitudinal data with autocorrelated random effects. The subjects are still independent.

Since the data is simulated from an AR(1) model, we know all the true parameter values. Thus, the DIC is not necessarily needed for the model comparison. We expected that the AR(1) model would perform the best. However, we would like to see if the two simpler models, ZIP and MIXED, can produce acceptable bias level on the fixed effects. In other words, we want to investigate how important the random effect parameters are for the estimation of the fixed effects.

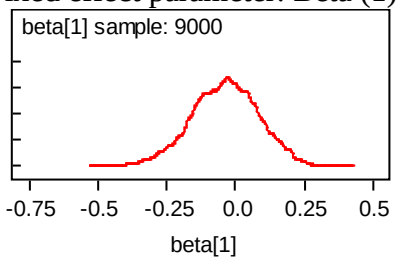
Results and Discussion

The results are presented in Table 5.1. For sample size $N = 2000$, the results are summarized as follows: in terms of bias, the AR(1) model is the best and the ZIP model is the worst one; in terms of variance, the ZIP model is the best and the AR(1) is the worst one; in terms of MSE, the MIXED model is the best. We also present the 95% credible intervals (CI95) for these models (Figure 5.1.1--Figure 5.1.5). For the ZIP model, four parameters were out of the CI95, and one parameter is on the boundary. For MIXED model, three parameters were in the CI95, one parameter is on the boundary, and one parameter is outside the boundary. For the AR(1) model, all five parameters were inside of the CI95.

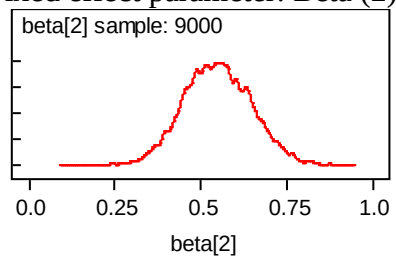
Since the AR(1) model is the true model, consequently its bias is smaller than the other two models. Note that, although AR(1) is the true model, there is some bias in its parameter estimates because of the priors. We expect these biases will decrease as the sample size increases. These three models have the same amount of information (data) so that the ZIP model performs the best in variance estimation because there are fewer parameters to estimate.

In the following, we present the posterior density of AR(1) model parameters (simulated data sample size $N = 5000$). Evidence of nonconvergence was not found for fixed effect parameters; however, the random effect parameters need more time to reach their stationary distribution.

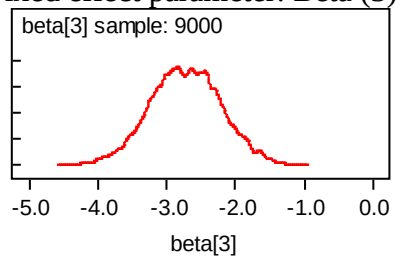
a) Fixed effect parameter: Beta (1)



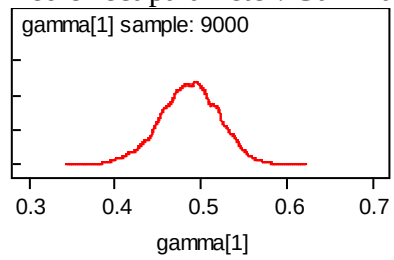
b) Fixed effect parameter: Beta (2)



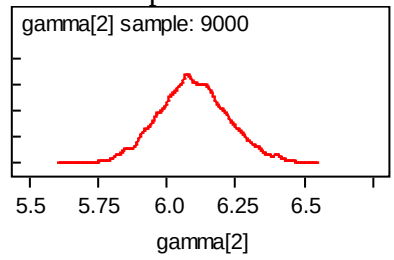
c) Fixed effect parameter: Beta (3)



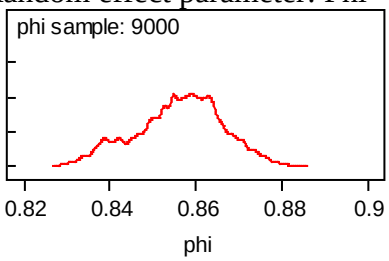
d) Fixed effect parameter: Gamma (1)



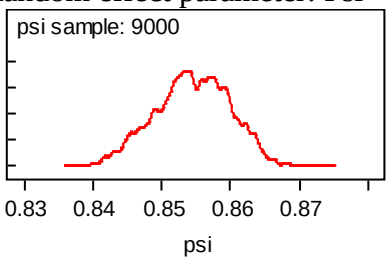
e) Fixed effect parameter: Gamma (2)



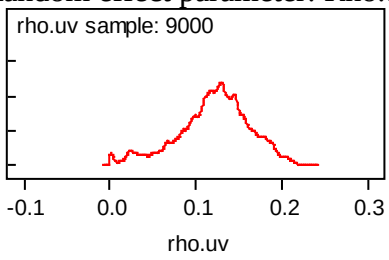
f) Random effect parameter: Phi



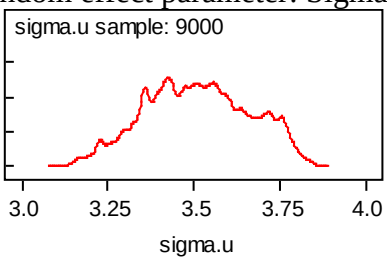
g) Random effect parameter: Psi



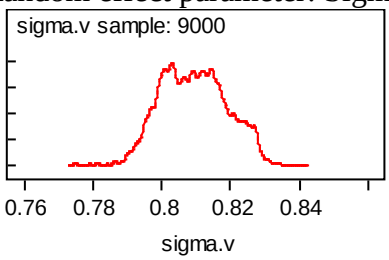
h) Random effect parameter: Rho.uv



i) Random effect parameter: Sigma U



j) Random effect parameter: Sigma V



For sample size N = 5000, the results are summarized as follows: in terms of bias, the AR(1) model is better; in terms of variance, the MIXED model is the better; in terms of MSE, the AR(1) model is better. We also present the 95% credible intervals (CI95) for these two models (Figure 5.1.1--Figure 5.1.5). For the MIXED model, two parameters were in the CI95, two parameters are on the boundary, and one parameter is outside the boundary. For AR(1) model, all five parameters were inside of the CI95.

When N increases from 2000 to 5000, we see that the posterior standard deviation decreases by one third. Recall the large sample theory for MLE, which states that the variance and sample size are related by:

$$Var_{N2}(\hat{\theta}) = \frac{N_1}{N_2} Var_{N1}(\hat{\theta}).$$

Thus, we expected the posterior standard deviation of sample size 5000 to be $(2/5)^{1/2}$ = 0.63 times that of the standard deviation when N=2000. Our results appear consistent with the large sample theory.

Is it possible that AR(1) model is the winner of MSE for all the five fixed effect parameters? We think the answer is YES when the sample size N is large. However we are unable to compute the sample size analytically. Recall that,

$$\text{MSE} = (bias)^2 + variance = [E(\hat{\theta}_n) - \theta]^2 + Var(\hat{\theta}_n)$$

Under regularity conditions and when the model is correctly specified,

$$E(\hat{\theta}_n) = \theta + \frac{b_1}{n} + o(1/n)$$

$$Var(\hat{\theta}_n) = \frac{v_1}{n} + o(1/n)$$

where b_1 and v_1 are some functions of the parameters

If the model is misspecified,

$$\mathrm{E}\left[\hat{\theta}_n^{mis}\right] = \theta + g(\theta) + \frac{b_2}{n} + o(1/n)$$

$$Var\left[\hat{\theta}_n^{mis}\right] = v_{mis} + o(1/n)$$

where $g(\theta)$ is effect of misspecification: $g(\theta) \neq 0$

We don't know whether $v_{mis} < v_1$,

Then $MSE\left(\hat{\theta}_n\right) \cong \left(\frac{b_1}{n}\right)^2 + \frac{v_1}{n}$

$$MSE\left(\hat{\theta}_n^{mis}\right) \cong \left(g(\theta) + \frac{b_2}{n}\right)^2 + \frac{v_{mis}}{n}$$

As $n \to \infty$, $MSE\left(\hat{\theta}_n^{mis}\right) - MSE\left(\hat{\theta}_n\right) \cong g(\theta)^2 + \frac{2b_2 g(\theta)}{n} + \frac{v_{mis}}{n} - \frac{b_1}{n} - \frac{v_1}{n} \to g(\theta)^2$

Table 5.1 Estimates of Fixed Effect of ZIP, MIXED, and AR(1) Models

Node	True value	ZIP (*N*=2000)			MIXED (*N*=2000)			AR(1) (*N*=2000)			
		mean	sd	mse	mean	sd	mse	mean	sd	mse	mse winner
β_0	.00	.32	.07	.11	.07	.11	.02	-.07	.18	.04	MIX
β_1	.50	.18	.04	.11	.27	.10	.06	.40	.15	.03	AR(1)
β_2	-3.00	-2.33	.34	.57	-2.08	.46	1.06	-2.57	.79	.81	ZIP
γ_0	.50	.92	.02	.18	.54	.03	. 00	.50	.06	. 00	MIX& AR(1)
γ_1	6.00	5.53	.10	.24	6.02	.11	.01	6.17	.20	.05	MIX
Node	True value				MIXED (*N*=5000)			AR(1) (*N*=5000)			
					mean	sd	mse	mean	sd	mse	mse winner
β_0	.00				0.089	0.08	0.014	-0.038	0.123	0.016	MIX
β_1	.50				0.356	0.071	0.026	0.548	0.102	0.013	AR(1)
β_2	-3.00				-2.132	0.313	0.852	-2.725	0.515	0.341	AR(1)
γ_0	.50				0.543	0.02	0.002	0.486	0.035	0.001	AR(1)
γ_1	6.00				5.953	0.072	0.007	6.097	0.128	0.026	MIX

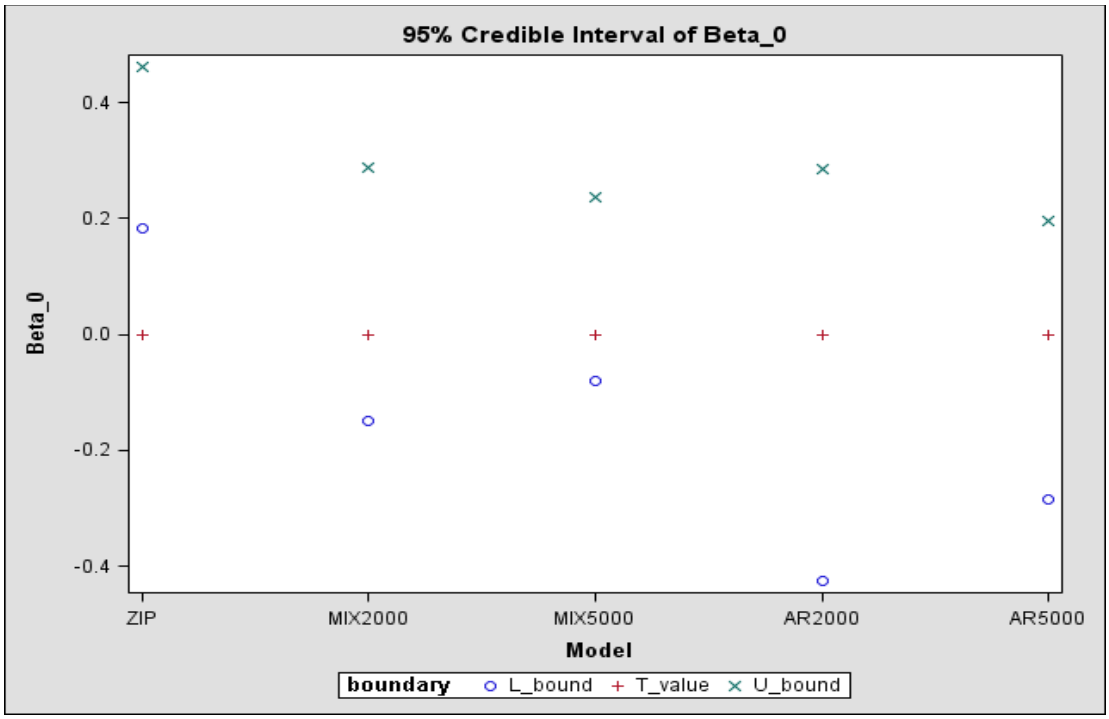


Figure 5.1 95% Credible Interval of Beta_0 Fixed Effects

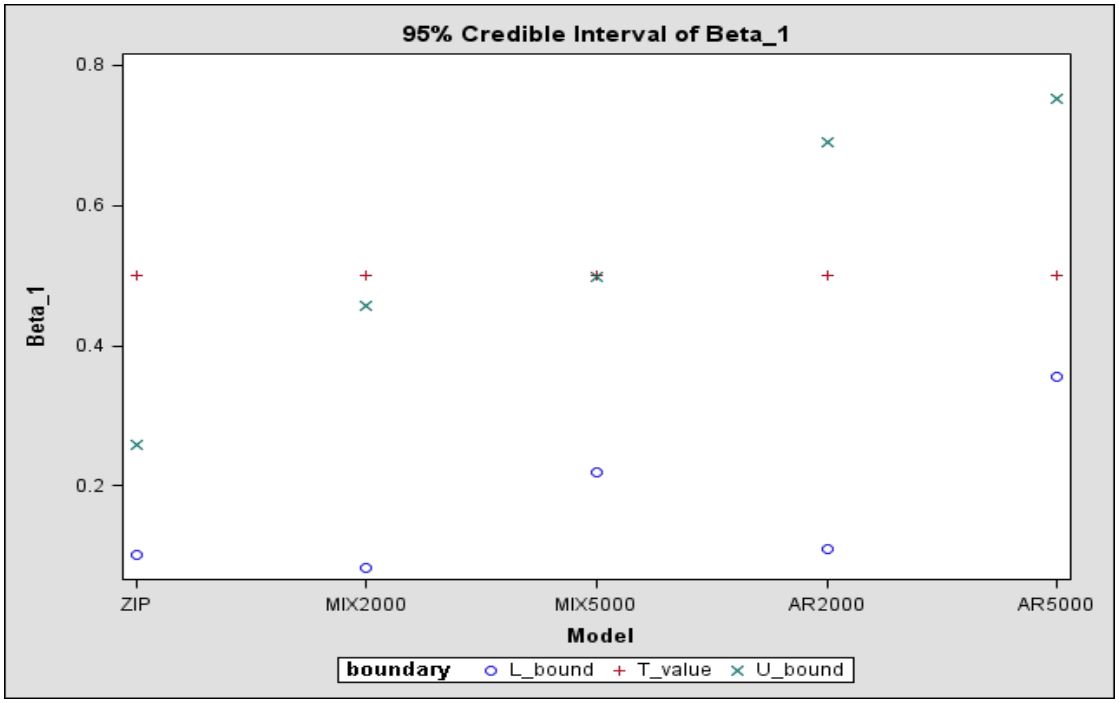


Figure 5.2 95% Credible Interval of Beta_1 Fixed Effects

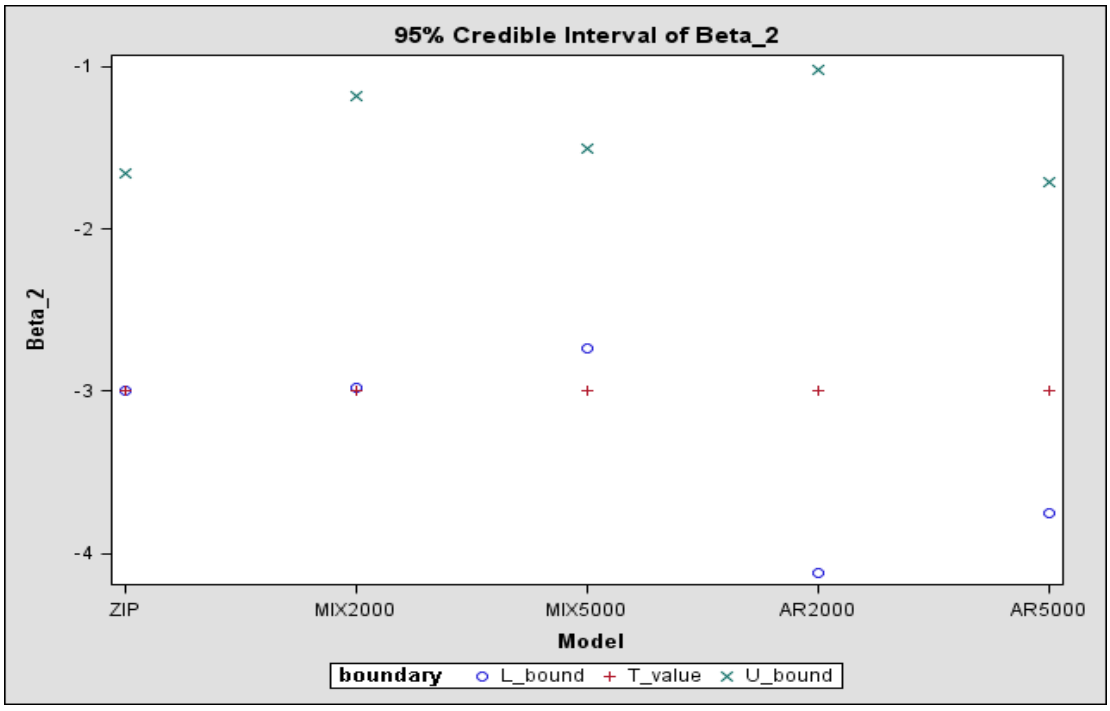


Figure 5.3 95% Credible Interval Beta_2 Fixed Effects

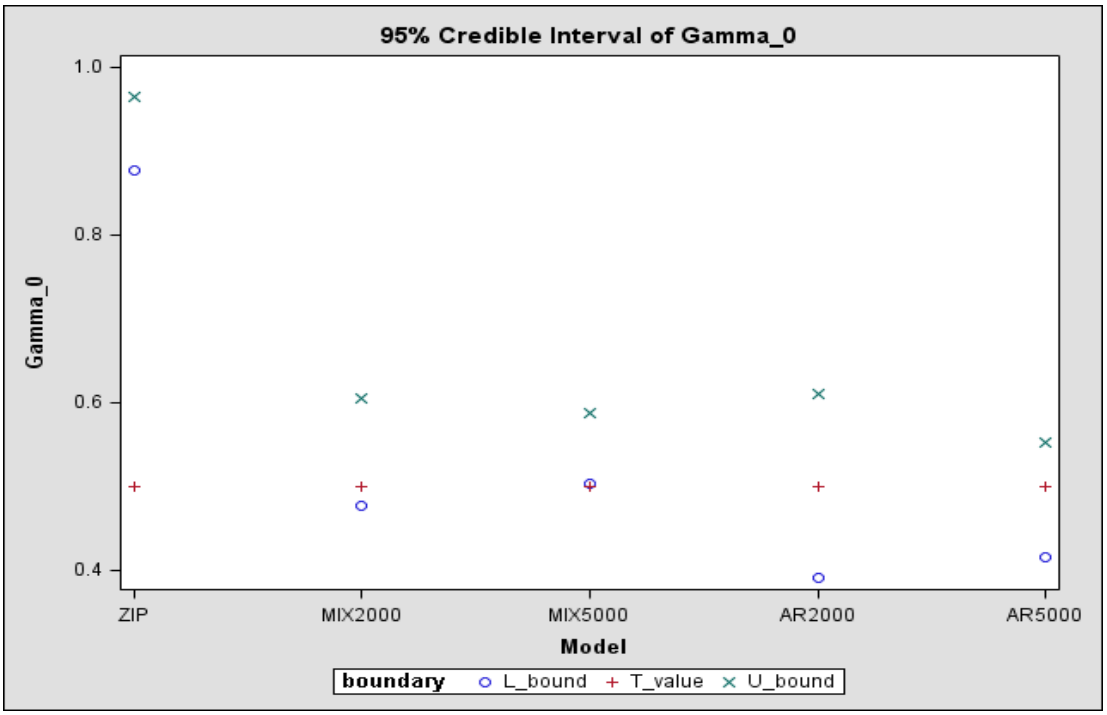


Figure 5.4 95% Credible Interval Gamma_0 Fixed Effects

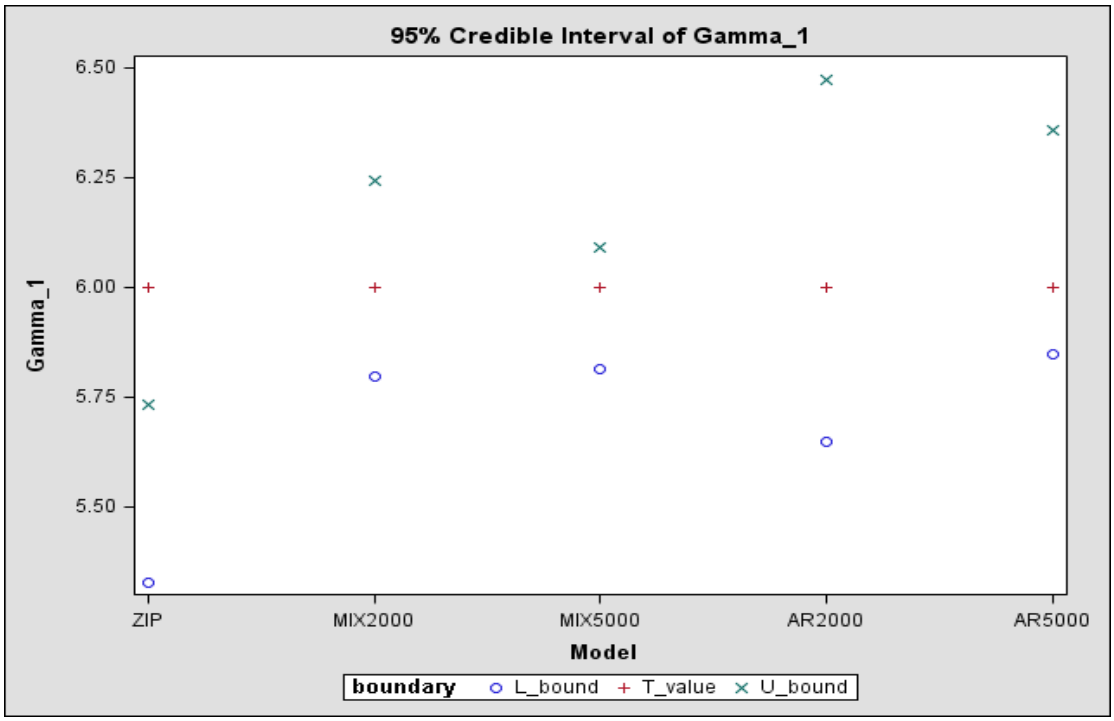


Figure 5.5 95% Credible Interval Gamma_1 Fixed Effects

5.2 NLSY97 Data Analysis

Imputation of Missing Values

We choose the NLSY97 data to illustrate the models. Note that the NLSY97 survey personnel do not, in general, impute missing values or perform internal consistency checks across waves. The following missing value conventions are used throughout the data: Noninterview, Valid Skip (respondent was not asked a question because it did not apply to him or her), Invalid Skip, Don't Know, and Refusal. Item nonresponse due to refusals, don't knows, or invalid skips is usually quite small, so the degree to which the weights are incorrect is probably quite small (NLSY, 1997). The weights are not considered here. Even though the sample size is almost 9000 subjects, we have a serious missing value problem. Only 600 self-reported smoking subjects

report cigarette consumption at all five observation times. Since the ZIP models require large sample size, an imputation is needed.

The following table is the NLSY97 coding for variables NUMBER OF CIGARETTES SMOKED 1997 and HAVE YOU EVER SMOKED 1997. The entries are frequencies of responses to each combination of values. Blanks represent zeros.

Table 5.2 NLSY97 Data Coding Example

		Have you ever smoked in 1997						
		D	I	R	N	V	0 (No)	1(Yes)
Number of cigarettes smoked in 1997	Don't Know (D)							6
	Invalid skip (I)							
	Refuse (R)							4
	Non interview (N)							
	Valid skip (V)	5	2	24			5436	1899
	Cigarettes smoked (Y)						0	1608
	Valid total							8943

Only the "Valid Skip" in "Number of Cigarettes Smoked" cell requires imputation. The other four types of missing values are deleted from the analysis. Records with missing values in the covariates, e.g., "Have You Ever Smoked," are also deleted from the analysis.

The 5436 subjects who responded "NO" to "Have you ever smoked" are non-smokers and their numbers of cigarettes smoked are simply imputed as $Y = 0$. The 1899 subjects who responded "YES" are smokers and we impute their numbers of cigarettes smoked as integers Y with Poisson distributions, given the covariates. Note that the Y values of these 3057 subjects (1899 + 1608) do not follow a standard Poisson distribution but an overdispersed Poisson. Since we don't know how to impute overdispersed Poisson data, the use of the Poisson distribution for Y will yield smaller variances compared to the original unobserved data.

Because the missing pattern is "Swiss cheese" (e.g., a subject may respond in year 1, refuse to respond in year 2, and then respond in year 3) and subjects do not always remain in the same class from year to year, we could not impute all five years at one time. Instead, we did the imputation year by year. As a result, our imputation yields smaller time correlations compared to original unobserved data.

For 1997 we used only the 1899 + 1608 = 3507 subjects who responded "YES" to "Have you ever smoked" to impute the missing values, using Poisson regression in WinBUGS. The covariates are: AGE (ordinal scale), SEX, PEER (ordinal scale), and RACE (nominal scale). If there were missing covariates, we deleted that case. The priors for the regression coefficients were normal with mean and variance estimated from observed data and initial values of the regression coefficients were drawn from their priors. No random effects were used in the imputation. BUGS imputed the missing counts Y_i as random draws from their predictive distribution. This analysis treats the missing data mechanism as ignorable, given the covariates (Gelman et al. 2003, pp. 517-518). For each of the later years we used exactly the same methods as described above. The imputations for each of the five years were performed independently.

From the descriptive statistics in Tables 5.3 ~ 5.5, the imputation has some degree of bias: We observe slightly smaller means, smaller standard deviations, and lower correlations. If we could provide information in the longitudinal zero inflated Poisson data setting--that is, AR(1) model imputation--the imputation might do a better job. However, this is beyond the scope of this research.

Table 5.3 Numbers of Cigarettes Smoked Per Day from Year 1997 to Year 2001

Before imputation					After imputation			
Variable	N	Mean	SD		Variable	N	Mean	SD
cig97	1608	5.0	6.0		cig_97	3485	4.6	4.5
cig98	2153	6.2	7.6		cig_98	2993	6.1	6.6
cig99	2418	7.3	8.3		cig_99	3026	7.2	7.6
cig00	2603	8.0	8.5		cig_00	3146	8.0	7.9
cig01	2749	8.5	8.8		cig_01	3226	8.4	8.3

Table 5.4 Correlations of Number of Cigarettes Smoked, from 1997 to 2001

Before imputation						After imputation					
Year	1997	1998	1999	2000	2001	Year	1997	1998	1999	2000	2001
1997	1.00	.45	.34	.35	.30	1997	1.00	.38	.30	.27	.26
1998		1.00	.51	.44	.35	1998		1.00	.47	.42	.31
1999			1.00	.55	.40	1999			1.00	.50	.38
2000				1.00	.51	2000				1.00	.47
2001					1.00	2001					1.00

Table 5.5 Numbers of Cigarettes Smoked Per Day in Year 1997

Before imputation		After imputation		Before imputation		After imputation	
cig97	Frequency	cig_97	Frequency	cig97	Frequency	cig_97	Frequency
0	146	0	232	16	8	16	8
1	450	1	609	17	8	17	9
2	232	2	531	18	7	18	7
3	139	3	467	19	1	19	1
4	94	4	370	20	68	20	67
5	95	5	301	21	0	21	0
6	48	6	233	22	1	22	1
7	30	7	146	23	1	23	1
8	32	8	127	24	5	24	5
9	8	9	71	25	7	25	7
10	124	10	152	26	1	26	1
11	5	11	18	27	0	27	0
12	16	12	26	28	0	28	0
13	5	13	14	29	0	29	0
14	2	14	5	30	14	30	14
15	61	15	62				

NLSY97 Data Analysis

The imputed NLSY97 data has N = 6932 complete cases, five time waves, and four covariates. The RACE variable was coded using three dummy variables, the indicators of "BLACK," "HISPANIC," and "MIXED." Now we have 19 parameters to estimate. The SEX variable was coded using one dummy variable, the indicator of SEX = "female". We explored the unimputed NLSY data using a GLMM Poisson model and GLMM logistic model to get some idea about the data. Then we used this knowledge to assign some of the priors (fixed effect, σ_u, and σ_v priors) and initial values (fixed effect, σ_u, and σ_v). We also used the knowledge from our previous simulation studies to assign the priors and some of the initial values. The remaining initial values were generated by BUGS from the prior distributions. All the prior distributions and model structure are the same as the model used in the simulation study. Only one chain is used in this analysis, but we still have some other tools to check convergence. The model runs smoothly but extremely slowly. Because of high autocorrelation, the iterations were thinned by a factor of 20. In each 24 hours OpenBUGS could only run 5000 iterations, which with the thinning factor produced 250 saved values in the chains. The AR(1) model ran for more than 15 days, performing 250 X 20 X 15 = 75000 (iterations) with a burn-in of 60000 iterations. The ratios (MC_error/Sd) are around 16%. Ideally, we would like these ratios to be no more than 1%.

Table 5.6 displays the results. Based on the results, we highlight important random effect parameters in yellow and the most important covariate parameters in green in terms of statistical significance. The random effect parameters, φ, ψ, σ_u, σ_v, and ρ_{uv} are overwhelmingly significant. We conclude that the NLSY data are strongly influenced

by the random effects, which are highly correlated. Evidently a simpler random effect structure would not adequately fit these data, so the ZIP model may produce serious bias for this NLSY data. However, MIXED model might produce milder bias than we saw in the simulation study of Chapter 5.1, in which $\varphi = .85$ and $\psi = .85$. The MIXED model will cause negligible bias and smaller variance when $\varphi = 1$ and $\psi = 1$.

Most of the covariates have a significant effect in the model. Based on the logistic regression, which estimates the probability that a youth is a nonsmoker, Hispanic and mixed race youths are more likely than whites to smoke because their coefficients have negative signs; the older youths are more likely than younger youths to be smokers; the youth having more peers who are smokers is more likely be a smoker. Based on the Poisson regression, which estimates the numbers of cigarettes that a youth smoked, black smokers smoke more cigarettes than white smokers; Hispanic and mixed smokers smoke fewer cigarettes than white smokers; the older youths are more likely than younger youths to smoke more cigarettes; the youth having more peers as smokers are more likely to smoke more cigarettes.

Table 5.6 NLSY 1997 ~ 2001 Data Analysis Results (AR(1) Model)

Logistic	Mean	Sd	MC_error	Poisson	Mean	Sd	MC_error
beta	5.282	.652	.115	gamma	-.514	.063	.011
age	-.251	.021	.004	age_p	.041	.002	.000
peer	-.438	.057	.008	peer_p	.248	.015	.003
sex	.912	.209	.028	sex_p	-.056	.042	.007
black	.860	.502	.088	black_p	1.949	.043	.007
hisp	-2.572	.219	.032	hisp_p	-.741	.059	.011
mixed	-12.58	.236	.029	mixed_p	-3.397	.030	.005
φ	.949	.000	.000	ψ	.868	.000	.000
σ_u	7.223	.054	.009	σ_v	3.635	.006	.001
ρ_{uv}	.886	.002	.000				

Since we don't know the true parameters of the NLSY97 data, we would like to use the deviance information criterion (DIC) in BUGS to make model fit comparisons. However the DIC assumes that the posterior mean can be used as a "good" summary of central location for description of the posterior distribution (Ntzoufras, 2009, pp 140-141). Since we did not run the model long enough to meet the model convergence assumption, we did not use DIC for these analyses. We present the other two model results in the following.

Table 5.7 NLSY 1997 ~ 2001 Data Analysis Results (MIXED Model)

Logistic	Mean	Sd	MC_error	Poisson	Mean	Sd	MC_error
beta	4.027	2.174	0.339	gamma	-0.909	0.338	0.052
age	-0.232	0.031	0.003	age_p	0.035	0.010	0.001
peer	-0.532	0.046	0.006	peer_p	0.199	0.022	0.003
sex	0.155	0.069	0.004	sex_p	-0.244	0.031	0.003
black	2.708	2.236	0.351	black_p	0.873	0.359	0.055
hisp	-0.820	0.122	0.009	hisp_p	0.114	0.057	0.006
mixed	-0.754	0.108	0.009	mixed_p	0.540	0.032	0.001
σ_u	2.464	0.255	0.038	σ_v	0.706	0.089	0.013
ρ_{uv}	-0.472	0.081	0.012				

Table 5.8 NLSY 1997 ~ 2001 Data Analysis Results (ZIP Model)

Logistic	Mean	Sd	MC_error	Poisson	Mean	Sd	MC_error
beta	3.578	.038	.008	gamma	-.609	.018	.003
age	-.116	.008	.002	age_p	.039	.001	.000
peer	-.308	.009	.001	peer_p	.144	.003	.000
sex	.103	.023	.001	sex_p	-.204	.007	.000
black	.108	.147	.032	black_p	1.299	.017	.004
hisp	-.472	.037	.004	hisp_p	.049	.013	.002
mixed	-.491	.030	.003	mixed_p	.427	.010	.001

These three models all have ratio (MC_error / SD) around 16%. Some parameters are not quite convergent yet; while some are not mixing well. Running a longer time might be helpful.

Even the ZIP model needs an extra one or two weeks of running time. Based on observing the trace plots and hist plots, we found that the estimation of Poisson regression is much worse than logistic regression in terms of the stability of the Markov chain. This is opposite to the results we saw in the simulation study. A possible reason is the bumps of numbers of cigarettes smoked, that the frequency of 5, 10, 15, or 20 cigarettes smoked are especially higher than other numbers. As for the estimation of variances, the ZIP model has smallest variances and the AR(1) model has largest variances. This finding agrees with our simulation study.

Although none of three models is the true model, we think AR(1) is closer to the truth. None of the parameter estimates are close together values in the three models. Some of parameters have the same sign and the magnitudes of the values occur in decreasing order of AR(1), MIX, and ZIP. For example, the estimated means of $\hat{\beta}$ are 5.282 (AR(1)), 4.027 (MIX), and 3.578 (ZIP). However, some parameters do even not meet this level of agreement (highlighted in Yellow). We also found that more than half of the MIX model estimates (highlighted in red) have big discrepancies from those of the AR(1) model. The prior in MIX model is one possible reason.

Table 5.9 Comparison of AR(1), MIX, and ZIP Model Estimates of Means

LOGISTIC	AR(1)	MIX	ZIP	POISSON	AR(1)	MIX	ZIP
beta	5.282	4.027	3.578	gamma	-.514	-0.909	-.609
age	-.251	-0.232	-.116	age_p	.041	0.035	.039
peer	-.438	-0.532	-.308	peer_p	.248	0.199	.144
sex	.912	0.155	.103	sex_p	-.056	-0.244	-.204
black	.860	2.708	.108	black_p	1.949	0.873	1.299
hisp	-2.572	-0.820	-.472	hisp_p	-.741	0.114	.049
mixed	-12.58	-0.754	-.491	mixed_p	-3.397	0.540	.427
σ_u	7.223	2.464		σ_v	3.635	0.706	
ρ_{uv}	.886	-0.472					

Chapter VI

Conclusions and Future Research

6.1 Summary

In this study, we developed the AR(1) model to handle longitudinal zero inflated Poisson data. We conducted simulation studies of the model and fitted the model to a real world data set, the National Longitudinal Survey of Youth. We used the theory of maximum likelihood estimation and Fisher information to develop estimates of the necessary sample size for our simulations. We compared the performance of the AR(1) model to simpler models to assess bias and sampling error.

For the simulated AR(1) data, which had features of the NLSY97 data, the ZIP model and MIXED model produced seriously biased estimates of the fixed effect parameters. However, these simpler models had less sampling variability than the AR(1) model, because fewer parameters had to be estimated.

The AR(1) model requires a large sample size to obtain adequate accuracy. Also, the large sample size, complicated model structure, and slowness of BUGS software require a tremendous length of time to analyze data using MCMC methods. The AR(1) model as implemented in BUGS, imposes a heavy computational burden.

6.2 Topics for Future Research

The AR(1) ZIP model entails a heavy computational burden. A more sophisticated MCMC algorithm might potentially reduce the computational requirements and make the model more useful to practitioners. Such algorithms might not be possible in BUGS and may require programming in more flexible languages, such as R or C++.

Maximum likelihood estimation does not seem like a practical way to analyze the model because of the need to compute high dimensional integrals for each subject in order to evaluate the likelihood.

It is possible that a Generalized Estimating Equation (GEE) approach might be devised. The GEE approach requires one to model the mean response accurately, but a correct model of the variance-covariance structure is not required (Agresti, 2002, Ch. 11). However, the mean response still involves integrals that can't be evaluated in closed form. Therefore GEE may also lead to computational difficulties.

Fitting the GLMM model instead of the AR(1) model is another alternative method. Our simulations suggest that estimates of regression coefficients may be biased under this model. One can not be sure whether the smaller standard errors and computational savings would offset the bias in estimating regression coefficients.

References

Agresti, A. (2002) *Categorical Data Analysis* 2nd ed. New York: John Wiley & Sons.

Berger, J. O. (1984). The Robust Bayesian Viewpoint. *In Robustness of Bayesian Analysis*, Joseph B. Kandane (Ed.) Amsterdam: North Holland, 63 – 144.

Best, N., Cowles, M. & Vines, K. (1996). *CODA: Convergence Diagnostics and Output Analysis Software for Gibbs Sampling Output, Version 0.30*, MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

Bickel, P. & Doksum, K. A. (2007). *Mathematical Statistics: Basic Ideas and Selected Topics.* New Jersey: Pearson Prentice Hall.

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, ***7*(4)**, 434-455.

Carlin, B. & Louis, T. (2000). *Bayesian and Empirical Bayes Methods for Data Analysis.* NY: Chapman & Hall/CRC.

Carriquiry, A. & Pawlovich, M. (2004). From Empirical Bayes to Full Bayes: Methods for Analyzing Traffic Safety Data. http://publications.iowa.gov/13274/1/eb_fb_comparison_whitepaper_october2004.pdf

Chao, M. T. (1970). The Asymptotic Behavior of Bayes' Estimators. *Annals of Mathematical Statistics* ***41****, 601 – 608.*

Chatfield, C. (2009). *The Analysis of Time Series: An Introduction* (6th *ed*.). NY: Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis*. Second Edition. New York: Chapman & Hall.

Gelman, A., & Rubin, D. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science.* **7**, 457-511.

Ghosh, S.K., Mukhopadhyay P, & Lu, JC. (2006). Bayesian Analysis of Zero-Inflated Regression Models. *Journal of Statistical Planning and Inference*, **136**, 1360–75.

Gilk, W., & Roberts, G. (1996). Strategies for Improving MCMC, in *Markov Chain Monte Carlo in Practice.* Gilk, W. R., Richardson, S., & Spiegelhalter, D. (Eds.) UK: Chapman & Wall. 89 -110.

Gill, J. (2008). *Bayesian Methods: A Social and Behavioral Sciences Approach.* Second Edition. FL: Chapman & Hall/CRC.

Hall, D. B. (2000). Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics.* **56,** 1030-1039.

Hedeker, D., & Gibbons, R. (2006). *Longitudinal Data Analysis*. New Jersey: Wiley.

Heibron, D. (1994). Zero-altered and Other Regression Models for Count Data with Added Zeros. *Biometrical Journal*, **36**, 531-547.

Jansen, M. G. H. (1986). A Bayesian Version of Rasch's Multiplicative Poisson Model for the Number of Errors of an Achievement Test. *Journal of Educational Statistics,* **11, 2**, 147-160.

Lambert, D. (1992). Zero-Inflated Poisson Regression Models with an Application to Defects in Manufacturing. *Technometrics*, **34**, 1-14.

Lindley, D. V. (1969). Discussion of Compound Decisions and Empirical Bayes, J. B. Copas. *Journal of the Royal Statistical Society, Series B* **31**, 397-425.

Lynch, S. M. (2007) *Introduction to Applied Bayesian Statistics and Estimation for*

Social Scientists. NY: Springer.

Miaou, S.-P. (1994). The Relationship Between Truck Accidents and Geometric Design of Road Sections. Poisson versus Negative Binomial Regressions. *Accident Analysis & Prevention,* **26**, 471-482.

Min, Y., & Agresti, A. (2005). Random Effect Models for Repeated Measures of Zero-Inflated Count Data. *Statistical Modelling*, **5**, 1–19.

Mullahy, J. (1986) Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics,* **33**, 341–65.

Neelon, B. H, O'Malley, A. J., & Normand, S. T. (2010). A Bayesian Model for Repeated Measures Zero-Inflated Count Data with Application to Outpatient Psychiatric Service Use. *Stat Modelling,* **10(4)**, 421–439.

Nelder, J. A., & Wedderburn, R. W. M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, **135**, 370-384.

NLSY97. (1997). National Longitudinal Survey of Youth 1997. http://www.bls.gov/nls/nlsy97.htm

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS.* USA: Wiley.

R Development Core Team (2008), *R: A language and environment for statistical Computing*. Vienna: R Foundation for Stat. Comp., http://WWW.R-project.org.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.

Ridout, M., Demetrio, C. G. B., & Hide, J. (1998). Models for Count Data with Many Zeros. Invited paper presented at the Nineteenth International Biometric Conference, Capetown, South Africa.

Roberts, G. (1996). Markov Chain Concepts Related to Sampling Algorithms. In Gilks, W., Richardson, S., & Spiegelhalter, D. (Eds.) *Markov Chain Monte Carlo in Practice.* Uk: Chapman & Hall. pp. 45-58.

Shao, J. (2003). *Mathematical Statistics.* NY: Springer.

Smith, B. (2005). *(B)ayseian (O)utput (A)nalysis Program (BOA) Version 1.1.5 User's Manual*, Technical Report, Department of Health. The University of Iowa, available at http://www.public-health.uiowa.edu/boa

Spiegelhalter, D. Thomas, A. Best, N. & Lunn, D. (2003) *WinBUGS User Manual, Version 1.4*. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK available at http://www.mrc-bsu.cam.ac.uk/bugs.

Sun, D., Speckman, P. L., & Tsutakawa, R. K. (2000). Random Effects in Generalized Linear Mixed Models. *Generalized Linear Models: A Bayesian Perspective.* New York: Marcel Dekker.

Weng, C. F. (2008). *Fixed versus Mixed Parameterization in Logistic Regression Models: Application to Meta-Analysis*. Unpublished Master Thesis. Department of Mathematics at University of Maryland, College Park.