IGNORABILTY CONDITIONS FOR INCOMPLETE DATA AND THE FIRST-ORDER MARKOV CONDITIONAL LINEAR EXPECTATION APPROACH FOR ANALYSIS OF LONGITUDINAL DISCRETE DATA WITH OVERDISPERSION

Shaun Bender

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

_____

Justine Shults

Professor of Biostatistics

Graduate Group Chairperson

_____

John H. Holmes, Professor of Medical Informatics in Epidemiology

Dissertation Committee

Warren Bilker, Professor of Biostatistics

Dawei Xie, Asst. Professor of Biostatistics

Yimei Li, Asst. Professor of Biostatistics

Peter Reese, Asst. Professor of Medicine (Renal-Electrolyte and Hypertension) & Epidemiology

IGNORABILTY CONDITIONS FOR INCOMPLETE DATA AND THE FIRST-ORDER MARKOV CONDITIONAL LINEAR EXPECTATION APPROACH FOR ANALYSIS OF LONGITUDINAL DISCRETE DATA WITH OVERDISPERSION

ACKNOWLEDGEMENT

I would like to thank my dissertation supervisor Professor Justine Shults; my committee chair Professor Warren Bilker; my committee members Assistant Professor Dawei Xie, Assistant Professor Yimei Li, and Assistant Professor Peter Reese; Professor Daniel Heitjan; the faculty and staff at the University of Pennsylvania Department of Biostatistics & Epidemiology; and all my friends and family for supporting me.

ABSTRACT

IGNORABILTY CONDITIONS FOR INCOMPLETE DATA AND THE FIRST-ORDER MARKOV CONDITIONAL LINEAR EXPECTATION APPROACH FOR ANALYSIS OF LONGITUDINAL DISCRETE DATA WITH OVERDISPERSION

Shaun Bender

Justine Shults

Medical researchers strive to collect complete information, but most studies will have some degree of missing data. We first study the situations in which we can relax well accepted conditions under which inferences that ignore missing data are valid. We partition a set of data into outcome, conditioning, and latent variables, all of which potentially affect the probability of a missing response. We describe sufficient conditions under which a complete-case estimate of the conditional cumulative distribution function of the outcome given the conditioning variable is unbiased. We use simulations on a renal transplant data set to illustrate the implications of these results. After describing when missing data can be ignored, we provide a likelihood based statistical approach that accounts for missing data in longitudinal studies, by fitting correlation structures that are plausible for measurements that may be unbalanced and unequally spaced in time. Our approach can be viewed as an extension of generalized linear models for longitudinal data that is in contrast to the generalized estimating equation approach that is semi-parametric. Key assumptions of our method include first-order ante-dependence within subjects; independence between subjects; exponential family distributions for the first outcome on each subject and for the subsequent conditional distributions; and linearity of the expectations of the conditional distributions. Our approach is appropriate for data with over-dispersion, which occurs when the variance is inflated relative to the assumed distribution. We consider a clinical trial to compare two treatments for seizures in patients using Poisson or Negative Binomial distributions. Next, we consider a study that evaluates the likelihood that a transplant center is flagged for poor performance using the Binomial distribution. For both studies, we perform simulations to assess the properties of our estimators and to compare our approach with GEE. We demonstrate that our method outperforms GEE, especially as the degree of over-dispersion increases. We also provide software in R so that the interested reader can implement our method in his or her own analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1. Overview of the Thesis

In this dissertation we discuss two related topics. The first concerns the ignorability conditions for frequentist nonparametric analysis of conditional distributions with incomplete data. Missing data is common in any sort of scientific investigation, although medical researchers work hard to obtain complete information. If missing observations occur completely at random throughout a data set, semi-parametric approaches such as generalized estimating equations (GEE) (Liang and Zeger, 1986) will be valid. However, analysis may be unbiased for when the data is not missing completely at random. We will explore a situation when missing observations can lead to biased results in Chapter 2. We will also explore conditions under which the reasons for missing data can be ignored.

The second topic we consider is the development of the first-order Markov conditional linear expectation approach that, like GEE, extends generalized linear models (GLM) to longitudinal data. However, unlike GEE, our proposed approach is likelihood based and is appropriate for longitudinal data with over-dispersion. The proposed method can also be used to address the problem of missing data that occur when participants in longitudinal studies drop out and do not return (drop out), or miss some visits but then return for their final measurement (intermittent missingness). Drop out and intermittently missing values result in a variable number of measurements per subject (unbalanced data) and in measurements that are unequally spaced in time. The proposed method addresses both situations, by allowing for the implementation of correlation structures that are appropriate for data that are unbalanced and unequally spaced in time. We present this approach for count data in Chapter 3 and for binomial type data in Chapter 4.

This thesis is organized as follows. This introductory Chapter provides a brief and general introduction, while Chapters 2,3, and 4 each represent distinct manuscripts. Finally, Chapter 5 summarizes our findings and discusses possible future research directions.

1.2. Ignorability Conditions for Incomplete Data

Protocols for longitudinal studies typically specify the number of measurements that will be collected on each of a pre-specified number of subjects. However, some measurements will invariably be missing from the final data set. Participants may drop out of the study, or may miss some visits. Or, some information may be lost, for example, if some samples in a laboratory are accidentally destroyed, or if questionnaires are lost.

In order to understand whether missing data will result in biased analytic results, it is important to assess the missing data mechanism, which is the random process that results in missing values in the final analytic dataset. Rubin (1976) provided general and weak conditions under which the missing data mechanism can be ignored and the analysis results will still be correct. The weakest general condition for frequentist ignorability of the missing data mechanism is when the missing data mechanism is such that any missing data are missing completely at random (MCAR). However, although the MCAR condition is sufficient for unbiased analysis, it is not always necessary. For example, in linear regression, if the missing data mechanism depends on the predictors of the outcome variables (so that the missing data mechanism depends on observed data and the MCAR assumption is violated), a complete case analysis that ignores subjects with any missing data will still yield results that are unbiased.

When the missing data mechanism depends on the predictors, the missing data are said to be missing at random (MAR). To be more precise, the missing data mechanism is MAR when the probability of observing the realized missing data pattern, given the missing and realized observed data, does not depend on the values of the missing data. If the MAR assumption is violated, the missing data mechanism is not missing at random (NMAR).

The conditions that are sufficient for correct analysis depend on the type of analysis that is being performed. In likelihood based analysis, if the missing data mechanism is MAR and the parameter spaces of the model and missing data mechanisms are distinct, then ignoring the missing data mechanism will lead to unbiased results. Under a Bayesian-based framework of analysis, if the data are MAR (along with the model parameters being *a priori* independent of the missing data mechanism parameters), then ignoring the missing data mechanism will lead to unbiased results. In general, ignoring the missing data mechanism will lead to unbiased results in a frequentist-based

approach if the data are MCAR.

In Chapter 2, we study the precise conditions under which frequentist nonparametric inference with missing data is correct for modeling conditional distributions, such as in linear regression. In this Chapter we rigorously describe the sufficient conditions for ignorability of the missing data mechanism in complete-case frequentist inference for a nonparametric model of conditional distributions when some outcomes may be missing. The conditions we describe may be weaker than MCAR. We illustrate these results using data from a renal transplant study ("Improvements in muscle mass and function following renal transplantation").

1.3. The First Order Markov Maximum Likelihood Based Approach for Count Data with Over-Dispersion

Count data are commonly encountered in medical research. For example, in Chapter 3 we consider a longitudinal study of repeated seizure counts on patients enrolled in a clinical trial Thall and Vail (1990). In general, we consider outcome variables that represent longitudinal counts that take value in $\{0, 1, 2, \cdots\}$ on subject i.

We also consider outcomes with over-dispersion, which is a common feature of count data that occurs when the variance of the outcome variable is inflated relative to the assumed distribution. Over-dispersion is easy to detect for an assumed Poisson distribution because the mean and variance of the outcome variable are equal for the Poisson distribution. Strong evidence of over-dispersion is provided when the sample variances are very large relative to the sample means.

We provide a likelihood based approach for analysis of over-dispersed longitudinal discrete data that is based on several assumptions. First, we assume that measurements on different subjects are independent, but measurements within subjects are correlated. We also assume first-order antedependence, which is also referred to as the first-order Markov property. Next, we assume that the distribution of the first outcome on each subjects, and the distribution of the subsequent conditional distributions, are members of the same exponential family. We also assume that the conditional expectations of the conditional distributions are linear.

The assumptions of linearity (of the conditional means) and of first-order antedependence induce decaying product structures that are plausible for longitudinal data (Guerra and Shults, 2014). The

decaying product structures include the AR(1) structure that is appropriate for equally spaced data; the Markov structure that takes unequal temporal spacing of measurements into account; and the first-order AD(1) structure that allows the correlation of adjacent measurements on subjects to vary over the course of the study. The Markov structure includes the AR(1) structure as a special case, when the measurements on a subject are equally spaced in time. The AD(1) structure includes the AR(1) structure as a special case, when the adjacent correlations on a subject are equal.

Our assumption of exponential distributions for the first observation and subsequent conditional distributions suggests that our method could be viewed as an approach that extends generalized linear models (GLM) (McCullagh and Nelder, 1989) to longitudinal data with over-dispersion. We therefore compared our approach with GEE, which extended the likelihood equations for exponential families to longitudinal data via the incorporation of working correlation structures that described the pattern of intra-subject association of measurements. However, while our approach is likelihood based, GEE is semi-parametric and only requires specification of the first and second moments of the distribution of the outcome variable. In addition, GEE ignores the over-dispersion that is commonly encountered in longitudinal discrete data.

Our comparison demonstrates the advantages of a likelihood based approach relative to a semi-parametric approach like GEE. Because likelihood based approaches are based on maximizing an objective function (the log-likelihood) we can more easily assess the goodness of fit of our models, via criteria that are based on the estimated log-likelihood. The criteria we consider include the Akaike's Information and Bayesian Information Criterion . We are also able to perform likelihood ratio tests that are useful for choosing between nested models. For analysis of the seizure data, the likelihood ratio test is also useful to justify the application of the negative Binomial versus the Poisson distributions.

We also perform simulations that indicate that the mean-square error and bias of our estimators decrease as the sample size increases. In addition, the estimated coverage probability of our estimators is appropriate, and approaches nominal levels as the sample size increases. Our simulations also demonstrate that our method outperforms GEE, with increasing improvement in performance as the degree of over-dispersion increases. GEE also requires the assumption that any missing data are MCAR, while our approach only requires the less restrictive assumption that the missing data are MAR. We also use simulation to assess the Likelihood Ratio Test for comparison of Pois-

son and Negative-Binomial models, in which we must use the results of Chernoff (1954) instead of the standard Likelihood Ratio Test.

1.4. The First Order Markov Maximum Likelihood Based Approach for Analysis of Binomial Type Variables

Chapter 4 considers Binomial type outcomes that take value in $\{0, 1, \cdots, n_i\}$ on subject i. Our motivating study for this Chapter is of the relationship between the U.S. News & World Report ranking of a hospital and the number of times the hospital was flagged for having poor performance with respect to organ transplant.

Transplant programs are flagged for poor performance based on information in the Scientific Registry of Transplant Recipients (SRTR), which is a database of organ transplantation statistics in the United States. Every six months, the SRTR releases publicly available transplant program reports for each transplant center that include information on waiting time, organ availability, and survival statistics. In addition, the reports include the number of observed and expected graft failures for each center during the first year after transplant. If the number of observed graft failures is large, then the Centers for Medicare & Medicaid Services (CMS) will flag a program for poor performance. (More details regarding the criteria are provided in Chapter 4.)

We consider transplant program reports (excluding pediatric and Veteran's hospitals) for kidney, lung, liver, and heart transplant programs during the years 2012-2015. Two reports are released each year, so that the maximum number of times that a transplant program can be flagged during a year is 8. However, not all transplant programs provide transplants for all organs, so that the number of times a program is flagged is a binomial type outcome with $n_i \in \{1, \cdots, 8\}$.

Our analysis goal is to relate the number of times a treatment program is flagged with the ranking of the program according to the U.S. News & World Report introduced "America's Best Hospitals" ranking system (Olmsted et al., 2015). Each year, the U.S. News & World Report provides rankings on 16 different adult specialties, 12 of which are based on the Donabedian model of health care: structure, process, and outcomes (Donabedian, 1966). The components of the score are then used to create a weighted score for each specialty at each hospital. We consider 1,897 hospitals that were eligible for at least 1 of the 12 score-driven specialties under the U.S. News & World Report

criteria.

For our analysis, we use a binary variable to indicate whether a hospital received a high enough score to earn a place in the U.S. News & World Honor Roll. Preliminary descriptive analysis suggested that transplant programs that were affiliated with hospitals that were not in the U.S. News & World Report Honor Roll had a higher number of occurrences of being flagged in the previous year than hospitals that were on the honor roll. A preliminary analysis also suggested that there was over-dispersion relative to the binomial distribution in the number of times a hospital was flagged. An appropriate analysis of data from this study should therefore account for the over-dispersion of the outcome variable, in addition to the intra-subject correlation within centers across the 4 years of follow-up.

As in the previous Chapter, we assume first-order antedependence, exponential distributions, and linearity of the conditional expectations. However, we now assume Binomial distributions for the first distribution and the subsequent conditional distributions. We obtain the form of the likelihood equations for the binomial distribution and fit models to relate being on the U.S. News & World Report Honor Roll with being flagged for poor performance. We also fit models with GEE, to make comparisons with the likelihood equations and developed software in R to implement our approach. We again make comparisons with GEE, both in the analysis and in simulations. Our simulations indicate that our approach outperforms GEE, with increasing superiority in relative performance as the degree of over-dispersion and the degree of violation of the MCAR assumption (in favor of an MAR assumption) increases. We perform our analyses in R and provide software so that interested readers can use our approach in their own analyses. We also use simulations to assess how well our Likelihood based analysis does in comparison to GEE methodology with data that are missing at random. We vary the portion of data that are missing and random and assess bias of each parameter.

CHAPTER 2

IGNORABILITY CONDITIONS FOR FREQUENTIST NONPARAMETRIC ANALYSIS OF CONDITIONAL DISTRIBUTIONS WITH INCOMPLETE DATA

2.1. Introduction

Missing data are common in scientific investigations of all kinds, and it has long been understood that randomness in the missing data mechanism (MDM) can induce bias in estimation. Rubin (1976) derived general conditions under which inference ignoring the MDM is correct. Rubin's conditions are, in his words, "the weakest general conditions under which ignoring the process that causes missing data always leads to correct inferences". That is, they represent general *sufficient* conditions for ignorability, but they may not be *necessary* in all contexts. Yet many subsequent papers assert that ignoring the MDM "requires" one or more of the conditions, when this is clearly not the case. For example, it is known that when conducting a linear regression of a variable Y on another variable X, one can select the points to include on the basis of their X values without biasing estimation of the regression coefficients. Data generated under such a selection method are not missing completely at random (MCAR), which Rubin presents (although not by that name) as the weakest general condition for frequentist ignorability; see Rubin (1976), Little and Rubin (2002). In this Chapter we study precise conditions under which frequentist nonparametric inference with missing data is correct for modeling conditional distributions.

2.2. Background

Let Y be a matrix of notional complete data whose (i, j) element is $Y_i^{(j)}$. Let R be the corresponding matrix of observation indicators, with $R_i^{(j)} = 1$ when Y_{ij} is observed and 0 when Y_{ij} is missing. Let y (r) be a realization of Y (R). The general selection model describes the MDM as the conditional distribution of R given Y, indexed by a parameter ψ:

$$\Pr_\psi(R = r|Y = y).$$

Rubin (1976) identified three types of MDM:

- Missing completely at random (MCAR, originally missing at random plus observed at random) holds when, for the given r and for every ψ, for all y^* and y^{**},

$$\Pr_\psi(R = r|Y = y^*) = \Pr_\psi(R = r|Y = y^{**}).$$

  That is, the probability of observing the realized missing data pattern r given Y does not depend on the value of Y.

- Missing at random (MAR). Let $o(y, r)$ denote the portion of Y consisting of elements whose corresponding elements of R equal 1 — i.e., the observed data. Then MAR holds when, for the given r and for every ψ, for all y^* such that $o(y^*, r) = o(y, r)$,

$$\Pr_\psi(R = r|Y = y) = \Pr_\psi(R = r|Y = y^*).$$

  That is, the probability of observing the realized missing data pattern r, given the missing data and the realized observed data $o(y, r)$, does not depend on the values of the missing data.

- Not MAR (NMAR) holds when the data are not MAR.

We emphasize that these definitions pertain only to the realized value $R = r$, not to all R; failure to recognize this distinction has led to much confusion in the literature (Seaman et al., 2013).

Sufficient conditions for correct inference vary depending on the mode of inference. In likelihood inference one specifies a model for Y indexed by a set of parameters θ. Within this framework, if the data are MAR and θ and ψ are distinct then ignoring the MDM does not impair inferences. The parameters are distinct if the joint parameter space is a Cartesian product of the individual parameter spaces. Similarly, if the data are MAR and θ and ψ are *a priori* independent, then one can safely ignore the MDM in Bayesian inferences.

The frequentist approach bases inference on the conditional distribution of Y given $R = r$ in repeated sampling. Within this framework, if the data are MCAR then the MDM is ignorable in the sense that the conditional sampling distribution of Y given $R = r$ and ignoring the MDM is equal to the correct conditional sampling distribution that accounts for the MDM.

Authors commonly cite MCAR as a sufficient condition for validity of a complete-case regression analysis with missing outcomes (Horton and Kleinman, 2007; Simonoff, 1988). Though this is true, such an analysis is generally valid under weaker assumptions. For example, simulations in Little (1992) show that complete-case linear regression estimates are unbiased if the MDM depends only on the predictor variables. Little and Rubin (2002) later discussed this in their Example 3.3. In an iid model, Galati and Seaton (2013) defined *available at random (AAR)* to mean that the conditional probability that a case is complete, given the Y values for that case, does not depend on Y. AAR is sufficient for the complete cases to be considered a random sample from the data, and therefore, like MCAR, justifies frequentist complete-case analyses. They moreover showed that AAR may hold under any of the conditions MCAR, MAR, or MNAR. This is slightly different from the results of Little and Rubin (2002) (and the results presented here) in that the MDM cannot depend on the regression covariates.

Robins, Rotnitzky, and Zhao (1994) examined general regression models of the form $\mathrm{E}(Y) = h(X, \beta)$, where β is the unknown parameter vector and h is a known function. They partitioned the regressors into $X = (W, V)$, where W may be missing and V is fully observed. They stated that if the probability of W being observed depends only on the vector V, then complete-case estimates of β are consistent. If the probability depends on both Y and V then the complete-case estimator may be biased. The first condition is MAR and the second condition is either MAR or NMAR (depending on whether Y can be missing).

In this Chapter we give a rigorous explication of sufficient conditions for ignorability of the MDM in complete-case frequentist inference for a nonparametric model of conditional distributions when some outcomes may be missing. Our conditions are substantially weaker than MAR. We illustrate the results using data from a renal transplant study ("Improvements in muscle mass and function following renal transplantation").

2.3. The Model and Ignorability Conditions

2.3.1. Assumptions

Suppose we sample N subjects from an infinite population and measure L discrete variables of interest. Let $Y_i^{(l)}$ be the value of the lth variable for the ith unit, and $Y_i = (Y_i^{(1)}, \cdots, Y_i^{(L)})$ be the

vector of observations for the ith unit. Let $R_i^{(l)} = 1$ if the lth variable for the ith unit is observed (0 if unobserved), and $R_i = (R_i^{(1)}, \cdots, R_i^{(L)})$ be the observation indicator vector for the ith unit. Thus we have the following notional data matrices:

$$Y = \begin{pmatrix} Y_1^{(1)} & \cdots & Y_1^{(L)} \\ \vdots & \ddots & \vdots \\ Y_N^{(1)} & \cdots & Y_N^{(L)} \end{pmatrix} \qquad R = \begin{pmatrix} R_1^{(1)} & \cdots & R_1^{(L)} \\ \vdots & \ddots & \vdots \\ R_N^{(1)} & \cdots & R_N^{(L)} \end{pmatrix}$$

Let y (r) be a realization of Y (R) and y_i (r_i) be its corresponding ith row. We assume that the MDM takes the form

$$\Pr(R = r|Y = y) = \prod_{i=1}^{N} \Pr(R_i = r_i|Y_i = y_i);$$

that is, units are observed or missing independently of each other. We also assume that units are iid with pmf

$$f(y) = \Pr(Y = y) = \prod_{i=1}^{N} h(y_i).$$

We seek to model the general situation of inference on outcome variables given conditioning variables when there are latent (unobserved or excluded) confounders. We therefore partition the column labels $\{1, \cdots, L\}$ into label sets $H(\neq \emptyset)$, K, and M representing outcome, conditioning, and latent variables, respectively. We partition each unit i in the same way: $Y_i = (Y_i^H, Y_i^K, Y_i^M)$ and $R_i = (R_i^H, R_i^K, R_i^M)$. Here, Y_i^H and R_i^H contain the elements of Y_i and R_i corresponding to the variables H, and similarly for K and M. We partition the realizations y_i and r_i in the same way: $y_i = (y_i^H, y_i^K, y_i^M)$ and $r_i = (r_i^H, r_i^K, r_i^M)$. Let Y^H be a submatrix of Y containing the columns corresponding to the variables H, and similarly for Y^K and Y^M, and define the realizations y^H, y^K, and y^M in the same way. Let Ω^H, Ω^K, and Ω^M be the sample spaces of Y_i^H, Y_i^K, and Y_i^M, respectively. We define the vector $\tilde{y}^K \in \Omega^K$ to be the generic conditioning vector, in that our objective is to make inferences regarding the conditional distribution of $Y_i^H|Y_i^K = \tilde{y}^K$. We define $K = \emptyset$ to refer to the situation where we are interested in the marginal distribution of Y^H.

Let $u \in \Omega^H$. Define $I(y_i^H < u) = 1$ if $y_i^{(j)} < u^{(j)}\ \forall j \in H$. Let 1^A be a $q \times 1$ vector of 1's where $q = \#A$. Let $n_{H,K,\tilde{y}^K} = \#\{i : r_i^H = 1^H, r_i^K = 1^K, y_i^K = \tilde{y}^K\}$ be the number of units fully observed for y_i^K and y_i^H with $y_i^K = \tilde{y}^K$. Let $T_{H,K,i} = 1$ indicate $r_i^H = 1^H$ and $r_i^K = 1^K$. Therefore

$$n_{H,K,\tilde{y}^K} = \sum_{i:Y_i^K=\tilde{y}^K} T_{H,K,i}.$$

Assuming $n_{H,K,\tilde{y}^K} > 0$, define

$$\begin{aligned}\bar{I}_{H,K,\tilde{y}^K,u} &= \frac{1}{\sum_{i:Y_i^K=\tilde{y}^K} T_{H,K,i}} \sum_{i:Y_i^K=\tilde{y}^K} T_{H,K,i} \cdot I(Y_i^H < u) \\ &= \frac{1}{n_{H,K,\tilde{y}^K}} \sum_{i:Y_i^K=\tilde{y}^K} T_{H,K,i} \cdot I(Y_i^H < u)\end{aligned}$$

as the sample mean of $I(Y_i^H < u)$ across the rows where $Y_i^K = \tilde{y}^K$. Let $F_{H|K}(u|\tilde{y}^K)$ be the cdf of $Y_i^H|Y_i^K = \tilde{y}^K$ evaluated at u. For $n_{H,K,\tilde{y}^K} > 0$, we seek to determine conditions that imply

$$\mathrm{E}\left(\bar{I}_{H,K,\tilde{y}^K,u} \middle| Y^K = \tilde{y}^K, R = r\right) = F_{H|K}(u|\tilde{y}^K).$$

This is a valuable nonparametric model because a range of parameters, including means and quantiles, are functionals of the cdf.

2.3.2. Illustration in the Renal Transplant Study

We illustrate the theorem using a subset of the data of Dienemann et al. (“Improvements in muscle mass and function following renal transplantation”), who studied changes in body composition and muscle function in renal transplant recipients. The study enrolled 60 patients and measured body composition at the time of transplantation. The data consist of sex-, race-, and age-specific Z-scores.

We suppose that the objective is to estimate the distribution of Muscle Strength Z-scores given positive or negative Muscle Area Z-score, taking Fat Area to be a latent variable. Hence, Muscle Strength Z-score is the outcome Y^H; dichotomized Muscle Area Z-score is the conditioning variable Y^K; and Fat Area Z-score is a latent variable Y^M. Here u represents any element in the space of Muscle Strength Z-score.

2.3.3. Summary of Lemmas

We summarize our results into four lemmas and their consequent theorem, whose proofs appear in Appendix A. Lemmas 1 and 2 give sufficient conditions under which the ith unit satisfies

$$\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) = F_{H|K}(u|\tilde{y}^K).$$

That is, Lemmas 1 and 2 give sufficient conditions under which the empirical conditional cdf for a single unit is unbiased for the population conditional cdf. The key difference between Lemmas 1 and 2 are in their assumptions. Lemma 1 requires that the probability the ith unit is observed is independent of the outcome and latent variables, whereas Lemma 2 requires that the probability is independent of only the outcome variables. Lemma 2 has the additional requirement that the outcome and latent variables are independent given the conditioning variables. Lemma 3 shows the equivalency in the statement $Y_i^H < u$ in Lemmas 1 and 2 with the more familiar statement $Y_i^H = u$. Lemma 4 states conditions sufficient for a sample cdf estimate $\bar{I}_{H,K,\tilde{y}^K,u}$ to be unbiased for $F_{H|K}(u|\tilde{y}^K)$.

2.3.4. Lemma 1

Given u and $\tilde{y}^K$, suppose the following conditions hold for all $y_i^M \in \Omega^M$:

1. $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K)$, and
2. $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$.

Then $\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) = F_{H|K}(u|\tilde{y}^K)$.

2.3.5. Lemma 2

Given u and $\tilde{y}^K$, suppose the following conditions hold for all $y_i^M \in \Omega^M$:

1. $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$,
2. $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$, and
3. $\Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K)$.

Then $\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) = F_{H|K}(u|\tilde{y}^K)$.

2.3.6. Lemma 3

Assume $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$ for all u, y_i^M. Then

$$\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \quad \text{for all } u, y_i^M$$

if and only if

$$\Pr(T_{H,K,i} = 1 | Y_i^H = u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \quad \text{for all } u, y_i^M$$

We note that, in the Lemma 3 statement and the proof, the expression $\Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$ is interchangeable with $\Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K)$.

2.3.7. Lemma 4

Suppose $n_{H,K,\tilde{y}^K} > 0$ and $\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) = F_{H|K}(u|\tilde{y}^K)$ for all units in the set $\{i : r_i^H = 1^H, r_i^K = 1^K, y_i^K = \tilde{y}^K\}$. Then $\mathrm{E}\left(\bar{I}_{H,K,\tilde{y}^K,u} \middle| Y^K = y^K, R = r\right) = F_{H|K}(u|\tilde{y}^K)$.

The Theorem below summarizes and integrates the lemmas.

2.3.8. Theorem

Given $\tilde{y}^K$ and suppose $n_{H,K,\tilde{y}^K} > 0$. Suppose further that for each unit i in the set $\{i : r_i^H = 1^H, r_i^K = 1^K, Y_i^K = \tilde{y}^K\}$ one of the following sets of conditions holds for all u and y_i^M:

1. (a) $\Pr(T_{H,K,i} = 1 | Y_i^H = u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K)$, and

   (b) $\Pr(T_{H,K,i} = 1 | Y_i^H = u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$; or

2. (a) $\Pr(T_{H,K,i} = 1 | Y_i^H = u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$,

   (b) $\Pr(T_{H,K,i} = 1 | Y_i^H = u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$, and

   (c) $\Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K)$.

Then $\mathrm{E}\left(\bar{I}_{H,K,\tilde{y}^K,u} \middle| Y^K = y^K, R = r\right) = F_{H|K}(u|\tilde{y}^K)$.

The Theorem states that the empirical conditional cdf is guaranteed unbiased if, for each unit with complete data in the outcome and conditioning variables, one of two sets of conditions holds. The first set of conditions has two elements:

1. The probability of being observed is independent of outcome and latent variables given the conditioning variables (i.e., the MDM can depend on conditioning variables).

2. Each combination of data has a non-zero probability of being observed.

The second set of conditions has three elements:

1. The probability of being observed is independent of outcome variables given conditioning and latent variables (i.e., the MDM can depend on conditioning and latent variables).

2. Each combination of data has a non-zero probability of being observed.

3. The outcome variables are independent of the latent variables given the conditioning variables.

The assumptions apply to the observed missingness matrix $R = r$, not to all possible values of R. In parallel with Seaman et al. (2013), the assumptions refer to "realised" rather than "everywhere" ignorability. See also Heitjan (1997).

2.3.9. Application of the Theorem to the Renal Transplant Data

Returning to the example, suppose there are N units from which we collect the Muscle Strength, Muscle Area, and Fat Area Z-scores. Assume our objective is to estimate the cdf of the Muscle Strength Z-score given a positive Muscle Area Z-score. From Rubin (1976), a sufficient condition for frequentist ignorability is that the missing data are MCAR. With our model, however, sufficient conditions are in some cases less restrictive.

Under our theorem, one sufficient condition for lack of bias is that the MDM depends only on the conditioning variable (Muscle Area Z-score). When the latent (Fat Area) and outcome (Muscle Strength) variables are independent given the conditioning variable (Muscle Area), one can permit the MDM to also depend on the latent variable. Essentially, this means that the latent variable is not a confounder.

If we are instead interested in the marginal cdf of Muscle Strength, both Muscle Area and Fat Area are latent variables, and our ignorability condition is equivalent to MCAR when the outcome Muscle Strength and latent variables are correlated. When the latent and outcome variables are independent of each other, the MDM can depend on the latent variables without inducing bias.

2.4. Simulation Study

We use simulation to illustrate the potential biases arising in this situation and to demonstrate the Theorem. For unit i, let MS_i, MA_i, and FA_i be the Muscle Strength, Muscle Area, and Fat Area Z-scores, respectively. After removing incomplete units there were 57 complete cases. We copied these cases to create a large dataset of 1824 units (32 copies per unit).

We performed a simulation on the data by generating Muscle Strength Z-score outcomes and artificially inducing data to be missing. We created Muscle Strength Z-score data as

$$\mathsf{MS}_i = -1 + a/2 \cdot \mathsf{I}(\mathsf{MA}_i > 0) + b/2 \cdot \mathsf{I}(\mathsf{FA}_i > 1) + \epsilon,$$

where $\epsilon \sim N(0, 1)$. We considered three models for the mean: i) $a = b = 1$, in which both Muscle Area and Fat Area predict Muscle Strength; ii) $a = 1$, $b = 0$, in which only Muscle Area predicts Muscle Strength; and iii) $a = 0$, $b = 1$, in which only Fat Area predicts Muscle Strength. We simulated the Fat Area latent variable as $\mathsf{FA}_i \sim N(1, 1/4)$.

We generated missing data indicators according to the MDM

$$\begin{aligned}\Pr(R_i = 0|\mathsf{MS}_i, \mathsf{MA}_i, \mathsf{FA}_i) = 3/10 + c/5 \cdot \mathsf{I}(\mathsf{MS}_i < -1/2) + d/5 \cdot \mathsf{I}(\mathsf{MA}_i > 0) \\ + e/5 \cdot \mathsf{I}(\mathsf{FA}_i > 1).\end{aligned}$$

We consider three missingness models: i) $c = d = e = 1$, in which Muscle Strength, Muscle Area, and Fat Area are associated with missingness; ii) $c = 0$, $d = e = 1$, in which only Muscle Area and Fat Area are associated with missingness; and iii) $c = e = 0$, $d = 1$, in which only Muscle Area is associated with missingness.

After creating the artificial Muscle Strength outcomes and generating the missing-data indicators, we computed the 25th, 50th, and 75th quantiles of the empirical distribution of Muscle Strength

Z-score given positive Muscle Area Z-score. We compared this to the quantile values of the true distribution, which is the two-component normal mixture

$$1/2N(-1+a/2,1)+1/2N(-1+a/2+b/2,1)$$

Table 1 shows the combinations of mean and missingness models. The first set of conditions in the Theorem applies to Scenarios 3, 5, and 7, where the missingness model depends only on the conditioning variable Muscle Area. The second set of conditions applies to Scenario 4, where the missingness model depends on both the conditioning variable (Muscle Area) and the latent variable (Fat Area), whereas the mean model is independent of the latent variable. Hence, for Scenarios 3, 4, 5, and 7, the empirical conditional cdf is unbiased. Scenarios 1, 2, and 6 do not satisfy the sufficient conditions because both the mean and missingness models depend on the latent variable. Scenario 1 also has the mean model dependent on the outcome variable Muscle Strength. Hence, for Scenarios 1, 2, and 6, estimation of the conditional cdf is potentially biased.

For each scenario, we simulated 10,000 data sets. Each mechanism resulted in roughly half to two-thirds of the units being missing. Table 2 shows descriptive statistics for the quantiles, including percent bias. As the Theorem predicts, Scenarios 3, 4, 5, and 7 all had minimal bias, whereas Scenarios 1, 2, and 6 had substantial bias.

2.5. Discussion

We have discussed sufficient conditions for correct analysis of frequentist nonparametric inference on conditional distributions subject to incomplete data. Our conditions relax those of Rubin (1976). That is, assuming that the MDM is dependent on conditioning variables only is sufficient for unbiased estimation of a conditional distribution. This can be relaxed further to have the MDM also depend on latent variables, provided the latent and outcome variables are independent given the conditioning variables. If the MDM depends on conditioning variables that are missing, generally the data are NMAR. For marginal distributions, the sufficient conditions are equivalent to those of Rubin (1976). That is, MCAR is sufficient for correct inference of a marginal distribution. Again, this can be relaxed to have the MDM depend on latent variables if the latent and outcome variables are independent. A strength of our model is that it is completely free of parametric assumptions.

Our analysis addresses complete-case inference, which can be inefficient as it forfeits information from incomplete cases. Statistical methods that make use of the incomplete units will likely require more restrictive assumptions on the MDM (Little and Zhang, 2011; Rubin, 1987). White and Carlin (2010) showed that multiple imputation is preferable to complete-case analysis under MAR mechanisms, though complete-case is often preferable under MNAR mechanisms. Bartlett et al. (2014) argued that in studies in which missingness occurs only in the variable of interest, methods that use the incomplete units offer little efficiency gain for analysis of the covariate effect.

Our analysis begs the question of what to do when ignorability conditions are not satisfied. We seldom are in a position to know the MDM, still less to know that it is ignorable. A possible next step is to collect an additional sample of observations that is free of missing observations, which we could then use to "patch" inferences and thereby eliminate bias. We will pursue this approach in ensuing work.

Table 2.1: Combinations of mean model and missingness model considered for simulation of renal transplant data ("Improvements in muscle mass and function following renal transplantation"). For a given scenario, a check mark denotes that the variable is associated with the outcome Muscle Strength Z-score or the variable is associated with missingness. The final column indicates whether the theorem predicts absence of bias.

Scenario	Mean Model		Missingness Model			Predicted Unbiased?
	Muscle Area Z-score Covariate	Fat Area Z-score Latent	Muscle Strength Z-score Outcome	Muscle Area Z-score Covariate	Fat Area Z-score Latent	
1	✓	✓	✓	✓	✓	
2	✓	✓		✓	✓	
3	✓	✓		✓		✓
4	✓			✓	✓	✓
5	✓			✓		✓
6		✓		✓	✓	
7		✓		✓		✓

Table 2.2: For a given Scenario, listed are whether the theorem predicts absence of bias and the simulated percent bias in the 25, 50, and 75th centiles of the empirical distribution of Muscle Strength Z-score given positive Muscle Area Z-score.

Scenario	Predicted Unbiased?	% Quantile Bias		
		25^{th}	50^{th}	75^{th}
1		8.83	35.18	-20.44
2		6.25	24.71	-14.48
3	✓	-0.22	-0.46	0.09
4	✓	-0.15	-0.02	-1.21
5	✓	-0.17	-0.07	-0.94
6		4.08	8.19	116.57
7	✓	-0.15	-0.26	-0.66

CHAPTER 3

THE FIRST ORDER MARKOV MAXIMUM LIKELIHOOD BASED APPROACH FOR COUNT DATA WITH OVER-DISPERSION

3.1. Introduction

Longitudinal data are commonly encountered in medical studies. For example, clinical trials often collect repeated measurements on patients at pre-specified measurement occasions, in order to compare the effectiveness of two or more treatments. When analyzing the data from these trials, it is important to account for the intra-subject correlation of measurements, in order to avoid the loss in efficiency in estimation of the regression parameter that can occur when models are specified under an incorrect assumption of independence.

In this Chapter we consider a clinical study that evaluated the effectiveness of the drug progabide relative to placebo in the treatment of seizures (Thall and Vail, 1990). The primary outcome in this trial was a count that represented the number of seizures that each patient experienced during consecutive time periods during follow-up. In addition to the anticipated similarity between the repeated counts on each subject (which would result in positive intra-subject correlations), it was observed that the patient level sample variances of seizure counts greatly exceeded the sample means. The seizure counts were therefore over-dispersed relative to the Poisson distribution that is often applied for analysis of count data, but which assumes equality of the means and variances. Appropriate analysis of the seizure data from Thall and Vail (1990) should therefore account for over-dispersion in addition to the intra-subject correlation of measurements.

Our goal was to provide an approach for analysis of over-dispersed longitudinal discrete data in the framework of generalized linear models, with a focus in this Chapter on outcome variables that are counts. Perhaps the most widely used statistical method that provides a unified framework for analysis of correlated variables that may be continuous or discrete is the generalized estimating equation (GEE) approach Liang and Zeger (1986). GEE extended generalized linear models (GLM) McCullagh and Nelder (1989) to longitudinal data by incorporating patterned correlation matrices into likelihood equations that were originally obtained under an assumption of independence. Par-

ticular patterned correlation matrices (Liang and Zeger, 1986) could be specified, depending on the nature of the study. For example, an exchangeable structure (with equal intra-cluster correlations) might be appropriate for clustered data in a cross-sectional study, while an AR(1) structure (with intra-subject correlations that decay with increasing separation in time) might be appropriate for longitudinal trials.

The GEE approach has a number of attractive features, in addition to providing a unified approach for analysis of correlated data that is in the framework of GLM. GEE models are straightforward to specify because they include the usual regression models for a two parameter exponential family (that include linear, logistic, or Poisson regression models) coupled with the patterned correlation matrix to describe the pairwise association of measurements within subjects, or clusters. They are also viewed as being relatively robust to misspecification of the patterned correlation matrix (working correlation structure) because the GEE estimator of the regression parameter will be consistent even if the choice of patterned structure is not correct.

However, there are some limitations to GEE. First, GEE is a framework for estimation that involves application of moment based estimates of the correlation parameters that are functions of the Pearson residuals. The approach does not specify which moment estimator should be used in an analysis. For example, SAS, Stata, the geepack package in R, and the original manuscript of Liang and Zeger (1986) all differ with respect to their choice of estimator for the AR(1) structure. When implementing GEE for a particular patterned correlation structure, it may not be readily apparent which estimator of the correlation parameter was implemented in the analysis, which can have negative consequences. For example, Stata's implementation of the AR(1) structure drops subjects with only one measurement from the analysis, which does not occur for other correlation structures in Stata, or when implementing the AR(1) structure for GEE in SAS or R.

Another limitation of GEE is that although the approach is robust with respect to choice of working structure with respect to consistency of the regression parameter estimators, the estimators of the correlation parameters may fail to be consistent if the working structure is misspecified. Crowder (1995) demonstrated that the GEE estimation procedure can fail even in simple misspecification scenarios, such as when the AR(1) structure is misspecified as exchangeable. If the limiting value of the estimator of the correlation parameter tends to a value outside the feasible interval (interval that yields a positive definite structure) this can result in a breakdown of the estimation procedure

(Crowder, 1995). Or, if the estimator of the correlation parameter tends to a feasible value that is not the true value (i.e. the estimator is not consistent but tends to a value that yields a positive definite structure) then the covariance matrix of the estimators will not be estimated consistently (Sutradhar and Das, 1999, 2000) and the p-values for tests involving the regression parameters will be invalid. In general, there may loss in precision of estimation if the assumed and true patterns of association are not close (Diggle et al., 2002; Fitzmaurice, Laird, and Ware, 2011).

It is also possible to perform a GEE analysis and unknowingly obtain estimates that are invalid because they are not compatible with any valid parent distribution in which case it is impossible to construct a valid multivariable distribution that has the fitted marginal means and correlation structures. For example, Prentice (1988) described additional constraints for the correlation estimates that are necessary (although not necessarily sufficient) to guarantee the existence of a valid multivariable parent distribution for longitudinal binary data. (The fitted marginal means and pairwise correlations completely determine the bivariate Bernoulli distributions. If the Prentice constraints (Prentice, 1988) are not satisfied, some of the bivariate probabilities will be negative.)

Assessing goodness of fit can also be more challenging for GEE and other estimating equation based approaches that do not start with an objective function (the log-likelihood), although attempts have been made to extend goodness of fit criteria such as Akaike's Information Criterion for GEE (Pan, 2001). The lack of a log-likelihood also means that it is not possible to perform likelihood ratio tests for comparison of nested models. The likelihood ratio test is especially useful for the analysis of count data, to distinguish between the Poisson and negative binomial distributions.

GEE also requires the assumption that any missing data are missing completely at random (MCAR). The MCAR assumption is often unreasonable for clinical trials, because patients who drop out of a study may tend to do so for reasons related to their treatment (e.g. worsening symptoms) or to other measured variables such as age or gender.

A final limitation of GEE is that it assumes that there is no over-dispersion, so that no adjustments are made during the GEE iterative estimation procedure to account for over (or under) dispersion.

Other authors suggested estimating equation based approaches for analysis of over-dispersed count data, including Thall and Vail, 1990, who presented a heuristic derivation of covariance matrices for count data that was based on random effects coupled with estimating equations for the

regression and correlation parameters. However, the approach of Thall and Vail, 1990 does not allow for implementation of all plausible structures, including the AR(1) structure that we will consider in this Chapter. All estimating equation based approaches share the same limitations, namely lack of a likelihood for likelihood ratio testing and assessment of goodness of fit; no guarantee of a valid of a parent distribution; and requirement of the MCAR assumption for missing data.

In contrast to estimating equation based approaches, another class of available models are generalized linear mixed-effects models that include random-effects Poisson and Negative Binomial models (Frees, 2004). The likelihoods for mixed-effects models can be complex and usually require integration over the random effects distributions. As a result, failure to converge can be an issue for these models. The suitability of the distributional assumptions regarding the random effects can be difficult to assess. It is also difficult to implement a plausible correlation structure for the outcome variable for random-effects models, because an assumed correlation structure for the random effects does not induce the same correlation structure for the outcome variable. For example, assuming an AR(1) structure for the random effects does not yield an AR(1) structure for the outcome variable, when the outcome variable is discrete and a non-identity link function is used to relate the marginal means with covariates. Zhang et al. (2011) also argues that GEE and generalized linear mixed-effects models are the two most popular paradigms that extend methods from cross-sectional to correlated data. However, there are conceptual differences between GEEs and generalized linear mixed-effects models that can make it difficult to interpret and estimate the regression parameters in mixed-effects models (Zhang et al., 2011).

In this Chapter we present an approach for analysis of over-dispersed longitudinal data that like GEE, extends generalized linear models to correlated data. However, unlike GEE, our approach is likelihood based. We assume distributions that are members of the exponential family. In order to account for intra-subject correlations with structures that are plausible for longitudinal data, we also assume first-order antedependence and linearity of the expectations of the conditional distributions. Our approach allows us to specify the usual generalized linear model for the outcome variable coupled with a decaying product working correlation structure that specifies the correlation between two measurements as the product of the correlations of intermediate and adjacement measurements. For example, for a decaying product correlation structure the correlation between the second and fourth measurement on a subject is the product of the correlation between the second and third

measurements and the correlation between the third and fourth measurements. If the adjacent correlations are assumed to be constant, then the data has a first-order autoregressive (AR(1)) correlation structure. The AR(1) correlation structure forces a decline in the correlations with increasing separation in time. This structure is plausible for longitudinal studies because we often expect that two outcomes measured closer in time will be more correlated than if they are measured farther apart in time. However, assuming constant adjacent correlation may not be plausible for data that are unequally spaced in time. Unequal spacing of measurements can be accounted for by allowing the adjacent correlations to depend on their separation in time, as in a Markov correlation structure. The AR(1) and Markov structures are not appropriate, however, when we expect different adjacent correlations for different periods. For example, we might expect the correlation between the first and second measurements to be different than the correlation between the second and third, regardless of the differences in time. For these situations, we assume the adjacent correlation is unique to the time period; this induces the first-order ante-dependent correlation structure (AD(1)).

This Chapter is organized as follows. In Section 3.2, we provide the assumptions and derived likelihood. In Section 3.2.2, we describe our procedure for estimating the log-likelihood. We present an analysis of the seizure count data from Thall and Vail (1990) to demonstrate application of the methods in Section 3.3. Simulation results are presented in Section 3.4. Finally, discussion and concluding remarks are presented in Section 3.5.

3.2. Methods

3.2.1. Assumptions and Likelihood

Suppose we collect longitudinal data on m units. The ith unit ($i = 1, \cdots, m$) has n_i measurements ($j = 1, \cdots, n_i$). For the ith unit, outcome measurements $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$ are collected at the corresponding times $T_i = (t_{i1}, \cdots, t_{in_i})^T$. At each time point, covariate data $x_{ij} = (x_{ij1}, \cdots, x_{ijp})$ is collected. Let $y_i = (y_{i1}, \cdots, y_{in_i})^T$ be a realization of Y_i. Let $\mu_{ij} = E(Y_{ij})$, $\sigma_{ij}^2 = Var(Y_{ij})$, and $C_{ijk} = Corr(Y_{ij}, Y_{ik})$.

First, we assume first-order antependence (Gabriel, 1962), so that the likelihood of $(Y_i, \cdots, Y_m)$ can be expressed as

$$\prod_{i=1}^{m} f(Y_{i1}) \prod_{j=2}^{n_i} f(y_{ij}|f(y_{ij-1})). \tag{3.1}$$

Next, we assume that the distribution of Y_{i1}, $f(Y_{i1})$, and the *conditional distributions* of Y_{ij} given Y_{ij-1}, $f(Y_{ij}|Y_{ij-1})$, $(j = 2, \ldots, n_i)$ are members of the same exponential family. The likelihood (3.1) can then be expressed as

$$\begin{aligned} L(\beta, \alpha) &= \prod_{i=1}^{m} f(Y_{i1} = y_{i1}) \prod_{j=2}^{n_i} f(Y_{ij} = y_{ij} | Y_{ij-1} = y_{ij-1}) \\ &= \prod_{i=1}^{m} \exp\left(\frac{y_{i1}\theta_{i1} - b(\theta_{i1})}{a(\phi)} - c(y_{i1}, \phi) \right) \prod_{j=2}^{n_i} \exp\left(\frac{y_{ij}\theta^*_{ij} - b(\theta^*_{ij})}{a(\phi^*)} - c(y_{ij}, \phi^*) \right) \end{aligned} \tag{3.2}$$

where $a()$, $b()$, $c()$ are functions specific to the assumed distribution and ϕ, ϕ^* are the dispersion parameters (unrelated to model overdispersion, these are parameters of an exponential family). For a particular distribution, let $g()$ be the link function, for which $\theta_{ij} = g(\mu_{ij})$ $(i = 1, \ldots, m; j = 1, \ldots, n_i)$ and $\theta^*_{ij} = g(\mu^*_{ij})$ $(i = 2, \ldots, m; j = 1, \ldots, n_i)$.

We next assume that the conditional expectation $E(Y_{ij}|Y_{ij-1})$ is a linear function of Y_{ij-1}. The assumptions of first-order antedependence and linearity of the conditional expectations imply the following (Theorems 2.1 and 2.2 of Guerra and Shults (2014)). First, for $j = 2, \cdots, n_i$,

$$E(Y_{ij}|Y_{ij-1}) \equiv \mu^*_{ij} = \mu_{ij} + C_{ijj-1}\frac{\sigma_{ij}}{\sigma_{ij-1}}(Y_{ij-1} - \mu_{ij-1}) \tag{3.3}$$

where

$$\sigma^2_{ij} = \frac{E(Var(Y_{ij}|Y_{ij-1}))}{1 - C^2_{ijj-1}}.$$

There are two important points to make about this results. The first is that the expectation of the conditional expectation is μ_{ij}:

$$\begin{aligned} E(E(Y_{ij}|Y_{ij-1})) &= \mu_{ij} + C_{ijj-1}\frac{\sigma_{ij}}{\sigma_{ij-1}}E(Y_{ij-1} - \mu_{ij-1}) \\ &= \mu_{ij} \end{aligned}$$

This suggest the marginal means are averages of the conditional expectations. The second point is that overdispersion is induced in the marginal means. Table 3.5 lists the marginal variances (derivations will come later in this Chapter and in Chapter 4). We will discuss this in further detail when discussing the individual distribution assumptions.

Next, the correlation between Y_{ij} and Y_{ij+t} for $t > 0$ has decaying product form

$$C_{ijj+t} = \prod_{k=j}^{j+t-1} C_{ikk+1}. \tag{3.4}$$

As noted earlier, product correlation structures are often plausible for longitudinal data because they force a decline in the (absolute) correlations with increasing separation in measurement occasion. If $C_{ijj-1} = \alpha$ in (3.4) the structure is the auto-regressive correlation structure of order 1, or AR(1). With its assumption of equal adjacent correlations, the AR(1) structure is often plausible for measurements that are equally spaced in time. Next, if $C_{ijj-1} = \alpha^{t_{ij}-t_{ij-1}}$ the structure is Markov, which is appropriate for unequally spaced data and includes the AR(1) structure as a special case (when $t_{i1}, \ldots, t_{in_i} = 1, \ldots, n_i \; \forall \; i$ and $j = 1, \ldots, n_i$). Finally, if $C_{ijj-1} = \alpha_{j-1}$ the structure is first-order antedependent, or AD(1). The AD(1) structure is plausible when the adjacent correlations vary with time. It includes the AR(1) structure as a special case (when $\alpha_1 = \ldots = \alpha_N$ for $N = max\{n_i\}$).

3.2.2. Likelihood Equations

General Form of Estimating Equations

To obtain the log likelihood function we take the natural log of (3.2), to obtain

$$\ln(L(\beta, \alpha)) = \sum_{i=1}^{m} \left(\frac{y_{i1}\theta_{i1} - b(\theta_{i1})}{a(\phi)} - c(y_{i1}, \phi) + \sum_{j=2}^{n_i} \left(\frac{y_{ij}\theta^*_{ij} - b(\theta^*_{ij})}{a(\phi^*)} - c(y_{ij}, \phi^*) \right) \right). \tag{3.5}$$

Differentiating the log likelihood function (3.5) with respect to β, we obtain

$$\begin{aligned} \frac{\partial \ln(L(\beta, \alpha))}{\partial \beta} &= \sum_{i=1}^{m} \left(\frac{y_{i1} - b'(\theta_{i1})}{a(\phi)} \frac{\partial \theta_{i1}}{\partial \beta} + \sum_{j=2}^{n_i} \frac{y_{ij} - b'(\theta^*_{ij})}{a(\phi^*)} \frac{\partial \theta^*_{ij}}{\partial \beta} \right) \\ &= \sum_{i=1}^{m} \left(\frac{y_{i1} - \mu_{i1}}{a(\phi)} \frac{\partial g(\gamma)}{\partial \gamma}\Big|_{\gamma=\mu_{i1}} \frac{\partial \mu_{i1}}{\partial \beta} + \sum_{j=2}^{n_i} \frac{y_{ij} - \mu^*_{ij}}{a(\phi^*)} \frac{\partial g(\gamma)}{\partial \gamma}\Big|_{\gamma=\mu^*_{ij}} \frac{\partial \mu^*_{ij}}{\partial \beta} \right) \end{aligned}$$

The last component, $\frac{\partial \mu_{ij}^*}{\partial \beta}$, takes value (for $j = 2$ and $j > 2$, respectively):

$$\frac{\partial \mu_{i2}^*}{\partial \beta} = \frac{\partial \mu_{i2}}{\partial \beta} + \frac{C_{i21}}{\sqrt{1 - C_{i21}^2}} \frac{\sqrt{E(Var(Y_{i2}|Y_{i1}))}}{\sqrt{Var(Y_{i1})}} \left(\frac{Y_{i1} - \mu_{i1}}{2} \right.$$
$$\times \left(\frac{1}{E(Var(Y_{i2}|Y_{i1}))} \frac{\partial E(Var(Y_{i2}|Y_{i1}))}{\partial \beta} - \frac{1}{Var(Y_{i1})} \frac{\partial Var(Y_{i1})}{\partial \beta} \right)$$
$$\left. - \frac{\partial \mu_{i1}}{\partial \beta} \right)$$

$$\frac{\partial \mu_{ij}^*}{\partial \beta} = \frac{\partial \mu_{ij}}{\partial \beta} + C_{ijj-1} \frac{\sqrt{1 - C_{ij-1j-2}^2}}{\sqrt{1 - C_{ijj-1}^2}} \frac{\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}{\sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))}}$$
$$\times \left(\frac{Y_{ij-1} - \mu_{ij-1}}{2} \left(\frac{1}{E(Var(Y_{ij}|Y_{ij-1}))} \frac{\partial E(Var(Y_{ij}|Y_{ij-1}))}{\partial \beta} \right. \right.$$
$$\left. \left. - \frac{1}{E(Var(Y_{ij-1}|Y_{ij-2}))} \frac{\partial E(Var(Y_{ij-1}|Y_{ij-2}))}{\partial \beta} \right) - \frac{\partial \mu_{ij-1}}{\partial \beta} \right)$$

See Appendix B.1 for the full derivation of the β estimating equation. Differentiating the log likelihood with respect to α, we obtain

$$\frac{\partial \ln(L(\beta, \alpha))}{\partial \alpha} = \sum_{i=1}^{m} \left(\frac{y_{i1} - b'(\theta_{i1})}{a(\phi)} \frac{\partial \theta_{i1}}{\partial \alpha} + \sum_{j=2}^{n_i} \frac{y_{ij} - b'(\theta_{ij}^*)}{a(\phi^*)} \frac{\partial \theta_{ij}^*}{\partial \alpha} \right)$$
$$= \sum_{i=1}^{m} \sum_{j=2}^{n_i} \left(\frac{y_{ij} - \mu_{ij}^*}{a(\phi^*)} \frac{\partial g(\gamma)}{\partial \gamma} \Big|_{\gamma = \mu_{ij}^*} \frac{\partial \mu_{ij}^*}{\partial \alpha} \right)$$

See Appendix B.2 for the full derivation of the α estimating equation. The value of the last component, $\frac{\partial \mu_{ij}^*}{\partial \alpha}$, depends on the induced correlation structure.

When the true correlation structure is AR(1), $\frac{\partial \mu_{ij}^*}{\partial \alpha}$ takes value (for $j = 2$ and $j > 2$, respectively; see Appendix B.2.1 for full derivation):

$$\frac{\partial \mu_{i2}^*}{\partial \alpha} = \sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}} \frac{Y_{i1} - \mu_{i1}}{(1 - \alpha^2)^{3/2}}$$

$$\frac{\partial \mu_{ij}^*}{\partial \alpha} = \sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}} (Y_{ij-1} - \mu_{ij-1})$$

When the true correlation structure is Markov, $\frac{\partial \mu_{ij}^*}{\partial \alpha}$ takes value (for $j = 2$ and $j > 2$, respectively; see Appendix B.2.2 for full derivation):

$$\frac{\partial \mu_{i2}^*}{\partial \alpha} = \frac{Y_{i1} - \mu_{i1}}{\sqrt{1 - \alpha^{2t_{i2} - 2t_{i1}}}} \sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}} (t_{i2} - t_{i1}) \alpha^{t_{i2} - t_{i1} - 1} \left(1 + \frac{\alpha^{t_{i2} - t_{i1}}}{1 - \alpha^{2t_{i2} - 2t_{i1}}} \right)$$

$$\begin{aligned} \frac{\partial \mu_{ij}^*}{\partial \alpha} = (Y_{ij-1} - \mu_{ij-1}) & \sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}} \sqrt{\frac{1 - \alpha^{2t_{ij-1} - 2t_{ij-2}}}{1 - \alpha^{2t_{ij} - 2t_{ij-1}}}} \alpha^{t_{ij} - t_{ij-1} - 1} \\ & \times \left(\frac{t_{ij} - t_{ij-1}}{1 - \alpha^{2t_{ij} - 2t_{ij-1}}} - \frac{(t_{ij-1} - t_{ij-2}) \alpha^{2t_{ij-1} - 2t_{ij-2}}}{1 - \alpha^{2t_{ij-1} - 2t_{ij-2}}} \right) \end{aligned}$$

Let $\hat{I}_j$ denote a vector containing a 1 in the jth element and 0 elsewhere and $\alpha = (\alpha_1, \cdots, \alpha_n)$. When AD(1) is the true correlation structure, $\frac{\partial \mu_{ij}^*}{\partial \alpha}$ takes value (for $j = 2$ and $j > 2$, respectively; see Appendix B.2.3 for full derivation):

$$\frac{\partial \mu_{i2}^*}{\partial \alpha} = (Y_{i1} - \mu_{i1}) \sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}} \frac{\hat{I}_1}{(1 - \alpha_1^2)^{3/2}}$$

$$\begin{aligned} \frac{\partial \mu_{ij}^*}{\partial \alpha} = (Y_{ij-1} - \mu_{ij-1}) & \sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}} \\ & \times \left(\hat{I}_{j-1} \frac{\sqrt{1 - \alpha_{j-2}^2}}{(1 - \alpha_{j-1}^2)^{3/2}} - \hat{I}_{j-2} \frac{\alpha_{j-1} \alpha_{j-2}}{\sqrt{(1 - \alpha_{j-1}^2)(1 - \alpha_{j-2}^2)}} \right) \end{aligned}$$

Estimation

In order to find the maximum likelihood estimators for our set of estimating equations, we use the package 'alabama' (Varadhan, 2015) that is available in the software R (R Core Team, 2013). The alabama package performs an augmented Lagrangian minimization algorithm for optimizing nonlinear objective functions with linear or nonlinear constraints.

The Augmented Lagrangian minimization algorithm finds the vector of variables x that minimize the

function $f(x)$, subject the the constraints

$$c_i(x) = 0,\ i = 1, \cdots, h$$
$$c_i(x) \geq 0,\ i = h+1, \cdots, k$$

The first h constraints are equality constraints whereas the remaining $k - h$ are inequality constraints. The Augmented Lagrangian method is similar to the penalty method in that it replaces the constrained optimization problem with a series of unconstrained problems. It does this by adding a penalty term to the objective function f and a second term that mimics a Lagrange multiplier. The resulting objective function is

$$g(x, \lambda, \sigma) = f(x) - \lambda^T d(x) + \frac{1}{2}\sigma d(x)^T d(x) \tag{3.6}$$

where

$$d_i(x) = \begin{cases} c_i(x) & \text{if } i \leq h \text{ or } c_i(x) \leq \frac{1}{\sigma}\lambda_i \\ \frac{1}{\sigma}\lambda_i & \text{if } i > h \text{ and } c_i(x) > \frac{1}{\sigma}\lambda_i \end{cases}$$

λ is the term that mimics a Lagrange multiplier and σ is the term that is the penalty term. A general description of the algorithm steps are as follows:

1. Choose initial values for x, λ, σ.
2. Next, repeat these steps until the stopping criteria are satisfied:
   (a) Compute the value of x that minimizes g in (3.6).
   (b) Update λ and σ.

Step 2a, that finds the x that minimizes g, is an unconstrained optimization problem. It is referred to as the "inner" loop and can be specified for different algorithms. For our purposes, we use the the Broyden - Fletcher - Goldfarb - Shanno (BFGS) algorithm as the inner loop, which is an unconstrained optimization algorithm that approximates Newton's method (Broyden, 1970a,b; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970). The BFGS routine is included in the standard package of R.

The solution to the constrained optimization problem is the resulting x is the steps above.

For the alabama package, the user provides starting values, the function to be minimized, the gradient (optional), constraints, and control parameters on the optimization algorithm. Since we seek to maximize the log likelihood equation, we simply reverse the sign of the log-likelihood. Because our log likelihood equation depends on many different factors, we developed R functions in which the user specifies the assumptions and the value of the log-likelihood is returned. The assumptions that the user specifies include the formula for the mean model expressed as a function of covariates, the correlation structure, and the distribution.

After running the alabama function, numerous components of the algorithm are returned, including whether the algorithm converged successfully. The log likelihood, the parameters $x = (\beta, \alpha), \lambda, \sigma$, the Gradient and Hessian of the Lagrangian function (3.6) are all returned. In addition, information regarding convergence is included, such as the number of iterations required to achieve convergence, and the values of the constraints. The code for the algorithm is provided in Appendix C.

Asymptotic distribution

Under weak regularity conditions (Bradley and Gart, 1962), the maximum likelihood estimator $\hat{\theta}$ (for variable $\theta = (\alpha, \beta)$) has an asymptotically normal distribution:

$$\sqrt{m}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N(0, I^{-1})$$

where

$$I = E\begin{pmatrix} \dfrac{\partial^2 \ln(L(\beta,\alpha))}{\partial \alpha^2} & \dfrac{\partial^2 \ln(L(\beta,\alpha))}{\partial \beta \partial \alpha} \\ \dfrac{\partial^2 \ln(L(\beta,\alpha))}{\partial \beta^T \partial \alpha} & \dfrac{\partial^2 \ln(L(\beta,\alpha))}{\partial \beta^T \partial \beta} \end{pmatrix}.$$

For our purposes, we estimate I by utilizing the Hessian matrix which is provided as output in optimization software.

3.2.3. Special Cases

Poisson

Here, we assume Y_{i1} and $Y_{ij}|Y_{ij-1}$ are distributed as Poisson. Taking u as a placeholder for μ_{i1} and μ^*_{ij}, the pdf is

$$\begin{aligned} f &= \frac{u_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!} \\ &= \exp\left(y_{ij}\ln(u_{ij}) - u_{ij} - \ln(y_{ij}!)\right) \end{aligned}$$

From here, we recognize the following component functions

$$\begin{aligned} \theta_{ij} &= \ln(u_{ij}) \\ a(\phi) &= 1 \\ b(\theta_{ij}) &= \mu_{ij} \\ c(y_{ij},\phi) &= \ln(y_{ij}!) \\ g(\gamma) &= \ln(\gamma) \\ g'(\gamma) &= \frac{1}{\gamma} \\ \mu_{ij} = \exp(x_i'\beta) Var(Y_{i1}) &= \mu_{i1} \\ \frac{\partial Var(Y_{i1})}{\partial\beta} &= \frac{\partial\mu_{ij}}{\partial\beta} \end{aligned}$$

We note above that we use the canonical inverse log-link function, $\mu_{ij} = \exp(x_i'\beta)$, which is standard practice for Poisson regression. Furthermore, for $j > 1$

$$\begin{aligned} E(Var(Y_{ij}|Y_{ij-1})) &= E(\mu^*_{ij}) = \mu_{ij} \\ \frac{E(Var(Y_{ij}|Y_{ij-1}))}{\partial\beta} &= \frac{\partial\mu_{ij}}{\partial\beta} \end{aligned}$$

Hence the marginal variance for $j > 1$ is $Var(Y_{ij}) = \frac{\mu_{ij}}{1 - C^2_{ijj-1}}$. Since $Var(Y_{ij}) > E(Y_{ij}) = \mu_{ij}$, it is clear the over dispersion is induced for the marginal distributions for $j > 1$.

We note that the Poisson distribution with AR(1) working structure was implemented in Gamerman,

Guerra, and Shults (2016).

Negative Binomial

Here, we assume Y_{i1} and $Y_{ij}|Y_{ij-1}$ are Negative-Binomial distributed with second parameter r. Taking u as a placeholder for μ_{i1} and μ^*_{ij}, the pdf is

$$\begin{aligned} f &= \left(\frac{ru_{ij}}{1+ru_{ij}}\right)^{y_{ij}} \left(\frac{1}{1+ru_{ij}}\right)^{1/r} \frac{\Gamma(y_{ij}+\frac{1}{r})}{\Gamma(y_{ij}+1)\Gamma(\frac{1}{r})} \\ &= \exp\left(y_{ij}\ln\left(\frac{ru_{ij}}{1+ru_{ij}}\right) - \frac{1}{r}\ln\left(1+ru_{ij}\right) + \ln\Gamma\left(y_{ij}+\frac{1}{r}\right) - \ln\Gamma(y_{ij}+1) - \ln\Gamma\left(\frac{1}{r}\right)\right) \end{aligned}$$

From here, we recognize the component functions to be

$$\begin{aligned} \theta_{ij} &= \ln\left(\frac{ru_{ij}}{1+ru_{ij}}\right) \\ a(\phi) &= 1 \\ b(\theta_{ij}) &= \frac{1}{r}\ln\left(1+ru_{ij}\right) \\ c(y_{ij},\phi) &= -\ln\Gamma\left(y_{ij}+\frac{1}{r}\right) + \ln\Gamma(y_{ij}+1) + \ln\Gamma\left(\frac{1}{r}\right) \\ g(\gamma) &= \ln\left(\frac{r\gamma}{1+r\gamma}\right) \\ g'(\gamma) &= \frac{1}{\gamma(1+r\gamma)} \\ \mu_{ij} &= \exp(x_i'\beta) \\ Var(Y_{i1}) &= \mu_{i1} + r\mu_{i1}^2 \\ \frac{\partial Var(Y_{i1})}{\partial\beta} &= (2r\mu_{i1}+1)\frac{\partial\mu_{i1}}{\partial\beta} \end{aligned}$$

We note above that use the non-canonical inverse log-link function, $\mu_{ij} = \exp(x_i'\beta)$, which is standard practice for Negative-Binomial regression. The canonical link function for the Negative-Binomial is not appropriate for modeling overdispersed Poisson data and the interpretation of the β coefficients would not be comparable to those found from Poisson regression. For a more thorough discussion of the different Negative-Binomial parameterizations see Hilbe (2011). Next, for $j > 1$ (see Appendix B.4 for derivation of $E((\mu^*_{ij})^2)$),

$$E(Var(Y_{ij}|Y_{ij-1})) = E\left(\mu_{ij} + r\mu_{ij}^2\right)$$

$$
\begin{aligned}
&= E(\mu_{ij}^*) + rE((\mu_{ij}^*)^2) \\
&= \mu_{ij} + r\left(\mu_{ij}^2 + \frac{C_{ijj-1}^2}{1 - C_{ijj-1}^2} E(Var(Y_{ij}|Y_{ij-1}))\right)
\end{aligned}
$$

Solving for $E(Var(Y_{ij}|Y_{ij-1}))$, we have

$$
\begin{aligned}
E(Var(Y_{ij}|Y_{ij-1}))\left(1 - r\frac{C_{ijj-1}^2}{1 - C_{ijj-1}^2}\right) &= \mu_{ij} + r\mu_{ij}^2 \\
E(Var(Y_{ij}|Y_{ij-1})) &= (\mu_{ij} + r\mu_{ij}^2)\left(1 - r\frac{C_{ijj-1}^2}{1 - C_{ijj-1}^2}\right)^{-1} \\
\frac{\partial E(Var(Y_{ij}|Y_{ij-1}))}{\partial \beta} &= \left(1 - r\frac{C_{ijj-1}^2}{1 - C_{ijj-1}^2}\right)^{-1} (2r\mu_{ij} + 1)\frac{\partial \mu_{ij}}{\partial \beta}
\end{aligned}
$$

Hence the marginal variance for $j > 1$ is $Var(Y_{ij}) = \dfrac{(\mu_{ij} + r\mu_{ij}^2)\left(1 - r\frac{C_{ijj-1}^2}{1 - C_{ijj-1}^2}\right)^{-1}}{1 - C_{ijj-1}^2}$. Since $Var(Y_{ij}) > E(Y_{ij}) = \mu_{ij}$, it is clear the over dispersion is induced for the marginal distributions for $j > 1$.

We note that this specifies a constraint on α. Since $E(Var(Y_{ij}|Y_{ij-1})) > 0$, that forces $1 - r\dfrac{C_{ijj-1}^2}{1 - C_{ijj-1}^2}$ to be positive as well. This means that

$$
\begin{aligned}
1 - r\frac{C_{ijj-1}^2}{1 - C_{ijj-1}^2} &> 0 \\
1 &> r\frac{C_{ijj-1}^2}{1 - C_{ijj-1}^2} \\
1 - C_{ijj-1}^2 &> rC_{ijj-1}^2 \\
1 &> rC_{ijj-1}^2 + C_{ijj-1}^2 \\
1 &> (r+1)C_{ijj-1}^2 \\
\frac{1}{r+1} &> C_{ijj-1}^2 \\
\sqrt{\frac{1}{r+1}} &> |C_{ijj-1}|
\end{aligned}
$$

If the correlation structure is AR(1), then the inequality is rewritten as $\sqrt{\dfrac{1}{r+1}} > |\alpha|$.

If the correlation structure is Markov, then the inequality is rewritten as $\sqrt{\dfrac{1}{r+1}} > |\alpha^{t_{ij} - t_{ij-1}}|$.

This is equivalent to $\left(\frac{1}{r+1}\right)^{\frac{1}{2(t_{ij}-t_{ij-1})}} > |\alpha|$. Since the equality is for each $j > 1$, α is constrained by the smallest difference in adjacent times. Written down algebraically, the constraint is $\left(\frac{1}{r+1}\right)^{\frac{1}{2 \cdot min(t_{ij}-t_{ij-1})}} > |\alpha|$

If the correlation structure is AD(1), then the inequality is rewritten as $\sqrt{\frac{1}{r+1}} > |\alpha_{j-1}|$. Since, the equality is for each $j > 1$, the constraint becomes $\sqrt{\frac{1}{r+1}}\hat{1} > |\alpha|$, where $|\alpha| = (|\alpha_1|, \cdots, |\alpha_{n-1}|)^T$ and $\hat{1} = (1, \cdots, 1)^T$ is a vector of length $n-1$ containing only 1's.

r is estimated using the maximum likelihood estimate. Finding the score equation with respect to r, we have (see Appendix B.3 for full derivations)

$$\begin{aligned}\frac{\partial \ln L}{\partial r} &= \sum_{i=1}^{m}\left((y_{i1} - b'(\theta_{i1}))\frac{\partial \theta_{i1}}{\partial r} - \frac{\partial C(y_{i1},\phi)}{\partial r} + \sum_{j=2}^{n_i}\left((y_{ij} - b'(\theta^*_{ij}))\frac{\partial \theta^*_{ij}}{\partial r} - \frac{\partial C(y_{ij},\phi^*)}{\partial r}\right)\right) \\ &= \sum_{i=1}^{m}\left((y_{i1} - \mu_{i1})\frac{\partial \theta_{i1}}{\partial r} - \frac{\partial C(y_{i1},\phi)}{\partial r} + \sum_{j=2}^{n_i}\left((y_{ij} - \mu^*_{ij})\frac{\partial \theta^*_{ij}}{\partial r} - \frac{\partial C(y_{ij},\phi)}{\partial r}\right)\right)\end{aligned}$$

where

$$\begin{aligned}\frac{\partial \theta_{i1}}{\partial r} &= \frac{1}{r(1+r\mu_{i1})} \\ \frac{\partial C(y_{ij},\phi)}{\partial r} &= \begin{cases} 0 & \text{if } y_{ij} = 0 \\ \frac{1}{r^2}\sum_{k=0}^{y_{ij}-1}\frac{1}{\frac{1}{r}+k} & \text{if } y_{ij} \neq 0 \end{cases} \\ \frac{\partial \theta^*_{ij}}{\partial r} &= \frac{r\frac{\partial \mu^*_{ij}}{\partial r} + \mu^*_{ij}}{r\mu^*_{ij}(1+r\mu^*_{ij})}\end{aligned}$$

For $j = 2$,

$$\frac{\partial \mu^*_{i2}}{\partial r} = \frac{C_{i21}}{\sqrt{1-C^2_{i21}}}(y_{i1} - \mu_{i1})\frac{\sqrt{Var(Y_{i1})}\frac{\partial\sqrt{E(Var(Y_{i2}|Y_{i1}))}}{\partial r} - \sqrt{E(Var(Y_{i2}|Y_{i1}))}\frac{\partial\sqrt{Var(Y_{i1})}}{\partial r}}{Var(Y_{i1})}$$

with

$$\frac{\partial\sqrt{Var(Y_{i1})}}{\partial r} = \frac{\mu^2_{i1}}{2\sqrt{\mu_{i1} + r\mu^2_{i1}}}$$

And, for $j > 2$,

$$\frac{\partial \mu_{ij}^*}{\partial r} = C_{ijj-1}\frac{\sqrt{1-C_{ij-1j-2}^2}}{\sqrt{1-C_{ijj-1}^2}}(y_{ij-1}-\mu_{ij-1}) \\ \times \frac{\sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))}\frac{\partial\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}{\partial r} - \sqrt{E(Var(Y_{ij}|Y_{ij-1}))}\frac{\partial\sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))}}{\partial r}}{E(Var(Y_{ij-1}|Y_{ij-2}))}$$

with

$$\frac{\partial\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}{\partial r} = \frac{1}{2\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}\left(\mu_{ij}^2 + \mu_{ij}\frac{C_{ijj-1}^2}{1-C_{ijj-1}^2}\right)\left(1 - r\frac{C_{ijj-1}^2}{1-C_{ijj-1}^2}\right)^{-2}$$

3.2.4. Comparison of Models

To compare models with different assumed distribution or correlation structures we use the Likelihood Ratio test. In the standard Likelihood-Ratio test, the goal is to make a comparison of the alternative model and the null model. The null model has less parameters and will be contained in the alternative models (i.e., by fixing the free parameters of the alternative model to some value, the null model is obtained). Let Lik_{alt} and Lik_{null} be the likelihood of the alternative and null models, respectively. The test statistic is twice the difference of the log likelihoods.

$$D = 2 \times (\ln(\text{Lik}_{\text{alt}}) - \ln(\text{Lik}_{\text{null}}))$$

The distribution of D is χ^2 with degrees of freedom equal to the number of free parameters.

The Likelihood-Ratio test is very useful but can only be used in limited circumstances. As explained, the null model must be contained in the alternative model. This is true when testing whether the addition of covariates improves the fit of the data, assuming the same distribution and correlation assumptions. Suppose the null model has covariates (X_1, X_2) corresponding to coefficients $(\beta_0, \beta_1, \beta_2)$ and the alternative model has covariates (X_1, X_2, X_3) corresponding to coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$. Under this set-up, the null model is the same as the alternative model when $\beta_3 = 0$. Hence the Likelihood-Ratio test can be used to compare the fit of the models. The degrees of freedom is simply the number of additional parameters in the alternative model.

Another scenario is in comparing models with different correlation structure assumptions. Suppose the null model assumes the AR(1) correlation structure and the alternative model assumes the AD(1) correlation structure. Here, the null model is the same as the alternative model when $\alpha_1 = \cdots = \alpha_{n-1}$. I.e., when all the adjacent correlations in the alternative model are equal to each other, the null model is contained within the alternative model. Hence, the Likelihood-Ratio test can be used to compare models in this case. However, the same is not true in comparing a model assuming Markov correlation structure with AR(1) or AD(1): the models are not contained within each other.

Suppose the null model assumes the Poisson distribution and the alternative model assumes a Negative-Binomial distribution. In this case, we are not able use the standard Likelihood-Ratio test for our comparison because the Poisson distribution is on the boundary of the Negative-Binomial. I.e., as $r \to 0$ in the Negative-Binomial distribution, we obtain the Poisson Distribution. However, 0 does not exists in the parameter space of r. Therefore, the asymptotic distribution of the likelihood ratio is not the standard χ^2 with a degree of freedom that is the difference in parameters. Instead, we need to use the results from Chernoff (1954), who derived the asymptotic distribution for the likelihood ratio when the value of the parameter is a boundary point. In short, Chernoff (1954) showed the the asymptotic distribution is zero half the time and χ^2 with one degree of freedom the other half of the time. Therefore, the corresponding p-value to this test - the probability of observing a more extreme likelihood ratio - is simply half the probability that a χ^2_1 distribution is greater the the likelihood ratio.

3.3. Application

Here we demonstrate our approach on data provided by Thall and Vail (1990); they analyzed data from a crossover trial of 59 epileptics who were were randomized to receive placebo or the anti-epileptic drug progabide. The drug was being tested as an adjuvant, a treatment applied after chemotherapy. Here, and in Thall and Vail (1990), we analyze only the pre-crossover responses. Upon starting the trial, the number of seizures in the 8 weeks prior to starting was recorded as a baseline measurement. At 2-week intervals, the number of seizures occurring in the last 2 weeks was reported. There were 4 total post-randomization clinic visits. In addition to the seizure counts, the patients age was recorded.

Table 3.2 reports the descriptive statistics of the epileptics dataset. For each of the four clinical visits, the mean and variance of the seizure counts for the different treatment groups are listed. The mean and variance of the ages and the baseline seizure counts are listed as well. For each of the visits and treatment groups, the variance is much larger than the mean, suggesting overdispersion.

Let Y_{ij} be the seizure count for patient i at time period j. Let Base_i be the baseline seizure count for patient i. Let Age_i be the age of patient i. Let Trt_i be the treatment assignment for patient i: 0 for placebo and 1 for progabide.

We assumed

$$\mu_{ij} = \exp(\beta_0 + \beta_1 \cdot \mathsf{Trt}_i + \beta_2 \cdot \mathsf{Base}_i + \beta_3 \cdot \mathsf{Age}_i + \beta_4 \cdot j)$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ are the coefficient values. We considered two distributions: i) Poisson, and ii) Negative-Binomial. We considered three structures for the correlation structure: i) AR(1), ii) Markov, and iii) AD(1).

After fitting our regression methodology to the data, we calculate several goodness-of-fit statistics, β coefficient estimates and their p-values, α correlation parameter estimates. For Negative-Binomial, the ancillary parameter r is also estimated.

The goodness-of-fit statistics provided are the log likelihood, the AIC, and the BIC. The log likelihood is simply the value of the objective function that the optimizing algorithm converges to. A higher log-likelihood indicates a better fitting model. The AIC and BIC are derived from the log-likelihood as

$$\mathsf{AIC} = 2 \cdot (\text{Number of Variables} + 1) - 2 \cdot \text{Log-Likelihood}$$

$$\mathsf{BIC} = \log(\text{Number of Subjects}) \cdot (\text{Number of Variables} + 1) - 2 \cdot \text{Log-Likelihood}$$

The AIC and BIC penalize models for having complex models with a large number of parameters. A lower AIC and BIC indicate a better fitting model.

The p-value is computed by utilizing the Hessian matrix which is provided as output in the optimization software. As explained in Section 3.2.2, the covariance is equal to the inverse of the Hessian

matrix. The Wald Statistic, with a null hypothesis that $\beta = 0$, is simply the square of the estimate divided by the corresponding variance. The p-value is the probability of a more extreme Wald Statistic with 1 degree of freedom.

We also fit GEE to the same model. Again, we have models for which the distribution is either Poisson or Negative-Binomial. For the correlation structure assumption, we assume Independence, AR(1), and Exchangeable correlation structures. In the Negative-Binomial cases, the ancillary parameter r must be specified. Here, we specify r as the estimate of r found in the Likelihood-based approach under an AR(1) correlation structure. We note that r is the ancillary parameter of the conditional distributions, which may not hold for the marginal distributions that are modeled in GEE.

3.3.1. Poisson

We first fit the epilepsy dataset to a model assuming a Poisson distribution. Table 3.3 reports the goodness-of-fit statistics, coefficients estimates, standard errors, Wald statistics, and p-values under each of the correlation structures. We also fit a model with an interaction between treatment and period (results not shown); the interaction term coefficient estimate had a p-value that was not statistically significant (p-value $>$ 0.18 in each correlation structure assumption).

Across each correlation structure case, all covariates had statistically significant associations. The AR(1) and Markov correlation structure cases had the same results due to the time periods being equally spaced. The coefficient estimates under the AD(1) cases were within 0.05 of the AR(1)/Markov coefficient estimates, suggesting the coefficient estimates are robust to the correlation structure assumption.

The coefficient estimates can be interpreted as marginal effects. Examining the coefficient estimates for AR(1), we find that patients had, on average

- a decrease of 0.170 in the log of the expected number of seizures for patients on progabide;
- an increase of 0.023 in the log of the expected number of seizures for each additional seizure in baseline seizure count;
- an increase of 0.022 in the log of the expected number of seizures for each additional year in

age;

- a decrease of 0.064 in the log of the expected number of seizures for an the subsequent clinical visit.

The coefficient estimates for the Markov and AD(1) correlation structure cases are interpreted in the same way.

Each correlation structure case had a positive correlation parameter. The AD(1) correlation structure parameter estimate, $\hat{\alpha} = (0.281, 0.620, 0.363)$ suggests the adjacent correlation structure differs across time. For example, the adjacent correlation parameter estimate between visits 2 and 3 (0.620) is more than twice that of the adjacent correlation parameter estimate between visits 1 and 2 (0.281). The adjacent correlation parameter estimates between visits 1 and 2 and between visits 3 and 4 have overlapping 95% confidence intervals (not shown).

By all goodness-of-fit statistics, the AD(1) was the best fitting of the correlation structure cases. It had the highest Log-Likelihood, lowest AIC, and lowest BIC. This difference was statistically significant when evaluating a likelihood ratio test (p-value $<$ 0.001).

By comparison, in the GEE analysis, only the coefficient estimate for baseline seizure count was found to be statistically significant. In addition, the Independence case had the Intercept coefficient statistically different from 0; the AR(1) case had the Age coefficient statistically different from 0; and the Exchangeable case had the Intercept and Age coefficients statistically different from 0. In terms of the magnitude and direction of the coefficient estimates, the GEE coefficients were very similar to those found in the Likelihood-based approach. The directions were all consistent and the difference in magnitude was up to 0.12. The estimates of α were similar to the Likelihood-based approach, with 0.510 for AR(1) and 0.399 for Exchangeable correlation structures.

3.3.2. Negative Binomial

Next we fit the epilepsy dataset to a model assuming a Negative-Binomial distribution. Table 3.4 reports the goodness-of-fit statistics, coefficients estimates, standard errors, Wald statistics, and p-values under each of the correlation structures. We also fit a model with an interaction between treatment and period (results not shown); the interaction term coefficient estimate had a p-value that was not statistically significant (p-value $>$ 0.48 in each correlation structure assumption).

Not all covariates had statistically significant associations. AR(1) and Markov correlation structure cases only had statistically significant coefficient estimates for the Intercept and Baseline. The AR(1) and Markov correlation structure cases had the same results due to the time periods being equally spaced. The AD(1) correlation structure case had statistically significant results for all covariates except Age and Period. The coefficient estimates under the AD(1) cases were within 0.08 of the AR(1)/Markov coefficient estimates, suggesting the coefficient estimates are robust to the correlation structure assumption.

The coefficient estimates can be interpreted as marginal effects. Examining the coefficient estimates for AR(1), we find that patients had, on average

- a decrease of 0.236 in the log of the expected number of seizures for patients on progabide, though this was not statistically significant;
- an increase of 0.026 in the log of the expected number of seizures for each additional seizure in baseline seizure count;
- an increase of 0.015 in the log of the expected number of seizures for each additional year in age, though this was not statistically significant;
- a decrease of 0.048 in the log of the expected number of seizures for an the subsequent clinical visit, though this was not statistically significant.

The coefficient estimates for the Markov and AD(1) correlation structure cases are interpreted in the same way.

Each correlation structure case had a positive correlation parameter. The AD(1) correlation structure parameter estimate, $\hat{\alpha} = (0.265, 0.536, 0.357)$ suggests the adjacent correlation structure differs across time. For example, the adjacent correlation parameter estimate between visits 2 and 3 (0.536) is more than twice that of the adjacent correlation parameter estimate between visits 1 and 2 (0.265). The adjacent correlation parameter estimates between visits 1 and 2 and between visits 3 and 4 have overlapping 95% confidence intervals (not shown).

The estimates for the ancillary parameter r were similar across all correlation structure cases. AD(1)/Markov had an estimate of 0.311 whereas AD(1) had an estimate of 0.293. As r is a measure

of overdispersion, these estimates suggest there is some overdispersion.

By all goodness-of-fit statistics, the AD(1) was the best fitting of the correlation structure cases. It had the highest Log-Likelihood, lowest AIC, and lowest BIC. This difference was statistically significant when evaluating a likelihood ratio test (p-value $=$ 0.016).

By comparison, in the GEE analysis, only the coefficient estimate for baseline seizure count was found to be statistically significant. In addition, the Independence case had the Intercept coefficient statistically different from 0; the AR(1) case had the Age coefficient statistically different from 0; and the Exchangeable case had the Intercept coefficient statistically different from 0. In terms of the magnitude and direction of the coefficient estimates, the GEE coefficients were very similar to those found in the Likelihood-based approach. The directions were all consistent and the difference in magnitude was up to 0.24. The estimates of α were higher than the Likelihood-based approach, with 0.531 for AR(1) and 0.407 for Exchangeable correlation structures.

3.3.3. Comparisons of Results

Comparing the results between Poisson and Negative-Binomial assumptions, we find that the coefficient parameter estimates are at most 0.12 of each other and were always in the same direction. This suggests the analysis is robust to the distribution assumption.

Comparing the correlation parameter estimates, we find that the Negative-Binomial case had estimates that were deflated by 2-14%. This is likely caused by introduction of the ancillary parameter r, which put constraints on α.

Using the nonstandard Likelihood-Ratio test described in Section 3.2.4 to compare the Poisson and Negative-Binomial models under a AD(1) correlation assumption, we find that the Negative-Binomial model was the better-fitting model (p-value $<$ 0.001).

Hence, assuming a distribution of Negative-Binomial and a correlation structure of AD(1) led to the best fitting model. The coefficient estimates can be interpreted as marginal effects. Examining the coefficient estimates, we find that patients had, on average

- a decrease of 0.257 in the log of the expected number of seizures for patients on progabide;
- an increase of 0.025 in the log of the expected number of seizures for each additional seizure

in baseline seizure count;

- an increase of 0.013 in the log of the expected number of seizures for each additional year in age, though this was not statistically significant;
- a decrease of 0.053 in the log of the expected number of seizures for an the subsequent clinical visit, though this was not statistically significant.

The correlation structure parameter was estimated to be $\alpha = (0.265, 0.536, 0.357)$. This suggests the correlation is high between the second and third visits. The ancillary parameter was estimated to be $r = 0.293$, which suggests overdispersion.

The GEE models had similar coefficient estimates as the Likelihood-based approach. The estimates themselves had the same direction and the difference in magnitude was up to 0.24. However, many more coefficient estimates in the GEE model were not statistically different from 0. So the Likelihood-based approach was able to detect a statistically significant different in those parameters whereas the GEE methodology was more conservative.

3.4. Simulations

We use simulation to assess the characteristics of the estimators and to demonstrate the methodology. We simulate outcomes on the seizure count data discussed previously (Thall and Vail, 1990). In this dataset, the number of seizures was recorded for four two-week periods. A baseline measurement of eight weeks was recorded as well as the patients age. Patients were randomly assigned to two treatments groups: placebo and progabide.

Let Y_{ij} be the seizure count for patient i at time period j. Let Base_i be the baseline seizure count for patient i. Let Age_i be the age of patient i. Let Trt_i be the treatment assignment for patient i: 0 for placebo and 1 for progabide. There were 59 complete cases in the dataset. We change the time periods from (1, 2, 3, 4), corresponding to the two-week intervals, to (1, 2, 4, 8). This is done in order to have differences between AR(1) and Markov correlation structures.

We performed a simulation on the data by generating the bi-weekly seizure count outcomes. We created the seizure count outcomes by randomly drawing a value from one of the distributions described in Section 3.2.3. For $j = 1$, the distribution is defined by μ_{i1}. For $j > 1$, the distributions

are defined by the conditional mean (3.3):

$$E(Y_{ij}|Y_{ij-1}) = \mu_{ij} + C_{ijj-1}\frac{\sigma_{ij}}{\sigma_{ij-1}}(Y_{ij-1} - \mu_{ij-1}).$$

C_{ijj-1} is defined as the correlation between Y_{ij} and Y_{ij-1}. μ_{ij} is assumed to be:

$$\mu_{ij} = \exp(\beta_0 + \beta_1 \cdot \text{Trt}_i + \beta_2 \cdot \text{Base}_i + \beta_3 \cdot \text{Age}_i + \beta_4 \cdot j)$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ are the coefficient values. We set $\beta = (0.667, -0.169, 0.023, 0.022, -0.027)$. We considered two distributions: i) Poisson, and ii) Negative-Binomial. When the distribution is Negative-Binomial, we set $r = 0.310$. We considered three structures for the correlation structure: i) AR(1), ii) Markov, and iii) AD(1). We consider several different values for the correlation structure parameter α that depends on the correlation structure. For AR(1), we consider $\alpha \in \{0, 0.5\}$. For Markov, we consider $\alpha \in \{0.1, 0.5\}$. For AD(1), we consider $\alpha \in \{(0, 0.25, 0), (0, 0.5, 0.25)\}$. We consider varying sample sizes by multiplying the 59 patients and giving the copied patients unique identifiers.

After creating the artificial seizure count outcomes, we fit our regression methodology to the resulting dataset. We calculate the mean square error, percent bias, and the 95% confidence interval of the treatment coefficient estimate. In fitting our regression methodology, we choose different combinations of the fitted distribution and fitted correlation structure. If the true distribution is Poisson, then the fitted distribution is always Poisson. If the true distribution is Negative-Binomial, then the fitted distribution is either Poisson or Negative-Binomial. We consider three fitted correlation structures: i) AR(1), ii) Markov, and iii) AD(1).

Below, we lists the different factors from which we draw our simulation cases. Not all combinations are included in the simulations (e.g., the fitted distribution is only Poisson if the true distribution is Poisson). The Sample Size cases refer to the factor by which the sample size is increased: 1X has 59 patients; 2X has 118 patients; 3X has 177 patients.

1. True Distribution: Poisson, Negative-Binomial

2. True Correlation Structure: AR(1), Markov, AD(1)

3. True Correlation Parameter - Depends on True Correlation Structure:
    - If AR(1): $\alpha \in \{0, 0.5\}$
    - If Markov: $\alpha \in \{0.1, 0.5\}$
    - If AD(1): $\alpha \in \{(0, 0.5, 0.25), (0, 0.25, 0)\}$
4. Fitted Distribution: Poisson, Negative-Binomial
5. Fitted Correlation Structure: AR(1), Markov, AD(1)
6. Sample Size: 1X, 2X, 3X

We also used simulations to assess the performance of GEE on the same data. After simulating the seizure count outcomes in the same way, we apply GEE to the dataset and obtain the regression estimates. We assume the same cases as before except for the fitted correlation structure. For the fitted correlation structure, we assume independence. Because GEE regression estimates are robust to the fitted correlation structure, we chose the simplest one available. We note that, in testing a fitted correlation structure of AR(1), the GEE algorithm did not always converge (not shown). When the true distribution is Negative-Binomial, we assume the ancillary parameter r is 0.310. We use the R package geeM (McDaniel and Henderson, 2015), since Negative-Binomial is not an option in many other R GEE packages.

For each case, we simulated 100 data sets. Tables 3.7, 3.8, 3.9, 3.10, 3.11, and 3.12 shows the mean square error, average percent bias, and the coverage probability of the 95% confidence interval for the treatment coefficient estimate under each of the cases examined.

Table 3.7 has the simulation statistics under which the true distribution is Poisson and the fitted distribution is Poisson. First, we report the results when the true correlation structure is AR(1). The MSE ranged from 0.666×10^{-3} to 5.467×10^{-3}; the average percent bias ranged from 0.131% to 10.583%; and the coverage probability ranged from 0.848 to 0.970. Having a true correlation parameter of 0.5 as compared to 0 lead to the MSE increasing by a factor of 2-4, and no discernible pattern in either the average percent bias or coverage probability. When the true correlation parameter is 0, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is

0.5, there was no discernible patterns in the MSE, average percent bias, or coverage probability, in comparing the fitted correlation structures. As the sample size increased, the MSE decreased, the average percent bias had no noticeable pattern, and the coverage probability had no noticeable pattern. Next, we report the results when the true correlation structure is Markov. The MSE ranged from 0.736×10^{-3} to 3.603×10^{-3}; the average percent bias ranged from 0.104% to 5.541%; and the coverage probability ranged from 0.900 to 0.970. Having a true correlation parameter of 0.5 as compared to 0.1 lead to the MSE increasing by up to a factor of 2, no discernible pattern in the average percent bias, and no discernible pattern in the coverage probability. When the true correlation parameter is .1, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is 0.5, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. As the sample size increased the coverage probability tended to increase; there were no noticeable patterns in the MSE or the average percent bias. Finally, we report the results when the true correlation structure is AD(1). The MSE ranged from 0.869×10^{-3} to 4.018×10^{-3}; the average percent bias ranged from 0.513% to 5.872%; and the coverage probability ranged from 0.860 to 0.950. Having a true correlation parameter of (0, 0.5, 0.25) as compared to (0, 0.25, 0) lead to the MSE increasing by up to a factor of 2, no discernible pattern in the average percent bias, and no discernible pattern in the coverage probability. When the true correlation parameter is (0, 0.25, 0), there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is (0, 0.5, 0.25), there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. As the sample size increased, the MSE decreased, there was no discernible pattern in the average percent bias, and the coverage probability had no discernible pattern.

Table 3.8 has the simulation statistics under which the true distribution is Negative-Binomial and the fitted distribution is Poisson. First, we report the results for when the true correlation structure is AR(1). The MSE ranged from 4.191×10^{-3} to 37.113×10^{-3}; the average percent bias ranged from 0.325% to 8.986%; and the coverage probability ranged from 0.424 to 0.650. Having a true correlation parameter of 0.5 as compared to 0 led to an increase in the MSE by up to a factor of 3, the average percent bias had no discernible pattern, and the coverage probability had no discernible pattern. When the true correlation parameter is 0, there was no noticeable patterns in the MSE,

average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is 0.5, having a Markov fitted correlation structure led to an increase in MSE, no discernible change in average percent bias, and a decrease in coverage probability, as compared to having a AR(1) fitted correlation structure. When the true correlation parameter is 0.5, having a AD(1) fitted correlation structure had no discernible differences compared to AR(1) in MSE, average percent bias, or coverage probability. As the sample size increased, the MSE decreased, while there was no discernible patterns in average percent bias or coverage probability. Next, we report the results for when the true correlation structure is Markov. The MSE ranged from 4.510×10^{-3} to 23.333×10^{-3}; the average percent bias ranged from 0.402% to 11.910%; and the coverage probability ranged from 0.530 to 0.640. Having a true correlation parameter of 0.5 as compared to 0.1 tended to lead to the MSE increasing by up to a factor of 3, no discernible pattern in the average percent bias or coverage probability. When the true correlation parameter is .1, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is 0.5, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. As the sample size increased the MSE decreased, while there was no discernible difference in average percent bias or coverage probability. Finally, we report the results for when the true correlation structure is AD(1). The MSE ranged from 3.678×10^{-3} to 22.869×10^{-3}; the average percent bias ranged from 0.763% to 17.848%; and the coverage probability ranged from 0.480 to 0.640. Having a true correlation parameter of (0, 0.5, 0.25) as compared to (0, 0.25, 0) led to an increase in the MSE by up to a factor of 2, and tended to lead to decreases in average percent bias and the coverage probability. When the true correlation parameter is (0, 0.25, 0), there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is (0, 0.5, 0.25), there was no discernible pattern in the MSE, average percent bias, or the coverage probability, in comparing the fitted correlation structures. As the sample size increased, MSE decreased, the average percent bias had no discernible pattern, and there was no discernible pattern in the coverage probability.

Table 3.9 has the simulation statistics under which the true distribution is Negative-Binomial and the fitted distribution is Negative-Binomial. First, we report the results for when the true correlation structure is AR(1). The MSE ranged from 2.649×10^{-3} to 21.800×10^{-3}; the average percent bias ranged from 0.059% to 18.254%; and the coverage probability ranged from 0.880 to 0.990. Having

a true correlation parameter of 0.5 as compared to 0 lead to the MSE increasing by up to a factor of 4, the average percent bias had no discernible pattern, and the coverage probability had no discernible pattern. When the true correlation parameter is 0, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is 0.5, there was no discernible pattern in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. As the sample size increased, the MSE decreased, the average percent bias had no discernible pattern, and the coverage probability had no discernible pattern. Next, we report the results for when the true correlation structure is Markov. The MSE ranged from 2.992×10^{-3} to 18.593×10^{-3}; the average percent bias ranged from 0.290% to 14.089%; and the coverage probability ranged from 0.900 to 0.970. Having a true correlation parameter of 0.5 as compared to 0.1 lead to the MSE increasing by up to a factor of 3, no discernible pattern in the average percent bias, and no discernible change for coverage probability. When the true correlation parameter is .1, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is 0.5, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. As the sample size increased the MSE tended to decrease, though there were no discernible patterns in average percent bias or coverage probability. Finally, we report the results for when the true correlation structure is AD(1). The MSE ranged from 2.959×10^{-3} to 22.993×10^{-3}; the average percent bias ranged from 0.294% to 7.116%; and the coverage probability ranged from 0.840 to 0.980. Having a true correlation parameter of (0, 0.5, 0.25) as compared to (0, 0.25, 0) lead to the MSE increasing by up to a factor of 3, no discernible pattern in the average percent bias, and no discernible difference in the coverage probability. When the true correlation parameter is (0, 0.25, 0), there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is (0, 0.5, 0.25), there was no discernible patterns in the MSE, average percent bias, or converage probability in comparing the fitted correlation structures. As the sample size increased, MSE decreased, there was no discernible pattern in the average percent bias, and there was no discernible pattern in the coverage probability.

Table 3.10 has the GEE simulation statistics under which the true distribution is Poisson and the fitted distribution is Poisson. The MSE ranged from 0.754×10^{-3} to 5.556×10^{-3}; the percent bias

ranged from 0.058 to 9.050; and the coverage probability ranged from 0.860 to 0.980. There was no discernible pattern in how well the GEE model did across the different true correlation structures. As the true correlation parameter increased in value, the MSE increased, the percent bias had no discernible pattern, and there was no discernible pattern in the coverage probability. As the sample size increased the MSE decreased, the percent biased had no discernible pattern, and there was no discernible pattern in the coverage probability.

Table 3.11 has the GEE simulation statistics under which the true distribution is Negative-Binomial and the fitted distribution is Poisson. The MSE ranged from 4.297×10^{-3} to 39.937×10^{-3}; the percent bias ranged from 0.044 to 24.322; and the coverage probability ranged from 0.860 to 0.960. There was no discernible pattern in how well the GEE model did across the different true correlation structures. As the true correlation parameter increased in value, the MSE increased, the percent bias had no discernible pattern, and there was no discernible pattern in the coverage probability. As the sample size increased the MSE decreased, the percent biased had no discernible pattern, and there was no discernible pattern in the coverage probability.

Table 3.12 has the GEE simulation statistics under which the true distribution is Negative-Binomial and the fitted distribution is Negative-Binomial. The MSE ranged from 2.387×10^{-3} to 23.382×10^{-3}; the percent bias ranged from 0.481 to 11.507; and the coverage probability ranged from 0.900 to 0.970. There was no discernible pattern in how well the GEE model did across the different true correlation structures. As the true correlation parameter increased in value, the MSE tended to increase, the percent bias no discernible pattern, and there was no discernible pattern in the coverage probability. As the sample size increased the MSE decreased, the percent biased tended to decrease, and there was no discernible pattern in the coverage probability.

The model performed well when both the true distribution and fitted distribution was Poisson. The MSE was consistently low, the average percent bias never went higher than 11%, and the coverage probability was consistently near 0.95. There was no consistency in how the model did when the fitted correlation structure was misspecified. By comparison, the GEE had similar results. The MSE, percent bias, and coverage probability had similar values and patterns as the likelihood method.

As expected, the results were poor when the true distribution was Negative-Binomial and the fitted distribution was Poisson. The model failed to capture the overdispersion and this is seen in the

MSE, average percent bias, and coverage probability. The MSE was larger, relative to the other scenarios. The average percent bias was consistently high in the 20 and 30%. The coverage probability was consistently low, rarely going above 0.7. The model performed better when the correlation parameter was higher, which is consistent across the scenarios. There was no consistency in how the model did when the fitted correlation structure was misspecified. By comparison, the GEE model did well. The MSE were low, the percent bias was low, and the coverage probability was high in comparison. However, the misspecification of the distribution led to poor results, compared to the correctly-specified GEE results.

The model performed well when both the true distribution and fitted distribution was Negative-Binomial. The MSE was consistently low, the average percent bias never went higher than 7%, and the coverage probability was consistently near 0.95. This suggests the model is effective in capturing overdispersion in the model. There was no consistency in how the model did when the fitted correlation structure was misspecified, suggesting the model is robust to the fitted correlation structure. By comparison, the GEE had similar results. The MSE, percent bias, and coverage probability had similar values and patterns as the likelihood method.

3.5. Discussion

We developed a maximum-likelihood based analysis that i) extends GLM for correlated discrete data, and ii) induces over-dispersion and plausible correlation structures for longitudinal data.

We demonstrated implementation of the approaches we developed in an analysis of Thall and Vail (1990). In our analysis, we found the model assuming a Negative-Binomial distribution and a AD(1) correlation structure to be the statistically-significant best fitting model. In this model, receiving the treatment progabide led to, on average, a decrease in the number of seizures. Having a higher number of seizures in the baseline count led to, on average, an increase in the number of seizures for any given visit. Age and the visit period were not statistically significant in the model.

Through simulations, we demonstrated that model performed well when both the true distribution and fitted distribution was Negative-Binomial, well when the true and fitted distributions was Poisson, and poor when the true distribution was Negative-Binomial and the fitted distribution was Poisson. Across each of these scenarios, the model tended to do better when the true correlation

parameter was higher. There was no consistency in how well the model did when the fitted correlation structure was different from the true correlation structure. The GEE did better when i) both the true and fitted distributions were Poisson; and ii) the true distribution is Negative-Binomial and the fitted distribution is Poisson. The GEE performed similarly as our likelihood method when both the true and fitted distribution was Negative-Binomial.

Our maximum-likelihood based analysis has a lot of attractive features. Having a log-likelihood allows us to assess the fit of competing models and construct likelihood ratio tests. By contrast, GEE has no objective function, which complicates the process of comparing competing models and assessing goodness-of-fit. Our maximum-likelihood based analysis requires the data to be missing at random for unbiased analysis whereas GEE requires missing completely at random for valid analysis. However, our maximum-likelihood approach may be less robust to misspecifying the model. Whereas GEE analysis will yield a consistent estimator of the regression parameters even when incorrectly specified.

Distribution	$E(Y_{ij}) = \mu_{ij}$	$Var(Y_{i1}) = \sigma^2_{i1}$	$Var(Y_{ij}) = \sigma^2_{ij}\ (j > 1)$
Poisson	$e^{x_{ij}^T\beta}$	μ_{i1}	$\frac{\mu_{ij}}{1 - C^2_{ijj-1}}$
Negative-Binomial	$e^{x_{ij}^T\beta}$	$\mu_{i1} + r\mu^2_{i1}$	$\frac{(\mu_{ij} + r\mu^2_{ij})\left(1 - r\frac{C^2_{ijj-1}}{1-C^2_{ijj-1}}\right)^{-1}}{1 - C^2_{ijj-1}}$
Binomial	$l_ij\frac{e^{x_{ij}^T\beta}}{1 + e^{x_{ij}^T\beta}}$	$\frac{\mu_{i1}}{l_{i1}}(l_{i1} - \mu_{i1})$	$\frac{\frac{\mu_{ij}}{l_{ij}}(l_{ij} - \mu_{ij})\left(1 + \frac{1}{l_{ij}}\frac{C^2_{ijj-1}}{1-C^2_{ijj-1}}\right)^{-1}}{1 - C^2_{ijj-1}}$

Table 3.1: Marginal means and variances for different assumed distributions. The assumed distributions are for Y_{i1} and for the conditional distribution of Y_{ij} given Y_{ij-1} $(j > 1)$.

	Variable	Placebo ($n = 28$)	Progabide ($n = 31$)	Total ($n = 59$)
Seizure Count	Visit 1	9.36 (102.8)	8.58 (332.7)	8.95 (220.1)
	Visit 2	8.29 (66.67)	8.42 (140.7)	8.36 (103.8)
	Visit 3	8.71 (213.3)	8.13 (193.1)	8.41 (199.3)
	Visit 4	7.96 (58.2)	6.71 (126.9)	7.31 (93.1)
	Age	29.00 (35.0)	27.74 (42.5)	28.34 (39.2)
	Baseline	30.79 (663.0)	31.61 (763.9)	31.22 (713.2)

Table 3.2: Mean and variance across treatment groups and periods for the sample population and for each combination of assumptions.

AR(1)	Log-Likelihood	-777.169		
	AIC	1566.338		
	BIC	1578.804		
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.718	0.195	13.590	<0.001
Treatment	-0.170	0.067	6.493	0.011
Baseline	0.023	0.001	1097.448	<0.001
Age	0.022	0.006	15.604	<0.001
Period	-0.064	0.021	8.760	0.003
α	0.416	0.033		
Markov	Log-Likelihood	-777.169		
	AIC	1566.338		
	BIC	1578.804		
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.718	0.195	13.590	<0.001
Treatment	-0.170	0.067	6.493	0.011
Baseline	0.023	0.001	1097.448	<0.001
Age	0.022	0.006	15.604	<0.001
Period	-0.064	0.021	8.760	0.003
α	0.416	0.033		
AD(1)	Log-Likelihood	-758.172		
	AIC	1528.344		
	BIC	1540.809		
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.765	0.193	15.780	<0.001
Treatment	-0.180	0.066	7.389	0.007
Baseline	0.023	0.001	1078.864	<0.001
Age	0.020	0.006	13.351	<0.001
Period	-0.060	0.021	8.142	0.004
α_1	0.281	0.062		
α_2	0.620	0.028		
α_3	0.363	0.058		

Table 3.3: Goodness of fit statistics, coefficient estimates, standard errors, Wald statistics, and p-values for analysis of seizure counts in epileptics (Thall and Vail, 1990) under each correlation structure in a Poisson distribution assumption.

AR(1)	Log-Likelihood	-637.722		
	AIC	1287.444		
	BIC	1299.909		
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.804	0.342	5.528	0.019
Treatment	-0.236	0.122	3.745	0.053
Baseline	0.026	0.002	144.587	$<$0.001
Age	0.015	0.010	2.263	0.132
Period	-0.048	0.043	1.233	0.267
α	0.374	0.060		
r	0.311	0.047		
Markov	Log-Likelihood	-637.722		
	AIC	1287.444		
	BIC	1299.909		
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.804	0.342	5.528	0.019
Treatment	-0.236	0.122	3.745	0.053
Baseline	0.026	0.002	144.587	$<$0.001
Age	0.015	0.010	2.263	0.132
Period	-0.048	0.043	1.233	0.267
α	0.374	0.060		
r	0.311	0.047		
AD(1)	Log-Likelihood	-634.305		
	AIC	1280.610		
	BIC	1293.075		
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.882	0.338	6.817	0.009
Treatment	-0.257	0.120	4.598	0.032
Baseline	0.025	0.002	140.835	$<$0.001
Age	0.013	0.010	1.939	0.164
Period	-0.053	0.042	1.549	0.213
α_1	0.265	0.103		
α_2	0.536	0.057		
α_3	0.357	0.110		
r	0.293	0.045		

Table 3.4: Goodness of fit statistics, coefficient estimates, standard errors, Wald statistics, and p-values for analysis of seizure counts in epileptics (Thall and Vail, 1990) under each correlation structure in a Negative-Binomial distribution assumption.

Independence				
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.719	0.344	2.092	0.036
Treatment	-0.152	0.171	-0.887	0.375
Baseline	0.023	0.001	18.448	0.000
Age	0.022	0.011	1.960	0.050
Period	-0.059	0.035	-1.682	0.093
α	0			
AR(1)				
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.601	0.348	1.727	0.084
Treatment	-0.163	0.160	-1.020	0.307
Baseline	0.023	0.001	18.742	0.000
Age	0.026	0.012	2.200	0.028
Period	-0.064	0.034	-1.888	0.059
α	0.510			
Exchangeable				
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.687	0.350	1.966	0.049
Treatment	-0.147	0.169	-0.872	0.383
Baseline	0.023	0.001	18.405	0.000
Age	0.023	0.012	1.995	0.046
Period	-0.059	0.035	-1.681	0.093
α	0.399			

Table 3.5: Coefficient estimates, robust standard errors, Wald statistics, and p-values for GEE analysis of seizure counts in epileptics (Thall and Vail, 1990) under several correlation structures in a Poisson distribution assumption.

Independence				
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.667	0.329	2.027	0.043
Treatment	-0.188	0.169	-1.112	0.266
Baseline	0.027	0.002	11.171	0.000
Age	0.018	0.010	1.762	0.078
Period	-0.045	0.034	-1.308	0.191
α	0	0.060		
AR(1)				
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.562	0.320	1.753	0.080
Treatment	-0.225	0.157	-1.427	0.153
Baseline	0.027	0.002	12.167	0.000
Age	0.021	0.010	2.076	0.038
Period	-0.042	0.033	-1.285	0.199
α	0.531	0.060		
Exchangeable				
Variable	Estimate	Std. Error	Wald	$Pr(>\|W\|)$
Intercept	0.662	0.330	2.007	0.045
Treatment	-0.187	0.169	-1.111	0.267
Baseline	0.026	0.002	11.201	0.000
Age	0.018	0.010	1.782	0.075
Period	-0.045	0.034	-1.306	0.191
α	0.407	0.103		

Table 3.6: Coefficient estimates, robust standard errors, Wald statistics, and p-values for GEE analysis of seizure counts in epileptics (Thall and Vail, 1990) under several correlation structure in a Negative-Binomial distribution assumption.

True Correlation Structure	True Correlation Parameter	Fitted Correlation Structure	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias (%)	Coverage Probability
AR(1)	0.0	AR(1)	1X	2.104	0.658	0.920
AR(1)	0.0	AR(1)	2X	0.990	0.457	0.940
AR(1)	0.0	AR(1)	3X	0.825	2.088	0.920
AR(1)	0.0	Markov	1X	2.539	3.347	0.950
AR(1)	0.0	Markov	2X	1.186	4.221	0.960
AR(1)	0.0	Markov	3X	0.936	0.677	0.910
AR(1)	0.0	AD(1)	1X	2.369	3.147	0.910
AR(1)	0.0	AD(1)	2X	1.246	0.348	0.930
AR(1)	0.0	AD(1)	3X	0.666	4.190	0.960
AR(1)	0.5	AR(1)	1X	5.266	10.583	0.880
AR(1)	0.5	AR(1)	2X	3.305	6.581	0.900
AR(1)	0.5	AR(1)	3X	1.552	0.379	0.950
AR(1)	0.5	Markov	1X	5.467	9.308	0.848
AR(1)	0.5	Markov	2X	2.945	3.929	0.880
AR(1)	0.5	Markov	3X	2.066	0.131	0.860
AR(1)	0.5	AD(1)	1X	4.973	1.260	0.880
AR(1)	0.5	AD(1)	2X	2.698	6.144	0.970
AR(1)	0.5	AD(1)	3X	2.031	4.486	0.930
Markov	0.1	AR(1)	1X	2.630	0.364	0.940
Markov	0.1	AR(1)	2X	1.524	1.690	0.940
Markov	0.1	AR(1)	3X	0.756	1.411	0.950
Markov	0.1	Markov	1X	2.453	3.834	0.960
Markov	0.1	Markov	2X	1.177	0.478	0.960
Markov	0.1	Markov	3X	0.825	3.128	0.960
Markov	0.1	AD(1)	1X	2.606	0.938	0.920
Markov	0.1	AD(1)	2X	1.263	2.326	0.920
Markov	0.1	AD(1)	3X	0.736	0.104	0.930
Markov	0.5	AR(1)	1X	3.603	5.541	0.920
Markov	0.5	AR(1)	2X	2.071	1.288	0.940
Markov	0.5	AR(1)	3X	1.213	0.582	0.960
Markov	0.5	Markov	1X	2.675	1.111	0.940
Markov	0.5	Markov	2X	2.197	1.131	0.900
Markov	0.5	Markov	3X	1.153	1.231	0.960
Markov	0.5	AD(1)	1X	3.564	2.275	0.910
Markov	0.5	AD(1)	2X	1.791	1.705	0.940
Markov	0.5	AD(1)	3X	0.953	1.482	0.970

Table 3.7: Continued on next page

Continued from previous page

True Correlation Structure	True Correlation Parameter	Fitted Correlation Structure	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias (%)	Coverage Probability
AD(1)	(0, 0.25, 0)	AR(1)	1X	2.897	0.513	0.900
AD(1)	(0, 0.25, 0)	AR(1)	2X	1.211	1.360	0.940
AD(1)	(0, 0.25, 0)	AR(1)	3X	0.869	0.829	0.940
AD(1)	(0, 0.25, 0)	Markov	1X	2.715	2.959	0.940
AD(1)	(0, 0.25, 0)	Markov	2X	1.214	1.681	0.930
AD(1)	(0, 0.25, 0)	Markov	3X	0.969	2.881	0.930
AD(1)	(0, 0.25, 0)	AD(1)	1X	2.635	4.189	0.950
AD(1)	(0, 0.25, 0)	AD(1)	2X	1.325	2.388	0.910
AD(1)	(0, 0.25, 0)	AD(1)	3X	0.942	0.841	0.920
AD(1)	(0, 0.5, 0.25)	AR(1)	1X	3.511	1.099	0.900
AD(1)	(0, 0.5, 0.25)	AR(1)	2X	1.720	5.108	0.930
AD(1)	(0, 0.5, 0.25)	AR(1)	3X	1.181	3.446	0.930
AD(1)	(0, 0.5, 0.25)	Markov	1X	2.941	2.885	0.930
AD(1)	(0, 0.5, 0.25)	Markov	2X	1.890	0.703	0.860
AD(1)	(0, 0.5, 0.25)	Markov	3X	1.332	5.872	0.890
AD(1)	(0, 0.5, 0.25)	AD(1)	1X	4.018	1.402	0.910
AD(1)	(0, 0.5, 0.25)	AD(1)	2X	2.003	2.609	0.940
AD(1)	(0, 0.5, 0.25)	AD(1)	3X	1.133	1.493	0.930

Table 3.7: Simulation results of mean square error (MSE), absolute value of the average percent bias, and coverage probability of the treatment parameter coefficient for combinations of true correlation structure, true correlation parameter, fitted correlation structure, and sample size. Both the true distribution and the fitted distribution are Poisson.

True Correlation Structure	True Correlation Parameter	Fitted Correlation Structure	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias (%)	Coverage Probability
AR(1)	0.0	AR(1)	1X	12.918	1.427	0.520
AR(1)	0.0	AR(1)	2X	5.922	0.610	0.590
AR(1)	0.0	AR(1)	3X	4.191	1.181	0.600
AR(1)	0.0	Markov	1X	11.739	5.111	0.650
AR(1)	0.0	Markov	2X	6.504	4.018	0.550
AR(1)	0.0	Markov	3X	4.144	4.354	0.650
AR(1)	0.0	AD(1)	1X	14.355	6.516	0.600
AR(1)	0.0	AD(1)	2X	6.870	4.970	0.560
AR(1)	0.0	AD(1)	3X	5.302	1.301	0.570
AR(1)	0.5	AR(1)	1X	26.715	8.966	0.630
AR(1)	0.5	AR(1)	2X	13.047	0.325	0.560
AR(1)	0.5	AR(1)	3X	9.112	1.249	0.550
AR(1)	0.5	Markov	1X	30.023	8.986	0.424
AR(1)	0.5	Markov	2X	14.741	4.061	0.540
AR(1)	0.5	Markov	3X	10.665	4.850	0.530
AR(1)	0.5	AD(1)	1X	37.113	5.025	0.520
AR(1)	0.5	AD(1)	2X	12.874	5.958	0.590
AR(1)	0.5	AD(1)	3X	9.989	9.900	0.520
Markov	0.1	AR(1)	1X	18.026	1.918	0.550
Markov	0.1	AR(1)	2X	7.358	1.599	0.590
Markov	0.1	AR(1)	3X	4.510	1.423	0.600
Markov	0.1	Markov	1X	12.779	7.712	0.560
Markov	0.1	Markov	2X	8.292	5.710	0.530
Markov	0.1	Markov	3X	4.967	0.653	0.580
Markov	0.1	AD(1)	1X	11.945	0.534	0.610
Markov	0.1	AD(1)	2X	5.591	3.143	0.630
Markov	0.1	AD(1)	3X	5.687	9.504	0.540
Markov	0.5	AR(1)	1X	21.685	0.977	0.550
Markov	0.5	AR(1)	2X	11.055	11.910	0.530
Markov	0.5	AR(1)	3X	7.243	0.260	0.590
Markov	0.5	Markov	1X	23.333	1.987	0.550
Markov	0.5	Markov	2X	8.714	0.402	0.590
Markov	0.5	Markov	3X	8.517	1.347	0.540
Markov	0.5	AD(1)	1X	20.580	1.388	0.600
Markov	0.5	AD(1)	2X	14.417	9.454	0.640
Markov	0.5	AD(1)	3X	7.267	6.441	0.500

Table 3.8: Continued on next page

Continued from previous page

True Correlation Structure	True Correlation Parameter	Fitted Correlation Structure	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias (%)	Coverage Probability
AD(1)	(0, 0.25, 0)	AR(1)	1X	14.973	12.593	0.600
AD(1)	(0, 0.25, 0)	AR(1)	2X	9.104	0.879	0.590
AD(1)	(0, 0.25, 0)	AR(1)	3X	4.515	2.650	0.600
AD(1)	(0, 0.25, 0)	Markov	1X	18.089	15.005	0.550
AD(1)	(0, 0.25, 0)	Markov	2X	7.993	3.554	0.490
AD(1)	(0, 0.25, 0)	Markov	3X	4.426	0.763	0.610
AD(1)	(0, 0.25, 0)	AD(1)	1X	21.073	0.882	0.550
AD(1)	(0, 0.25, 0)	AD(1)	2X	7.714	6.740	0.560
AD(1)	(0, 0.25, 0)	AD(1)	3X	3.678	9.378	0.640
AD(1)	(0, 0.5, 0.25)	AR(1)	1X	20.929	3.240	0.560
AD(1)	(0, 0.5, 0.25)	AR(1)	2X	14.097	5.382	0.540
AD(1)	(0, 0.5, 0.25)	AR(1)	3X	6.978	1.840	0.590
AD(1)	(0, 0.5, 0.25)	Markov	1X	22.869	6.297	0.520
AD(1)	(0, 0.5, 0.25)	Markov	2X	9.661	1.531	0.500
AD(1)	(0, 0.5, 0.25)	Markov	3X	7.103	0.951	0.500
AD(1)	(0, 0.5, 0.25)	AD(1)	1X	21.594	17.848	0.480
AD(1)	(0, 0.5, 0.25)	AD(1)	2X	9.182	8.603	0.540
AD(1)	(0, 0.5, 0.25)	AD(1)	3X	7.190	3.117	0.480

Table 3.8: Simulation results of mean square error (MSE), absolute value of the average percent bias, and coverage probability of the treatment parameter coefficient for combinations of true correlation structure, true correlation parameter, fitted correlation structure, and sample size. The true distribution is Negative-Binomial and fitted distribution is Poisson.

True Correlation Structure	True Correlation Parameter	Fitted Correlation Structure	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias (%)	Coverage Probability
AR(1)	0.0	AR(1)	1X	6.915	0.059	0.970
AR(1)	0.0	AR(1)	2X	4.308	4.676	0.930
AR(1)	0.0	AR(1)	3X	2.707	4.479	0.970
AR(1)	0.0	Markov	1X	7.992	12.356	0.940
AR(1)	0.0	Markov	2X	3.838	2.690	0.960
AR(1)	0.0	Markov	3X	2.740	2.142	0.950
AR(1)	0.0	AD(1)	1X	10.602	3.501	0.920
AR(1)	0.0	AD(1)	2X	4.199	3.009	0.950
AR(1)	0.0	AD(1)	3X	2.649	0.089	0.930
AR(1)	0.5	AR(1)	1X	19.786	2.438	0.920
AR(1)	0.5	AR(1)	2X	7.287	4.381	0.990
AR(1)	0.5	AR(1)	3X	6.296	5.688	0.930
AR(1)	0.5	Markov	1X	21.800	6.906	0.880
AR(1)	0.5	Markov	2X	12.125	18.254	0.890
AR(1)	0.5	Markov	3X	6.338	0.782	0.910
AR(1)	0.5	AD(1)	1X	18.089	7.660	0.950
AR(1)	0.5	AD(1)	2X	11.339	4.847	0.930
AR(1)	0.5	AD(1)	3X	4.273	4.180	0.970
Markov	0.1	AR(1)	1X	9.198	0.864	0.950
Markov	0.1	AR(1)	2X	5.021	5.124	0.950
Markov	0.1	AR(1)	3X	2.992	0.290	0.950
Markov	0.1	Markov	1X	6.856	2.326	0.970
Markov	0.1	Markov	2X	5.162	5.594	0.920
Markov	0.1	Markov	3X	4.017	1.980	0.920
Markov	0.1	AD(1)	1X	17.061	4.799	0.940
Markov	0.1	AD(1)	2X	3.415	1.477	0.960
Markov	0.1	AD(1)	3X	4.360	4.230	0.930
Markov	0.5	AR(1)	1X	15.247	2.999	0.930
Markov	0.5	AR(1)	2X	6.852	1.789	0.940
Markov	0.5	AR(1)	3X	4.877	2.009	0.900
Markov	0.5	Markov	1X	14.019	14.089	0.920
Markov	0.5	Markov	2X	5.662	3.442	0.940
Markov	0.5	Markov	3X	4.421	1.645	0.930
Markov	0.5	AD(1)	1X	18.593	6.754	0.920
Markov	0.5	AD(1)	2X	5.437	4.794	0.960
Markov	0.5	AD(1)	3X	3.478	1.809	0.940

Table 3.9: Continued on next page

Continued from previous page

True Correlation Structure	True Correlation Parameter	Fitted Correlation Structure	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias (%)	Coverage Probability
AD(1)	(0, 0.25, 0)	AR(1)	1X	8.423	4.913	0.950
AD(1)	(0, 0.25, 0)	AR(1)	2X	3.868	4.390	0.960
AD(1)	(0, 0.25, 0)	AR(1)	3X	2.964	3.009	0.950
AD(1)	(0, 0.25, 0)	Markov	1X	12.122	3.999	0.869
AD(1)	(0, 0.25, 0)	Markov	2X	5.245	1.895	0.900
AD(1)	(0, 0.25, 0)	Markov	3X	3.421	3.133	0.910
AD(1)	(0, 0.25, 0)	AD(1)	1X	9.041	4.144	0.930
AD(1)	(0, 0.25, 0)	AD(1)	2X	4.389	4.807	0.930
AD(1)	(0, 0.25, 0)	AD(1)	3X	2.959	4.380	0.980
AD(1)	(0, 0.5, 0.25)	AR(1)	1X	15.785	4.441	0.920
AD(1)	(0, 0.5, 0.25)	AR(1)	2X	9.485	1.229	0.890
AD(1)	(0, 0.5, 0.25)	AR(1)	3X	4.383	0.546	0.950
AD(1)	(0, 0.5, 0.25)	Markov	1X	18.370	5.409	0.840
AD(1)	(0, 0.5, 0.25)	Markov	2X	6.933	0.294	0.879
AD(1)	(0, 0.5, 0.25)	Markov	3X	6.419	2.097	0.800
AD(1)	(0, 0.5, 0.25)	AD(1)	1X	22.993	9.089	0.930
AD(1)	(0, 0.5, 0.25)	AD(1)	2X	4.366	7.116	0.980
AD(1)	(0, 0.5, 0.25)	AD(1)	3X	3.651	2.645	0.960

Table 3.9: Simulation results of mean square error (MSE), absolute value of the average percent bias, and coverage probability of the treatment parameter coefficient for combinations of true correlation structure, true correlation parameter, fitted correlation structure, and sample size. Both the true distribution and fitted distribution are Negative-Binomial.

True Correlation Structure	True Correlation Parameter	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias	Coverage Probability
AR(1)	0.0	1X	2.151	2.541	0.910
AR(1)	0.0	2X	1.002	2.118	0.930
AR(1)	0.0	3X	0.818	0.492	0.930
AR(1)	0.5	1X	5.556	9.050	0.930
AR(1)	0.5	2X	3.629	4.962	0.890
AR(1)	0.5	3X	2.584	3.243	0.900
Markov	0.1	1X	2.132	0.108	0.960
Markov	0.1	2X	1.090	7.146	0.940
Markov	0.1	3X	0.797	0.880	0.930
Markov	0.5	1X	4.284	2.713	0.920
Markov	0.5	2X	2.629	1.757	0.860
Markov	0.5	3X	1.524	1.572	0.940
AD(1)	(0, 0.25, 0)	1X	2.224	0.408	0.920
AD(1)	(0, 0.25, 0)	2X	1.683	2.506	0.910
AD(1)	(0, 0.25, 0)	3X	0.754	1.512	0.980
AD(1)	(0, 0.5, 0.25)	1X	3.343	2.197	0.930
AD(1)	(0, 0.5, 0.25)	2X	2.723	6.750	0.880
AD(1)	(0, 0.5, 0.25)	3X	0.994	0.058	0.970

Table 3.10: GEE simulation results of mean square error (MSE), absolute value of the average percent bias, and coverage probability of the treatment parameter coefficient for combinations of true correlation structure, true correlation parameter, and sample size. Both the true distribution and fitted distribution are Poisson.

True Correlation Structure	True Correlation Parameter	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias	Coverage Probability
AR(1)	0.0	1X	10.851	5.588	0.920
AR(1)	0.0	2X	5.849	3.110	0.920
AR(1)	0.0	3X	4.297	4.013	0.950
AR(1)	0.5	1X	39.937	24.322	0.890
AR(1)	0.5	2X	19.703	8.126	0.860
AR(1)	0.5	3X	10.074	2.038	0.940
Markov	0.1	1X	12.796	9.486	0.920
Markov	0.1	2X	5.394	0.968	0.960
Markov	0.1	3X	4.749	7.320	0.950
Markov	0.5	1X	18.200	6.525	0.920
Markov	0.5	2X	11.735	8.350	0.900
Markov	0.5	3X	6.961	1.891	0.920
AD(1)	(0, 0.25, 0)	1X	17.888	1.912	0.880
AD(1)	(0, 0.25, 0)	2X	8.619	4.786	0.910
AD(1)	(0, 0.25, 0)	3X	5.952	0.044	0.910
AD(1)	(0, 0.5, 0.25)	1X	18.338	5.211	0.930
AD(1)	(0, 0.5, 0.25)	2X	7.302	3.606	0.950
AD(1)	(0, 0.5, 0.25)	3X	7.602	10.307	0.930

Table 3.11: GEE simulation results of mean square error (MSE), absolute value of the average percent bias, and coverage probability of the treatment parameter coefficient for combinations of true correlation structure, true correlation parameter, and sample size. The true distribution is Negative-Binomial and the fitted distribution is Poisson.

True Correlation Structure	True Correlation Parameter	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias	Coverage Probability
AR(1)	0.0	1X	7.418	0.481	0.970
AR(1)	0.0	2X	4.831	2.898	0.930
AR(1)	0.0	3X	2.387	1.234	0.970
AR(1)	0.5	1X	23.382	1.294	0.930
AR(1)	0.5	2X	10.384	6.055	0.970
AR(1)	0.5	3X	8.145	0.500	0.950
Markov	0.1	1X	9.217	4.067	0.910
Markov	0.1	2X	4.494	8.263	0.910
Markov	0.1	3X	2.440	3.433	0.960
Markov	0.5	1X	17.571	8.588	0.910
Markov	0.5	2X	8.632	11.507	0.900
Markov	0.5	3X	5.043	5.690	0.920
AD(1)	(0, 0.25, 0)	1X	8.407	2.981	0.960
AD(1)	(0, 0.25, 0)	2X	5.183	3.069	0.930
AD(1)	(0, 0.25, 0)	3X	3.455	2.215	0.910
AD(1)	(0, 0.5, 0.25)	1X	11.693	6.138	0.960
AD(1)	(0, 0.5, 0.25)	2X	7.044	4.250	0.930
AD(1)	(0, 0.5, 0.25)	3X	4.464	4.599	0.930

Table 3.12: GEE simulation results of mean square error (MSE), absolute value of the average percent bias, and coverage probability of the treatment parameter coefficient for combinations of true correlation structure, true correlation parameter, and sample size. Both the true distribution and fitted distribution are Negative-Binomial.

CHAPTER 4

THE FIRST ORDER MARKOV MAXIMUM LIKELIHOOD BASED APPROACH FOR ANALYSIS OF BINOMIAL TYPE VARIABLES

4.1. Introduction

Longitudinal binomial outcomes are often encountered in medical research. We investigate once such example in this paper. We explore the associations between hospital quality evaluations developed by U.S. News & World Report and the transplant quality evaluations developed by the Scientific Registry of Transplant Recipients (SRTR). Twice a year, the SRTR releases reports on how well a hospital's transplant program is doing for each organ category. The Center for Medicare & Medicaid Services (CMS) uses the SRTR reports to "flag" a hospital for review. We treat the number of "flags" that a hospital receives in a year, which can be up to 8, as a Binomial response. Likewise, the U.S. News & World Report releases "America's Best Hospitals," a ranking of all the hospitals in the United States. Included is the Honor Roll, a list of hospitals that are in the top 20 across at least 6 medical specialties. In this paper, we explore the association of the number of "flagged" SRTR reports with the status of a hospital being on the Honor Roll. We use the maximum-likelihood based methodology developed in Chapter 3, which we extend for Binomial outcome data.

The methodology developed in Chapter 3 makes several assumptions. First, they assume the first-order Markov property for outcomes within hospitals. That is to say, the value of an outcome for a hospital at a particular measurement only depends on the value at the immediate previous measurement. The methodology allows several different assumptions on the adjacent correlation. We assume adjacent correlations that induce three different types of correlation structures: AR(1), Markov, and AD(1). The AR(1) correlation structure assumes a decline in the correlations with increasing separation in time. The Markov correlation structure assumes the adjcent correlation depends on the separation in time. The AD(1) correlation structure assumes the adjacent correlation is unique to the time period.

There are many advantages and disadvantages to maximum-likelihood based analysis. An advantage is that it provides an objective function. This allows easy comparison between models

through the likelihood ratio test and to assess the goodness-of-fit of a model. Generalized estimating equations (GEE) have no objective function, which makes assessing these properties difficult. Maximum-likelihood based analysis requires the data to be missing at random whereas GEE requires the data to be missing completely at random (Liang and Zeger, 1986). However, maximum-likelihood based analysis may be less robust to mis-specification of the model. Whereass GEE yields a consistent estimator of the regression parameters even when the fited correlation structure is not the true correlation structure. This comes with a cost for GEE: there may be a loss in precision if the fitted and true correlation structures are not close (Diggle et al., 2002; Fitzmaurice, Laird, and Ware, 2011).

Guerra et al. (2012) showed that the maximum-likelihood based analysis of this type are robust to mis-specification of the true correlation structure. However, there are caveats with that statement. For example, the maximum-likelihood model cannot be the true model for exchangeable data, under which all pairwise correlations are constant. As long as the funcitonal forms for the off-diagonal elements of the assumed correlation structure are correctly specified, the analysis will be robust to mis-specification. With this in mind, the assumption will be correct if the true correlation structure is among exchangelable, tri-diagonal, AR(1), or identity. With that said, it is not to the same degree as GEE, which often yields consistent estimates of the regression parameter for any combination of true and fitted correlation structure. In addition, mis-specifying the correlation structure may cause a loss of efficiency in estimation of the regression parameter.

In this Chapter we explore associations between hospital quality evaluations developed by U.S. News & World Report and the transplant quality evaluations developed by the Scientific Registry of Transplant Recipients (SRTR). The Chapter is organized as follows. Section 4.2 discuss the data, with Sections 4.2.1 and 4.2.2 providing details on the source and Sections 4.2.3 providing descriptive statistics. In Section 4.3, we review the assumptions and likelihood that was derived in Chapter 3. We present an analysis of the U.S. News / SRTR dataset to demonstrate application of the methods in Section 4.4. Simulation results are presented in Section 4.5. Finally, discussion and concluding remarks are presented in Section 4.7.

4.2. SRTR / U.S. News Dataset

4.2.1. Scientific Registry of Transplant Recipients

The Scientific Registry of Transplant Recipients (SRTR) is a database of organ transplantation statistics in the United States that started in 1987. The registry was designed to provide useful information in the evaluation of solid organ transplantation, which can include kidney, heart, liver, lung, intestine, and pancreas. The data for organ transplantation in the United States is collected by the Organ Procurement and Transplantation Network (OPTN), which also manages the national transplant waiting list and matches the organ donors to recipients. The goal of the SRTR is to provide information relevant to evidence-based policy, provide analysis of transplant programs, and to support transplantation research.

The SRTR releases publicly available transplant program reports for each transplant center every six months that provide waiting time, organ availability, and survival statistics. Included in the report card is the number of observed and expected graft failures during the first year after transplant. The observed number of graft failures is simply the count for the cohort that corresponds to that particular report.

The number of expected graft failures is calculated based on the national data for donor recipients similar to those at a particular transplant program. A Cox proportional hazards regression model for time to graft failure was fit to nation-wide data. The covariates included various patient, donor, and transplant characteristics. The covariates included differ depending on the organ. The exact list of covariates used, and the resulting β estimates, can be found at http://www.srtr.org/csr/current/modtabs.aspx. From the model, we obtain $S_i(1)$, the probability of graft survival to 1 year for patient i with characteristics x_i. Supposing there are n patients, the expected number of graft failures would then be $\sum -\ln S_i(1)$.

The SRTR transplant program reports are used by the Centers for Medicare & Medicaid Services (CMS) in evaluating the effectiveness of hospital transplant programs. CMS has criteria for whether a transplant program is approved for Medicare or Medicaid. Let O be the observed number of graft failures and E be the expected number of graft failures. CMS will review a transplant program if all of the following three criteria are met:

1. the number of observed graft failures is 3 more than the expected number of graft failures ($O - E > 3$);

2. the number of observed graft failures is at least 50% more than the expected number of graft failures ($O/E > 1.50$);

3. the one-sided P value of the statistical hypothesis test that $O = E$ is less than 0.05 (one-sided P < 0.05).

The hypothesis that $O = E$ is tested using an exact Poisson test. If a transplant program meets these criteria, we say that the program is "flagged."

For our analysis, we exclude pediatric and Veteran's hospitals. We include transplant program reports for Kidney, Lung, Liver, and Heart in years 2012-2015. For a given year, we count the number of flagged reports for a given transplant program. Because there are two reports released each year, the number of flagged reports can be up to 8. Some transplant programs may not provide transplants for all the organs, so their maximum count may be less than 8.

4.2.2. US News & World Report

Beginning in 1990, the U.S. News & World Report introduced "America's Best Hospitals," in order to aid patients and families facing serious or complex medical problems in finding a hospital. The first year of the program took the form of an alphabetically ordered list of hospitals that were rated. After the first year, beginning in 1991, the hospitals were ranked. The rankings were developed to help patients determine the best hospitals for providing care for serious or complicated medical conditions. The data is provided on their website at www.usnews.com/besthospitals/rankings. The methodology is summarized below and can be found in more detail in Olmsted et al. (2015).

Each year, the U.S. News & World Report provides rankings on 16 different adult specialties which are based on data from several sources. The specialties examined over the years has changed; for example, AIDS care was removed in 1998 after it became clear that the care had shifted to outpatient settings.

The rankings for 12 of the 16 specialties are based on the Donabedian model of health care: structure, process, and outcomes (Donabedian, 1966). The structure refers to hospital resources

that are directly related to patient care such as nurse staffing and availability of technologies and patient services. The process refers to delivering care, including diagnosis, treatment, prevention, and patient education. The outcomes include death, harm to patients, incidence of preventable re-admissions, etc. The patient's condition and complexity are taken into account in measuring the outcomes. The U.S. News & World Report also include patient safety in calculating the rankings for each specialty. These include any complications that may compromise the safety of a patient. From these four components, a weighted score is calculated for each of the specialties at each hospital.

The structure component of a hospital is measured based on the AHA Annual Survey Database, as well as the Medicare Provider Analysis and Review (MedPAR) that is provided by CMS. The process component of a hospital is measured based on a hospital's reputation. The reputation is found by averaging the responses of the three most recent annual surveys of physicians. The surveyed physicians were asked to list up to five hospitals that they considered best for complicated conditions. The outcomes component of a hospital is measured based on the mortality 30 days after admission and is based on the MedPAR data. The patient safety component is measured based on the MedPAR data.

The rankings for the remaining 4 of the 16 specialties are based on a reputation survey alone. These specialties include ophthalmology, psychiatry, rehabilitation, and rheumatology. Because care for these specialties is largely outpatient and poses little risk of death, the structural and outcomes measures of the Donabedian model are not appropriate.

1,897 hospitals were eligible for at least 1 of the 12 score-driven specialties under the U.S. News & World Report criteria. For each specialty and each hospital the rankings are calculated. The Honor Roll hospitals are calculated by a point system. If a hospital is in the top 10 rankings in a specialty, they receive 2 points. If a hospital is in the next 10, 11-20, they receive 1 point. Any hospital with points in at least six specialties are included in the honor roll. They are ranked by the number of points they receive. For our analysis, we use an binary variable to indicate whether a hospital is in the Honor Roll or not, regardless of where they are ranked in the Honor Roll.

4.2.3. Descriptive Statistics

Table 4.1 lists across each year the number of hospitals, the number of hospitals ranked in the Honor Roll by the U.S. News & World Report, number of hospitals with a given number of transplant

reports, and the mean percentage of flagged reports for hospitals both in and not in the Honor Roll of the U.S. News & World Report. There were 220 hospitals that met the SRTR and US News & World Report inclusion/exclusion criteria; each year had 208-213 hospitals. For a given year, the number of hospitals ranked in the Honor Roll was between 7.2 and 8.5% of the hospitals included in the analysis. The number of transplant reports for a given transplant program ranged from 1-8 but was most concentrated on the even numbers. The mean percentage of flagged reports for hospitals not ranked in the U.S. News & World Report Honor Roll was 2.6 to 6.4 times the mean percentage of flagged reports for hospitals ranked in the U.S. News & World Report.

4.3. Methods

4.3.1. Assumptions and Likelihood

We assume the same notation and assumptions as Chapter 3, which is summarized here. See Appendix B for derivations.

We collect data on m hospitals. For the ith hospital, the number of flagged reports $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$ and the total number of reports $l_i = (l_{i1}, \cdots, l_{in_i})^T$ are collected at the corresponding years $T_i = (t_{i1}, \cdots, t_{in_i})'$. We note that, for each i and j, $Y_{ij} < l_{ij}$. At each year, the Ranked and Year data $x_{ij} = (x_{ij1}, , x_{ij2})$ are collected. Let $y_i = (y_{i1}, \cdots, y_{in_i})^T$ be a realization of Y_i.

By the assumptions made in Chapter 3, the distribution of $(Y_1, \cdots, Y_m)$ takes form

$$f(Y_1 = y_1, \cdots, Y_m = y_m) = \prod_{i=1}^{m} f(Y_{i1} = y_{i1}) \prod_{j=2}^{n_i} f(Y_{ij} = y_{ij} | Y_{ij-1} = y_{ij-1})$$

Furthermore, by assuming the expectation of $Y_{ij}|Y_{ij-1}$ is a linear function of Y_{ij-1} we have, for $j = 2, \cdots, n_i$,

$$\begin{aligned} E(Y_{ij}|Y_{ij-1}) &= \mu_{ij} + C_{ijj-1}\frac{\sigma_{ij}}{\sigma_{ij-1}}(Y_{ij-1} - \mu_{ij-1}) \\ &\equiv \mu^*_{ij} \end{aligned}$$

where

$$\sigma_{ij}^2 = \frac{E(Var(Y_{ij}|Y_{ij-1}))}{1 - C_{ijj-1}^2}$$

As mentioned previously, we assume Y_{i1} and $Y_{ij}|Y_{ij-1}$ are Binomial distributed with l_{ij} as the total number of trials. Taking u as a placeholder for μ_{i1} and μ_{ij}^*, the pdf is

$$\begin{aligned} f &= \binom{l_{ij}}{y_{ij}} \left(\frac{u_{ij}}{l_{ij}}\right)^{y_{ij}} \left(1 - \frac{u_{ij}}{l_{ij}}\right)^{l_{ij}-y_{ij}} \\ &= \exp\left(y_{ij} \ln\left(\frac{u_{ij}}{l_{ij}}\right) + (l_{ij} - y_{ij}) \ln\left(1 - \frac{u_{ij}}{l_{ij}}\right) + \ln\binom{l_{ij}}{y_{ij}}\right) \\ &= \exp\left(y_{ij} \ln\left(\frac{u_{ij}}{l_{ij} - u_{ij}}\right) + l_{ij} \ln\left(\frac{l_{ij} - u_{ij}}{l_{ij}}\right) + \ln\binom{l_{ij}}{y_{ij}}\right) \end{aligned}$$

4.3.2. Likelihood Equations

The likelihood of $(Y_i, \cdots, Y_m)$, described in Chapter 3 is

$$L(\beta, \alpha) = \prod_{i=1}^{m} \exp\left(\frac{y_{i1}\theta_{i1} - b(\theta_{i1})}{a(\phi)} - c(y_{i1}, \phi)\right) \prod_{j=2}^{n_i} \exp\left(\frac{y_{ij}\theta_{ij}^* - b(\theta_{ij}^*)}{a(\phi^*)} - c(y_{ij}, \phi^*)\right)$$

where $\theta_{i1} = g(\mu_{i1})$ and $\theta_{ij}^* = g(\mu_{ij}^*)$. $a()$, $b()$, and $c()$ are functions specific to the assumed distribution; ϕ is the dispersion parameter. $g()$ is the link function, which relates the linear predictor to the expected value of the data. Taking the natural log, we obtain

$$\ln(L(\beta, \alpha)) = \sum_{i=1}^{m} \left(\frac{y_{i1}\theta_{i1} - b(\theta_{i1})}{a(\phi)} - c(y_{i1}, \phi) + \sum_{j=2}^{n_i} \left(\frac{y_{ij}\theta_{ij}^* - b(\theta_{ij}^*)}{a(\phi^*)} - c(y_{ij}, \phi^*)\right)\right)$$

From here, we recognize the different components of the Binomial Distribution as

$$\begin{aligned} \theta_{ij} &= \ln\left(\frac{u_{ij}}{l_{ij} - u_{ij}}\right) \\ a(\phi) &= 1 \\ b(\theta_{ij}) &= l_{ij} \ln\left(\frac{l_{ij} - u_{ij}}{l_{ij}}\right) \\ c(y_{ij}, \phi) &= -\ln\binom{l_{ij}}{y_{ij}} \end{aligned}$$

$$
\begin{aligned}
g(\gamma) &= \ln\left(\frac{\gamma}{l_{ij}-\gamma}\right) \\
g'(\gamma) &= \frac{l_{ij}}{\gamma(l_{ij}-\gamma)} \\
\mu_{ij} &= l_{ij}\,\text{expit}(x_i'\beta) \\
Var(Y_{i1}) &= \frac{\mu_{ij}}{l_{ij}}(l_{ij}-\mu_{ij}) \\
\frac{\partial Var(Y_{i1})}{\partial\beta} &= \frac{\mu_{i1}\frac{-\partial\mu_{i1}}{\partial\beta} + (l_{ij}-\mu_{i1})\frac{\partial\mu_{i1}}{\partial\beta}}{l_{ij}} \\
&= \frac{l_{ij}-2\mu_{i1}}{l_{ij}}\frac{\partial\mu_{i1}}{\partial\beta}
\end{aligned}
$$

We note above that we use the canonical inverse logit-link function, $\mu_{ij} = l_{ij}\,\text{expit}(x_i'\beta)$, which is standard practice for Binomial regression. Furthermore, for $j > 1$,

$$
\begin{aligned}
E(Var(Y_{ij}|Y_{ij-1})) &= E\left(\frac{\mu_{ij}^*}{l_{ij}}(l_{ij}-\mu_{ij}^*)\right) \\
&= E(\mu_{ij}^*) - \frac{1}{l_{ij}}E((\mu_{ij}^*)^2) \\
&= \mu_{ij} - \frac{1}{l_{ij}}\left(\mu_{ij}^2 + \frac{C_{ijj-1}^2}{1-C_{ijj-1}^2}E(Var(Y_{ij}|Y_{ij-1}))\right)
\end{aligned}
$$

Solving for $E(Var(Y_{ij}|Y_{ij-1}))$, we have

$$
\begin{aligned}
E(Var(Y_{ij}|Y_{ij-1}))\left(1+\frac{1}{l_{ij}}\frac{C_{ijj-1}^2}{1-C_{ijj-1}^2}\right) &= \frac{\mu_{ij}}{l_{ij}}(l_{ij}-\mu_{ij}) \\
E(Var(Y_{ij}|Y_{ij-1})) &= \frac{\mu_{ij}}{l_{ij}}(l_{ij}-\mu_{ij})\left(1+\frac{1}{l_{ij}}\frac{C_{ijj-1}^2}{1-C_{ijj-1}^2}\right)^{-1} \\
\frac{\partial E(Var(Y_{ij}|Y_{ij-1}))}{\partial\beta} &= \frac{\mu_{ij}\frac{-\partial\mu_{ij}}{\partial\beta} + (l_{ij}-\mu_{ij})\frac{\partial\mu_{ij}}{\partial\beta}}{l_{ij}}\left(1+\frac{1}{l_{ij}}\frac{C_{ijj-1}^2}{1-C_{ijj-1}^2}\right)^{-1} \\
&= \frac{l_{ij}-2\mu_{ij}}{l_{ij}}\frac{\partial\mu_{ij}}{\partial\beta}\left(1+\frac{1}{l_{ij}}\frac{C_{ijj-1}^2}{1-C_{ijj-1}^2}\right)^{-1}
\end{aligned}
$$

Hence the marginal variance for $j > 1$ is $Var(Y_{ij}) = \dfrac{\frac{\mu_{ij}}{l_{ij}}(l_{ij}-\mu_{ij})\left(1+\frac{1}{l_{ij}}\frac{C_{ijj-1}^2}{1-C_{ijj-1}^2}\right)^{-1}}{1-C_{ijj-1}^2}$. Since $Var(Y_{ij}) > E(Y_{ij}) = \mu_{ij}$, it is clear the over dispersion is induced for the marginal distributions for $j > 1$.

We note that, for the Bernoulli case, we simply set $l_{ij} = 1$ to obtain the component functions.

General Form of Estimating Equations

The estimating equations are the same as those listed in Chapter 3:

With respect to β, we have

$$\frac{\partial \ln(L(\beta,\alpha))}{\partial \beta} = \sum_{i=1}^{m} \left(\frac{y_{i1} - \mu_{i1}}{a(\phi)} \frac{\partial g(\gamma)}{\partial \gamma}\Big|_{\gamma=\mu_{i1}} \frac{\partial \mu_{i1}}{\partial \beta} + \sum_{j=2}^{n_i} \frac{y_{ij} - \mu^*_{ij}}{a(\phi^*)} \frac{\partial g(\gamma)}{\partial \gamma}\Big|_{\gamma=\mu^*_{ij}} \frac{\partial \mu^*_{ij}}{\partial \beta} \right)$$

With respect to α, we have

$$\frac{\partial \ln(L(\beta,\alpha))}{\partial \alpha} = \sum_{i=1}^{m} \sum_{j=2}^{n_i} \left(\frac{y_{ij} - \mu^*_{ij}}{a(\phi^*)} \frac{\partial g(\gamma)}{\partial \gamma}\Big|_{\gamma=\mu^*_{ij}} \frac{\partial \mu^*_{ij}}{\partial \alpha} \right)$$

Details on the different components in each estimating equation is provided in Chapter 3.

Again, we assume three different correlation structures: AR(1), Markov, and AD(1). AR(1) holds when the correlation between adjacent measurements takes the same value regardless of the preceding value or time measurement. Markov holds when the correlation between adjacent measurements is dependent on the time between those measurements. AD(1) holds when the correlation between adjacent measurements on a subject is dependent on a parameter that is unique to the previous measurement. Further details are provided in Chapter 3. The code describing the algorithm that finds the maximum likelihood estimates is provided in Appendix C.

4.4. Analysis of Real-World Data

4.4.1. Analysis

Here we demonstrate our approach on the SRTR / U.S. News dataset. Let Y_{ij} be the number of flagged reports for the jth observation at hospital i. Let l_{ij} be the total number of reports for the jth observation at hospital i. Let Ranked_{ij} indicate that the jth observation at hospital i is ranked in the Honor Roll by the U.S. News & World Report. Let Year_{ij} be the Year of the jth observation for the ith hospital.

We assumed

$$\mu_{ij} = l_{ij} \operatorname{expit}(\beta_0 + \beta_1 \cdot \text{Ranked}_{ij} + \beta_2 \cdot \text{Year}_{ij})$$

where $\beta = (\beta_0, \beta_1, \beta_2)$ are the coefficient values. We also consider a model in which the covariate Year is not included. We considered three structures for the correlation structure: i) AR(1), ii) Markov, and iii) AD(1).

After fitting our regression methodology to the data, we calculate several goodness-of-fit statistics, β coefficient estimates and their p-values, α correlation parameter estimates.

The goodness-of-fit statistics provided are the log likelihood, the AIC, and the BIC. The log likelihood is simply the value of the objective function that the optimizing algorithm converges to. A higher log-likelihood indicates a better fitting model. The AIC and BIC are derived from the log-likelihood as

$$\text{AIC} = 2 \cdot (\text{Number of Variables} + 1) - 2 \cdot \text{Log-Likelihood}$$
$$\text{BIC} = \log(\text{Number of Subjects}) \cdot (\text{Number of Variables} + 1) - 2 \cdot \text{Log-Likelihood}$$

The AIC and BIC penalize models for having complex models with a large number of parameters. A lower AIC and BIC indicate a better fitting model.

The p-value is computed by utilizing the Hessian matrix which is provided as output in the optimization software. The Hessian matrix is a matrix of second derivatives of a function. In the context of log-likelihoods, the Hessian matrix is equal to the inverse of the covariance matrix. The Wald Statistic, with a null hypothesis that $\beta = 0$, is simply the square of the estimate divided by the corresponding variance. The p-value is the probability of a more extreme Wald Statistic with 1 degree of freedom.

We fit the SRTR / U.S. News dataset to a model assuming a Binomial distribution. Table 4.2 reports the goodness-of-fit statistics, coefficients estimates, standard errors, Wald statistics, and p-values under each of the correlation structures.

Across each correlation structure case, only Ranked had an association that was statistically sig-

nificant whereas Year did not have an association that was statistically significant. The AR(1) and Markov correlation structure cases had the same results due to the time periods being equally spaced. The coefficient for Ranked under the AD(1) case was -0.749 as opposed to -1.172 for the AR(1) and Markov cases. The coefficient for the Intercept under the AD(1) case was 157 636 as opposed to 25.860 for the AR(1) and Markov cases, though the differences from 0 was not statistically significant. The coefficient for the Year under the AD(1) was -0.080 as opposed to -0.014 for the AR(1) and Markov cases, though the differences from 0 was not statistically significant.

The coefficient estimates can be interpreted as marginal effects. Examining the coefficient estimates for AD(1), we find that patients had, on average

- a decrease of 0.749 in the logit of the expected number of flagged report cards for hospitals that were ranked in the Honor Roll by the U.S. News & World Report;
- a decrease of 0.080 in the logit of the expected number of flagged report cards for an increase by one year.

The coefficient estimates for the Markov and AD(1) correlation structure cases are interpreted in the same way.

Each correlation structure case had a positive correlation parameter. The AD(1) correlation structure parameter estimate, $\hat{\alpha} = (0.437, 0.552, 0.455)$ suggests the adjacent correlation structure is rather consistent across time; the adjacent correlation parameter estimates between the visits have overlapping 95% confidence intervals. The correlation parameter estimate for AR(1) / Markov was 0.473, which was similar to the AD(1) correlation parameter estimates.

By all goodness-of-fit statistics, the AD(1) was the best fitting of the correlation structure cases: it had the highest Log-Likelihood, lowest AIC, and lowest BIC. This difference was statistically significant when evaluating a likelihood ratio test (p-value = 0.038).

Table 4.3 reports the goodness-of-fit statistics, coefficients estimates, standard errors, Wald statistics, and p-values under each of the correlation structures for the model without Year as a covariate.

Across each correlation structure case, both the intercept and Ranked had coefficient that were statistically different from 0. Again, the AR(1) and Markov correlation structure cases had the same

results due to the time periods being equally spaced. The coefficient for Ranked under the AD(1) case was -0.784 as opposed to -0.838 for the AR(1) and Markov cases. The coefficient for the Intercept under the AD(1) case was -2.680 as opposed to -2.699 for the AR(1) and Markov cases.

The coefficient estimates can be interpreted as marginal effects. Examining the coefficient estimates for AD(1), we find that patients had, on average

- a decrease of 0.784 in the logit of the expected number of flagged report cards for hospitals that were ranked in the Honor Roll by the U.S. News & World Report;

The coefficient estimates for the Markov and AD(1) correlation structure cases are interpreted in the same way.

Each correlation structure case had a positive correlation parameter. The AD(1) correlation structure parameter estimate, $\hat{\alpha} = (0.432, 0.563, 0.493)$ suggests the adjacent correlation structure is rather consistent across time; the adjacent correlation parameter estimates between the visits have overlapping 95% confidence intervals. The correlation parameter estimate for AR(1) / Markov was 0.479, which was similar to the AD(1) correlation parameter estimates.

There was no statistically significant difference in how well the AD(1) model fit in comparison to the AR(1) / Markov models (p-vlaue = 0.106).

Comparing the results between the models with and without Year, we find that the Ranked coefficient parameter estimates for models without year are 0.1 within the model with year assuming AD(1). The correlation parameter estimates were very consistent, only varying by up to 0.04 of the equivalent model between with / without. The correlation parameter estimates varied from 0.43 to 0.56. Comparing the models with and without Year under a AD(1) correlation assumption, we find there was no significant difference in how well the model with Year fit (p-value = 0.124).

By comparison, in the GEE analysis, only the Ranked coefficient estimate was found to be statistically significant from 0, and only in the cases with Independence or Exchangeable correlation structures. All other parameters were not statistically different from 0. In terms of the magnitude and direction of the coefficient estimates, the GEE coefficients were very similar to those found in the Likelihood-based approach. The directions were all consistent and the difference in magnitude was up to 0.6. The estimates of α were similar than the Likelihood-based approach, with 0.531 for

AR(1) and 0.407 for Exchangeable correlation structures. Hence, the GEE methodology was more conservative in regards to the statistical significance of the parameters.

4.5. Simulations

We use simulation to assess the characteristics of the estimators and to demonstrate the methodology under a Binomial assumption. We simulate the number of flagged reports discussed previously in Section 4.4. In this dataset, the number of flagged reports was recorded every year for four years. The status of a hospital being ranked in the Honor Roll of the U.S. News & World Report was the main covariate of interest.

There were 220 hospitals in the dataset. We performed a simulation on the data by generating the number of flagged report outcomes. We created the number of flagged report outcomes by randomly drawing a value from the Binomial distributions. For $j = 1$, the distribution is defined by μ_{i1}. For $j > 1$, the distributions are defined by the conditional mean (3.3):

$$E(Y_{ij}|Y_{ij-1}) = \mu_{ij} + C_{ijj-1}\frac{\sigma_{ij}}{\sigma_{ij-1}}(Y_{ij-1} - \mu_{ij-1}).$$

C_{ijj-1} is defined as the correlation between Y_{ij} and Y_{ij-1}. μ_{ij} is assumed to be:

$$\mu_{ij} = l_{ij}\,\mathrm{expit}(\beta_0 + \beta_1 \cdot \mathsf{Ranked}_{ij} + \beta_2 \cdot \mathsf{Year}_{ij})$$

where $\beta = (\beta_0, \beta_1, \beta_2)$ are the coefficient values. We set $\beta = (0.75, -0.5, -0.25)$. The Year values are changed from $\{2012, 2013, 2014, 2015\}$ to $\{0, 1, 2, 3\}$, respectively. We considered three structures for the correlation structure: i) AR(1), ii) Markov, and iii) AD(1). We consider several different values for the correlation structure parameter α that depends on the correlation structure. For AR(1), we consider $\alpha \in \{0, 0.5\}$. For Markov, we consider $\alpha \in \{0.1, 0.5\}$. For AD(1), we consider $\alpha \in \{(0, 0.25, 0), (0, 0.5, 0.25)\}$. We consider varying sample sizes by multiplying the x patients and giving the copied patients unique identifiers.

After creating the artificial number of flagged reports, we fit our regression methodology to the resulting dataset. We calculate the mean square error, percent bias, and the 95% confidence interval of the ranked coefficient estimate. In fitting our regression methodology, we choose different combinations of the fitted correlation structure. We consider three fitted correlation structures: i)

AR(1), ii) Markov, and iii) AD(1).

Below, we lists the different factors from which we draw our simulation cases. The Sample Size cases refer to the factor by which the sample size is increased: 1X has 220 hospitals; 2X has 440 hospitals; 3X has 660 hospitals.

1. True Correlation Structure: AR(1), Markov, AD(1)
2. True Correlation Parameter - Depends on True Correlation Structure:
   - If AR(1): $\alpha \in \{0, 0.5\}$
   - If Markov: $\alpha \in \{0.1, 0.5\}$
   - If AD(1): $\alpha \in \{(0, 0.5, 0.25), (0, 0.25, 0)\}$
3. Fitted Correlation Structure: AR(1), Markov, AD(1)
4. Sample Size: 1X, 2X, 3X

We also used simulations to assess the performance of GEE on the same data. After simulating the seizure count outcomes in the same way, we apply GEE to the dataset and obtain the regression estimates. We assume the same cases as before except for the fitted correlation structure. For the fitted correlation structure, we assume independence. Because GEE regression estimates are robust to the fitted correlation structure, we chose the simplest one available. We note that, in testing a fitted correlation structure of AR(1), the GEE algorithm did not aways converge (not shown).

For each case, we simulated 100 data sets. Tables 4.5 and 4.6 show the mean square error, average percent bias, and the coverage probability of the 95% confidence interval for the treatment coefficient estimate under each of the cases examined.

Table 4.5 has the simulation statistics under which the true and fitted distributions are Binomial. First, we report the results for when the true correlation structure is AR(1). The MSE ranged from 3.124×10^{-3} to 19.860×10^{-3}; the average percent bias ranged from 0.054% to 8.827%; and the coverage probability ranged from 0.910 to 1.000. Having a true correlation parameter of 0.5 as compared to 0 lead to the MSE increasing by up to a factor of 3, the average percent bias had

no discernible pattern, and the coverage probability had no discernible pattern. When the true correlation parameter is 0, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is 0.5, there was no discernible pattern in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. As the sample size increased, the MSE decreased, the average percent bias had no discernible pattern, and the coverage probability had no discernible pattern. Next, we report the results for when the true correlation structure is Markov. The MSE ranged from 3.432×10^{-3} to 20.099×10^{-3}; the average percent bias ranged from 0.253% to 2.861%; and the coverage probability ranged from 0.920 to 0.980. Having a true correlation parameter of 0.5 as compared to 0.1 lead to the MSE increasing by up to a factor of 3, no discernible pattern in the average percent bias, and no discernible change for coverage probability. When the true correlation parameter is .1, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is 0.5, there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. As the sample size increased the MSE tended to decrease, though there were no discernible patterns in average percent bias or coverage probability. Finally, we report the results for when the true correlation structure is AD(1). The MSE ranged from 3.552×10^{-3} to 14.061×10^{-3}; the average percent bias ranged from 0.122% to 2.813%; and the coverage probability ranged from 0.900 to 0.990. Having a true correlation parameter of (0, 0.5, 0.25) as compared to (0, 0.25, 0) lead to the MSE increasing by up to a factor of 2, no discernible pattern in the average percent bias, and no discernible difference in the coverage probability. When the true correlation parameter is (0, 0.25, 0), there was no noticeable patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. When the true correlation parameter is (0, 0.5, 0.25), there was no discernible patterns in the MSE, average percent bias, or coverage probability in comparing the fitted correlation structures. As the sample size increased, MSE decreased, there was no discernible pattern in the average percent bias, and there was no discernible pattern in the coverage probability.

Table 4.6 has the GEE simulation statistics under which the true and fitted distribution are Binomial. The MSE ranged from 3.409×10^{-3} to 27.164×10^{-3}; the percent bias ranged from 0.054 to 8.706; and the coverage probability ranged from 0.880 to 1.000. There was no discernible pattern in how well the GEE model did across the different true correlation structures. As the true correlation

parameter increased in value, the MSE tended to increase, the percent bias had no discernible pattern, and there was no discernible pattern in the coverage probability. As the sample size increased the MSE decreased, the percent biased tended to decrease, and there was no discernible pattern in the coverage probability.

The Likelihood model performed well under a Binomial assumption. The MSE was consistently low, the average percent bias never went higher than 8%, and the coverage probability was consistently near 0.95. There was no consistency in how the model did when the fitted correlation structure was misspecified. By comparison, the GEE had similar results. The MSE, percent bias, and coverage probability had similar values and patterns as the Likelihood method.

4.6. Missing Data Simulation

We use simulation to assess the characteristics of the estimators under a missing at random (MAR) missing data mechanism. We simulate the number of flagged reports, using the same method as described in Section 4.5. In short, we created the number of flagged report outcomes by randomly drawing a value from the Binomial distributions. Again, we set $\beta = (0.75, -0.5, -0.25)$. We set the true correlation structure as AR(1) and set $\alpha = 0.5$. Hospitals without data across all four years were excluded. The resulting dataset is the "base" dataset from which the MAR simulation is conducted.

With the base dataset created, we induced missingness onto the dataset. For each hospital, if a random number drawn from a $U(0,1)$ distribution is less than $a \times \frac{Y_{i2}}{l_{i2}}$, then the third and fourth visits (Y_{i3} and Y_{i4}) are missing. The probability of this happening is $a \times \frac{Y_{i2}}{l_{i2}}$. If $a = 0$, then this is a missing completely at random mechanism. If $a > 0$, then this is a missing at random mechanism. We considered $a \in \{0, 0.1, \cdots, 0.65, 0.66\}$. For each value of a, we simulated 100 missing datasets.

After inducing values to be missing from the dataset, we fit both GEE and our regression methodology to each resulting dataset. We fit both models assuming an AR(1) correlation structure. We extract the regression parameter estimates β and the correlation parameter estimate α from both models. We found the average bias for each parameter.

Figure 4.1 plots the average bias for each parameter under GEE and our developed likelihood methodology against a. We first discuss the results of the plot of average bias against a for the

Intercept coefficient estimates. For both GEE and the Likelihood methodology, as a increased so did the absolute average bias. Both methods had similar projectiles, starting at an average bias of roughly -0.3 for $a \approx 0$ and linearly decreasing to roughly -0.5 for $a \approx 0.66$. The average bias for the Likelihood methodology averaged roughly 0.03 less than the average bias for the GEE methodology.

Next we discuss the results of the plot of average bias against a for the Ranked coefficient estimate. Again, as a increased so did the absolute average bias, for both GEE and the Likelihood methodology. For the Likelihood methodology, the average bias was roughly -0.03 for $a \approx 0$ and linearly decreased to roughly -0.04 for $a \approx 0.66$. For the GEE methodology, the average bias was roughly -0.04 for $a \approx 0$ and linearly decreased to roughly -0.06 for $a \approx 0.66$. The absolute average bias for the Likelihood methodology averaged roughly 0.01 less than the absolute average bias for the GEE methodology.

Next we discuss the results of the plot of average bias against a for the Year coefficient estimate. As a increased so did the absolute average bias, for both GEE and the Likelihood methodology. For the Likelihood methodology, the average bias was roughly 0.09 for $a \approx 0$ and linearly increased to roughly 0.14 for $a \approx 0.66$. For the GEE methodology, the average bias was roughly 0.10 for $a \approx 0$ and linearly increased to roughly 0.15 for $a \approx 0.66$. The absolute average bias for the Likelihood methodology averaged roughly 0.015 less than the average bias for the GEE methodology.

Finally, we discuss the results of the plot of average bias against a for the correlation parameter α estimates. The Likelihood methodology had an average bias of roughly +0.01 and it stayed roughly constant. The GEE methodology had an average bias roughly around -0.08 for $a \approx 0$ and linearly increased to roughly -0.06 for $a \approx 0.66$. The Likelihood methodology had an absolute average bias roughly 0.085 less than that of the GEE methodology.

Overall, our developed Likelihood methodology performed better than GEE in estimating the parameters. In each of the model parameters, the Likelihood methodology has less absolute bias than did the GEE methodology. In each regression parameter of β, the absolute bias increased as a increased. For the correlation parameter estimate of α, the average bias stayed relatively constant as a increase, whereas the absolute average bias decreased as a increased. These simulations suggest the developed Likelihood methodology perform better than GEE in the presense of MAR

data.

4.7. Discussion

We developed a maximum-likelihood based analysis that i) extends GLM for correlated binomial data, and ii) induces over-dispersion and plausible correlation structures for longitudinal data. We applied our methodology in evaluating the association between hospital quality evaluations developed by U.S. News & World Report and the transplant quality evaluations developed by the Scientific Registry of Transplant Recipients.

In our analysis of the U.S. News / SRTR dataset, we found that hospitals ranked in the Honor Roll of the U.S. News & World Report had a lower probability of flagged reports in SRTR. There was no statistical significance in the association between year and the probability of flagged reports. Among models that included year, AD(1) was the best-fitting model. Among models that did not include year, there was no statistically significant difference in how well the models fit.

Through simulations, we demonstrated that model performed well under a Binomial distribution. Across each of these scenarios, the model tended to do better when the true correlation parameter was higher. There was no consistency in how well the model did when the fitted correlation structure was different from the true correlation structure. The GEE performed similarly as our likelihood method. We also used simulations to compare our Likelihood methodology and GEE in the presence of data that are missing at random. Overall, the Likelihood methodology has less absolute bias than did the GEE methodology. These simulations suggest the developed Likelihood methodology perform better than GEE in the presense of MAR data.

Our maximum-likelihood based analysis has a lot of attractive features. Having a log-likelihood allows us to assess the fit of competing models and construct likelihood ratio tests. By contrast, GEE has no objective function, which complicates the process of comparing competing models and assessing goodness-of-fit. Our maximum-likelihood based analysis requires the data to be missing at random for unbiased analysis whereas GEE requires missing completely at random for valid analysis. However, our maximum-likelihood approach may be less robust to misspecifying the model. Whereas GEE analysis will yield a consistent estimator of the regression parameters even when incorrectly specified.

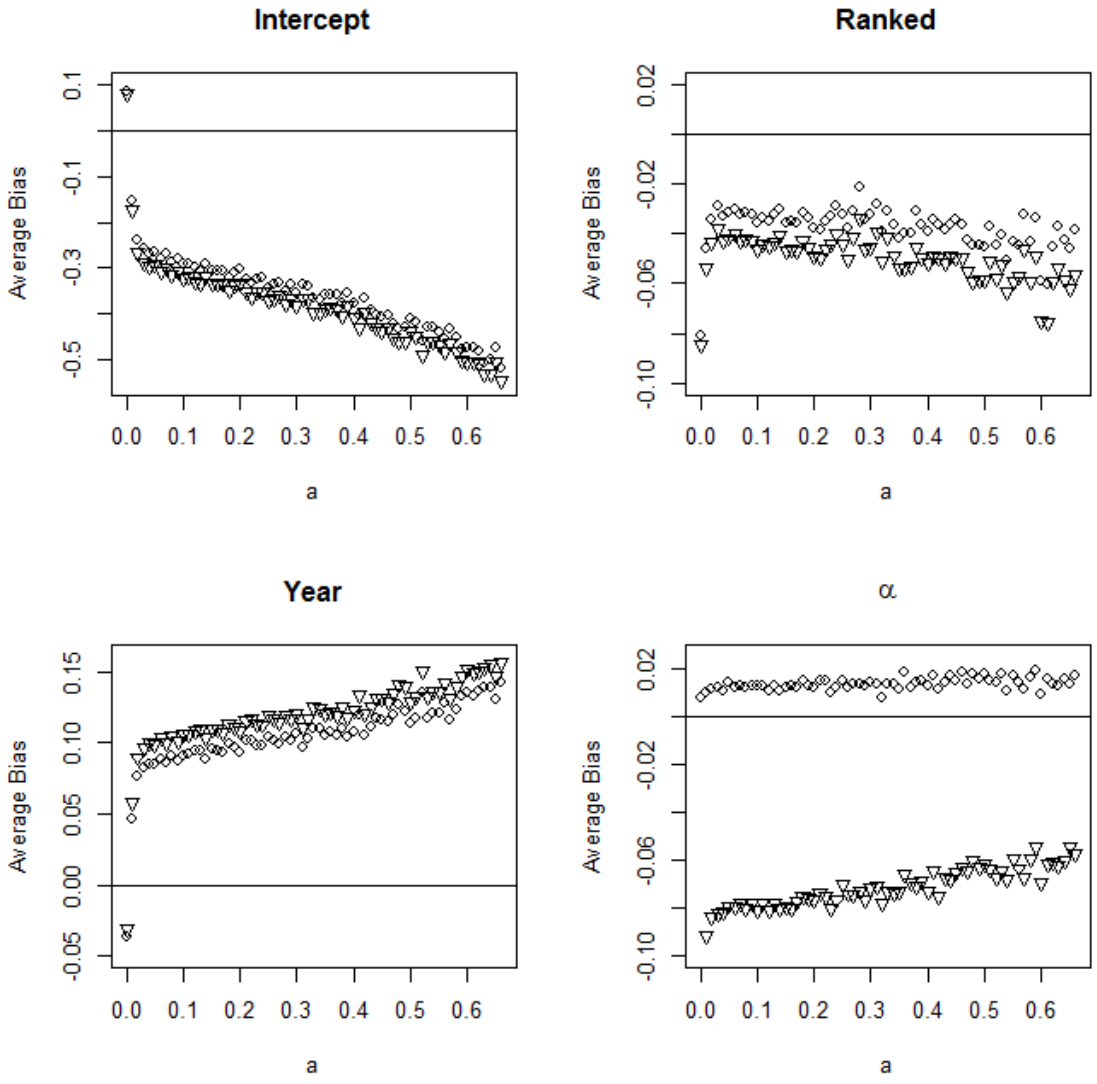

Figure 4.1: Average bias of each parameter estimate against MAR simulation parameter a for GEE (triangles) and the developed likelihood method (circles). A horizontal line at 0 has been added to each plot.

Year	n	Number of Ranked Programs	Number of Transplant Reports								Mean % of Flagged Reports	
			1	2	3	4	5	6	7	8	Ranked	Unranked
2012	213	17 (8.0%)	3	81	2	48	1	26	1	51	1.471	7.255
2013	213	18 (8.5%)	3	81	1	46	2	28	1	51	2.778	7.385
2014	213	17 (8.0%)	3	81	0	45	0	32	1	51	1.716	5.685
2015	208	15 (7.2%)	3	77	2	42	3	29	4	48	0.952	6.085
Total	220	67 (7.9%)	12	320	5	181	6	115	7	201	1.768	6.603

Table 4.1: Descriptive statistics across each year of the number of hospitals, number (percentage) of hospitals that were in the Honor Roll of the U.S. News & World Report, number of hospitals with a given number of transplant reports, and mean percentage of flagged reports for hospitals both in and not in the Honor Roll of the U.S. News & World Report.

AR(1)	Log-Likelihood	-496.053		
	AIC	1000.106		
	BIC	1013.680		
Variable	Estimate	Std. Error	Wald	Pr($>$|W|)
Intercept	25.860	65.592	0.155	0.693
Ranked	-1.172	0.395	8.822	0.003
Year	-0.014	0.031	0.206	0.650
α	0.473	0.047		
Markov	Log-Likelihood	-496.053		
	AIC	1000.106		
	BIC	1013.680		
Variable	Estimate	Std. Error	Wald	Pr($>$|W|)
Intercept	25.860	65.592	0.155	0.693
Ranked	-1.172	0.395	8.822	0.003
Year	-0.014	0.031	0.206	0.650
α	0.473	0.047		
AD(1)	Log-Likelihood	-493.475		
	AIC	994.950		
	BIC	1008.525		
Variable	Estimate	Std. Error	Wald	Pr($>$|W|)
Intercept	157.636	195.572	0.650	0.420
Ranked	-0.749	0.322	5.416	0.020
Year	-0.080	0.093	0.731	0.393
α_1	0.437	0.050		
α_2	0.552	0.086		
α_3	0.455	0.229		

Table 4.2: Goodness of fit statistics, coefficient estimates, standard errors, Wald statistics, and p-values for analysis of flagged hospitals under each correlation structure.

AR(1)	Log-Likelihood	-495.907		
	AIC	997.814		
	BIC	1007.995		
Variable	Estimate	Std. Error	Wald	Pr(>|W|)
Intercept	-2.699	0.102	706.125	< 0.001
Ranked	-0.838	0.350	5.731	0.017
α	0.479	0.040		
Markov	Log-Likelihood	-495.907		
	AIC	997.814		
	BIC	1007.995		
Variable	Estimate	Std. Error	Wald	Pr(>|W|)
Intercept	-2.699	0.102	706.125	< 0.001
Ranked	-0.838	0.350	5.731	0.017
α	0.479	0.040		
AD(1)	Log-Likelihood	-494.358		
	AIC	994.715		
	BIC	1004.896		
Variable	Estimate	Std. Error	Wald	Pr(>|W|)
Intercept	-2.680	0.104	666.995	<0.001
Ranked	-0.784	0.341	5.277	0.022
α_1	0.432	0.058		
α_2	0.563	0.057		
α_3	0.493	0.071		

Table 4.3: Goodness of fit statistics, coefficient estimates, standard errors, Wald statistics, and p-values for analysis of flagged hospitals under each correlation structure without year as a covariate.

Independence				
Variable	Estimate	Std. Error	Wald	$Pr(> \|W\|)$
Intercept	203.247	168.120	1.209	0.227
Ranked	-1.265	0.403	-3.143	0.002
Year	-0.102	0.084	-1.225	0.221
α	0	0.060		
AR(1)				
Variable	Estimate	Std. Error	Wald	$Pr(> \|W\|)$
Intercept	161.298	171.595	0.940	0.347
Ranked	-0.611	0.405	-1.507	0.132
Year	-0.081	0.085	-0.956	0.339
α	0.550			
Exchangeable				
Variable	Estimate	Std. Error	Wald	$Pr(> \|W\|)$
Intercept	191.698	169.472	1.131	0.258
Ranked	-0.799	0.380	-2.099	0.036
Year	-0.097	0.084	-1.147	0.251
α	0.385			

Table 4.4: Coefficient estimates, robust standard errors, Wald statistics, and p-values for GEE analysis of seizure counts in epileptics (Thall and Vail, 1990) under several correlation structure in a Negative-Binomial distribution assumption.

True Correlation Structure	True Correlation Parameter	Fitted Correlation Structure	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias (%)	Coverage Probability
AR(1)	0.0	AR(1)	1X	10.103	1.308	0.910
AR(1)	0.0	AR(1)	2X	3.802	0.054	1.000
AR(1)	0.0	AR(1)	3X	3.427	1.028	0.950
AR(1)	0.0	Markov	1X	11.141	1.422	0.960
AR(1)	0.0	Markov	2X	4.256	1.576	0.960
AR(1)	0.0	Markov	3X	3.124	1.029	0.950
AR(1)	0.0	AD(1)	1X	10.558	1.663	0.960
AR(1)	0.0	AD(1)	2X	4.816	0.507	0.940
AR(1)	0.0	AD(1)	3X	3.809	0.479	0.920
AR(1)	0.5	AR(1)	1X	19.860	8.827	0.930
AR(1)	0.5	AR(1)	2X	10.888	2.556	0.950
AR(1)	0.5	AR(1)	3X	5.657	3.316	0.950
AR(1)	0.5	Markov	1X	19.166	0.539	0.940
AR(1)	0.5	Markov	2X	8.561	0.101	0.950
AR(1)	0.5	Markov	3X	6.538	0.465	0.970
AR(1)	0.5	AD(1)	1X	17.752	0.293	0.940
AR(1)	0.5	AD(1)	2X	8.264	1.349	0.960
AR(1)	0.5	AD(1)	3X	5.120	0.349	0.970
Markov	0.1	AR(1)	1X	13.530	1.888	0.950
Markov	0.1	AR(1)	2X	5.313	0.326	0.970
Markov	0.1	AR(1)	3X	3.396	1.982	0.950
Markov	0.1	Markov	1X	8.775	1.127	0.980
Markov	0.1	Markov	2X	4.974	2.562	0.950
Markov	0.1	Markov	3X	5.305	0.654	0.930
Markov	0.1	AD(1)	1X	12.764	0.612	0.930
Markov	0.1	AD(1)	2X	5.500	0.327	0.970
Markov	0.1	AD(1)	3X	3.432	1.827	0.970
Markov	0.5	AR(1)	1X	21.925	0.253	0.920
Markov	0.5	AR(1)	2X	9.870	1.048	0.920
Markov	0.5	AR(1)	3X	7.507	2.192	0.930
Markov	0.5	Markov	1X	20.099	2.846	0.930
Markov	0.5	Markov	2X	10.633	2.861	0.940
Markov	0.5	Markov	3X	6.634	1.320	0.930
Markov	0.5	AD(1)	1X	17.205	0.467	0.950
Markov	0.5	AD(1)	2X	8.085	0.672	0.980
Markov	0.5	AD(1)	3X	5.217	0.354	0.980

Table 4.5: Continued on next page

Continued from previous page

True Correlation Structure	True Correlation Parameter	Fitted Correlation Structure	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias (%)	Coverage Probability
AD(1)	(0, 0.25, 0)	AR(1)	1X	13.523	1.277	0.910
AD(1)	(0, 0.25, 0)	AR(1)	2X	4.814	1.728	0.990
AD(1)	(0, 0.25, 0)	AR(1)	3X	3.552	0.391	0.940
AD(1)	(0, 0.25, 0)	Markov	1X	10.812	0.728	0.960
AD(1)	(0, 0.25, 0)	Markov	2X	6.146	0.376	0.930
AD(1)	(0, 0.25, 0)	Markov	3X	4.447	1.543	0.900
AD(1)	(0, 0.25, 0)	AD(1)	1X	9.704	0.143	0.960
AD(1)	(0, 0.25, 0)	AD(1)	2X	5.955	1.248	0.950
AD(1)	(0, 0.25, 0)	AD(1)	3X	3.642	0.122	0.960
AD(1)	(0, 0.5, 0.25)	AR(1)	1X	12.129	2.813	0.970
AD(1)	(0, 0.5, 0.25)	AR(1)	2X	6.270	0.625	0.950
AD(1)	(0, 0.5, 0.25)	AR(1)	3X	4.849	1.694	0.950
AD(1)	(0, 0.5, 0.25)	Markov	1X	12.047	0.253	0.970
AD(1)	(0, 0.5, 0.25)	Markov	2X	8.069	1.187	0.910
AD(1)	(0, 0.5, 0.25)	Markov	3X	4.721	1.730	0.940
AD(1)	(0, 0.5, 0.25)	AD(1)	1X	14.061	0.351	0.940
AD(1)	(0, 0.5, 0.25)	AD(1)	2X	7.483	0.403	0.920
AD(1)	(0, 0.5, 0.25)	AD(1)	3X	3.681	2.699	0.940

Table 4.5: Simulation results of mean square error (MSE), absolute value of the average percent bias, and coverage probability of the treatment parameter coefficient for combinations of true correlation structure, true correlation parameter, fitted correlation structure, and sample size.

True Correlation Structure	True Correlation Parameter	Sample Size	MSE ($\times 10^{-3}$)	Average Percent Bias	Coverage Probability
AR(1)	0.0	1X	10.116	1.410	0.900
AR(1)	0.0	2X	3.802	0.054	0.980
AR(1)	0.0	3X	3.423	0.986	0.940
AR(1)	0.5	1X	27.164	0.361	0.880
AR(1)	0.5	2X	9.768	0.746	0.960
AR(1)	0.5	3X	6.231	2.594	1.000
Markov	0.1	1X	10.890	3.231	0.950
Markov	0.1	2X	5.398	0.326	0.940
Markov	0.1	3X	3.409	1.047	0.940
Markov	0.5	1X	24.154	8.706	0.930
Markov	0.5	2X	13.692	2.414	0.930
Markov	0.5	3X	6.736	3.270	0.970
AD(1)	(0, 0.25, 0)	1X	13.322	2.594	0.950
AD(1)	(0, 0.25, 0)	2X	7.006	0.067	0.930
AD(1)	(0, 0.25, 0)	3X	3.452	0.576	0.970
AD(1)	(0, 0.5, 0.25)	1X	12.801	1.242	0.980
AD(1)	(0, 0.5, 0.25)	2X	7.422	1.118	0.920
AD(1)	(0, 0.5, 0.25)	3X	3.919	0.252	0.960

Table 4.6: GEE simulation results of mean square error (MSE), absolute value of the average percent bias, and coverage probability of the treatment parameter coefficient for combinations of true correlation structure, true correlation parameter, and sample size.

CHAPTER 5

DISCUSSION

In this dissertation we discussed two related topics. The first concerned the ignorability conditions for frequentist nonparametric analysis of conditional distributions with incomplete data. We discussed sufficient conditions for correct analysis of frequentist nonparametric inference on conditional distributions subject to incomplete data. We provided conditions that are weaker than the conditions that are usually considered necessary Rubin (1976). Assuming that the missing data mechanism is dependent on conditioning variables only, is known to be sufficient for unbiased estimation of a conditional distribution. However, we showed that this condition is not a necessary. We showed that we will have unbiased estimation when we ignore the missing data mechanism, if the missing data mechanism depends on conditioning variables only.

We also showed the condition of missing completely at random for correct inference for a marginal distribution can be relaxed to have the missing data mechanism also depend on latent variables if the latent and outcome variables are independent. A strength of our model and proofs is that they are completely free of any parametric assumptions.

A possible next step for this research is to evaluate what can be done when the ignorability conditions are not satisfied. In general, we will not know the missing data mechanism, let alone be able to determine that any missing data are ignorable. It would be of interest to attempt to correct our inference and thereby reduce bias, by collecting an additional sample of observations that is complete. This additional information could be used to adjust our analysis results with the goal of eliminating any bias due to missing data in the original analytic data set.

In the second part of this dissertation, we developed a maximum-likelihood based analysis that could be viewed as extending GLM for correlated discrete data with over-dispersion. Our approach assumed first-order antedependence of outcomes within subjects, exponential family distributions for the first and conditional distributions, and linearity of the conditional expectations. The assumptions of first-order antedependence and linearity induced decaying product correlation structures that are plausible for longitudinal data because they force the correlation on two measurements to decrease as the two measurements become further apart in time.

We proposed an approach for analysis of count data, by implementing the Poisson and Negative Binomial distributions. Our likelihood based approach allowed for easier assessment of goodness of fit (when compared with GEE) and also allowed for implementation of the likelihood ratio test that was useful for choosing between the Poisson and negative Binomial distributions in our analysis of seizure data. We also proposed an approach for analysis of binomial type outcomes, by implementing the Binomial distribution. We demonstrated our approach in an analysis of the association of SRTR transplant reports and the status of being in the U.S. News & World Report Honor Roll. Our analysis suggested that hospitals that were in the U.S. News & World Report Honor Roll of Hospitals had fewer occurrences of being flagged for poor performance with respect to organ transplantation.

Through simulations, we demonstrated the model performed well for when the distribution is correctly specified. Across each of these scenarios explored in the simulations, the model tended to do better when the true correlation parameter was higher. There was no consistency in how well the model did when the fitted correlation structure was different from the true correlation structure. The GEE performed similarly as our likelihood method. We also used simulations to assess our Likelihood methodology and GEE when data are missing at random. Overall, we found that the Likelihood methodology has less absolute bias than did the GEE methodology suggesting the developed Likelihood methodology perform better than GEE in the presense of MAR data.

We provided estimating equations under the assumption of exponential families, and then further simplified those equations for the Poisson, negative binomial, and binomial distributions. It should be relatively straightforward to implement our approach for other distributions that are members of an exponential family, such as the exponential or Dirichlet distributions. It might also be of interest to choose distributions for the first and then conditional distributions that are not members of the same family. For example, it might be of interest to allow the first measurement on a subject follow a Bernoulli distribution (to represent membership in a particular class), followed by conditional distributions that are negative binomial (to model counts).

Also, we are interested in further exploring the properties of the marginal distributions. Although we derived the means and variances of the marginal distribution, it would be of interest to learn more of the distribution. This can potentially be explored through simulations to obtain an estimate of the distribution. We are also interested in seeing how the method does in modeling data created

from alternative models. In the simulations, we modeled data that was simulated from the correct structure. It would be of interest to see how it does in other structures. Furthermore, we are interested in comparing the methodology with other alternative models beyond GEE. In addition, we are interested in appropriate goodness-of-fit statistics for our methodology.

APPENDIX A

CHAPTER 2 DERIVATIONS

A.1. Lemma 1

Given u and $\tilde{y}^K$, suppose the following conditions hold for all $y_i^M \in \Omega^M$:

1. $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K)$, and
2. $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$.

Then $\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) = F_{H|K}(u|\tilde{y}^K)$.

Proof:

$$
\begin{aligned}
&\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) \\
&\quad = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1) \\
&\quad = \frac{\Pr(Y_i^H < u, T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K)}{\Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K)} \\
&\quad = \frac{\Pr(Y_i^H < u | Y_i^K = \tilde{y}^K) \Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K)}{\Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K)} \\
&\quad = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K) \\
&\qquad \cdot \frac{\sum_{y_i^M} \Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \Pr(Y_i^M = y_i^M | Y_i^H < u, Y_i^K = \tilde{y}^K)}{\Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K)} \\
&\quad = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K) \\
&\qquad \cdot \sum_{y_i^M} \frac{\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)}{\Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K)} \Pr(Y_i^M = y_i^M | Y_i^H < u, Y_i^K = \tilde{y}^K) \\
&\quad = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K) \sum_{y_i^M} \Pr(Y_i^M = y_i^M | Y_i^H < u, Y_i^K = \tilde{y}^K) \\
&\quad = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K) \cdot 1 \\
&\quad = \mathrm{E}(Y_i^H < u | Y_i^K = \tilde{y}^K) \\
&\quad = F_{H|K}(u|\tilde{y}^K),
\end{aligned}
$$

which is the desired quantity.

A.2. Lemma 2

Given u and $\tilde{y}^K$, suppose the following conditions hold for all $y_i^M \in \Omega^M$:

1. $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$,
2. $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$, and
3. $\Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K)$.

Then $\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) = F_{H|K}(u|\tilde{y}^K)$.

Proof:

$$
\begin{aligned}
&\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) \\
&\quad = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1) \\
&\quad = \sum_{y_i^M} \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1, Y_i^M = y_i^M) \Pr(Y_i^M = y_i^M | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1) \\
&\quad = \sum_{y_i^M} \frac{\Pr(Y_i^H < u, T_{H,K,i} = 1, Y_i^M = y_i^M | Y_i^K = \tilde{y}^K)}{\Pr(T_{H,K,i} = 1, Y_i^M = y_i^M | Y_i^K = \tilde{y}^K)} \Pr(Y_i^M = y_i^M | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1) \\
&\quad = \sum_{y_i^M} \frac{\begin{array}{r}\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \Pr(Y_i^H < u, Y_i^M = y_i^M | Y_i^K = \tilde{y}^K) \\ \cdot \Pr(Y_i^M = y_i^M | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1)\end{array}}{\Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \Pr(Y_i^M = y_i^M | Y_i^K = \tilde{y}^K)} \\
&\quad = \sum_{y_i^M} \frac{\begin{array}{r}\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \Pr(Y_i^H < u | Y_i^M = y_i^M, Y_i^K = \tilde{y}^K) \\ \cdot \Pr(Y_i^M = y_i^M | Y_i^K = y_i^K) \Pr(Y_i^M = y_i^M | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1)\end{array}}{\Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \Pr(Y_i^M = y_i^M | Y_i^K = \tilde{y}^K)} \\
&\quad = \sum_{y_i^M} \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \Pr(Y_i^M = y_i^M | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1) \\
&\quad = \sum_{y_i^M} \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K) \Pr(Y_i^M = y_i^M | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1) \\
&\quad = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K) \sum_{y_i^M} \Pr(Y_i^M = y_i^M | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1) \\
&\quad = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K) \cdot 1
\end{aligned}
$$

$$= \mathrm{E}(Y_i^H < u | Y_i^K = \tilde{y}^K)$$
$$= F_{H|K}(u|\tilde{y}^K),$$

which is the desired quantity.

A.3. Lemma 3

Assume $\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$ for all u, y_i^M. Then

$$\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \quad \text{for all } u, y_i^M$$

if and only if

$$\Pr(T_{H,K,i} = 1 | Y_i^H = u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \quad \text{for all } u, y_i^M$$

Proof: We show that the former statement implies the latter. To show the converse, apply the steps in reverse. By assumption we have that, for any u,

$$\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

We manipulate this equality below to arrive at the desired equality.

$$\Pr(T_{H,K,i} = 1 | Y_i^H < u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

$$\frac{\Pr(T_{H,K,i} = 1, Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)}{\Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)} = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

$$\Pr(T_{H,K,i} = 1, Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$
$$= \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \cdot \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

Define u_1 as the largest element of Ω^H such that $u_1 < u$.

$$\Pr(T_{H,K,i} = 1, Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$
$$= \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \cdot \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

$$\Pr(T_{H,K,i} = 1, Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) - \Pr(T_{H,K,i} = 1, Y_i^H < u_1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$
$$= \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \cdot \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$
$$- \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \cdot \Pr(Y_i^H < u_1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

$$\sum_{y_i^H < u} \Pr(T_{H,K,i} = 1, Y_i^H = y_i^H | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$
$$- \sum_{y_i^H < u_1} \Pr(T_{H,K,i} = 1, Y_i^H = y_i^H | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$
$$= \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \cdot \sum_{y_i^H < u} \Pr(Y_i^H = y_i^H | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$
$$- \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) \cdot \sum_{y_i^H < u_1} \Pr(Y_i^H < y_i^H | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

$$\Pr(T_{H,K,i} = 1, Y_i^H = u_1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$
$$= \Pr(T_{H,K,i} = 1 | Y_i^K = y_i^K, Y_i^M = y_i^M) \cdot \Pr(Y_i^H = u_1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

$$\frac{\Pr(T_{H,K,i} = 1, Y_i^H = u_1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)}{\Pr(Y_i^H = u_1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)} = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

$$\Pr(T_{H,K,i} = 1 | Y_i^H = u_1, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$$

Repeat these same steps for all $u \in \Omega^H$ to obtain the full result.

A.4. Lemma 4

Suppose $n_{H,K,\tilde{y}^K} > 0$ and $\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) = F_{H|K}(u|\tilde{y}^K)$ for all units in the set $\{i : r_i^H = 1^H, r_i^K = 1^K, y_i^K = \tilde{y}^K\}$. Then $\mathrm{E}\left(\bar{I}_{H,K,\tilde{y}^K,u} \middle| Y^K = y^K, R = r\right) = F_{H|K}(u|\tilde{y}^K)$.

Proof:

$$\mathrm{E}\left(\bar{I}_{H,K,\tilde{y}^K,u}\middle|Y^K=y^K,R=r\right)=\mathrm{E}\left[\frac{1}{n_{H,K,\tilde{y}^K}}\sum_{i:y_i^K=\tilde{y}^K}T_{H,K,i}\cdot I(Y_i^H<u)\middle|Y^K=y^K,R=r\right]$$
$$=\frac{1}{n_{H,K,\tilde{y}^K}}\sum_{i:y_i^K=\tilde{y}^K}T_{H,K,i}\mathrm{E}\left[I(Y_i^H<u)|Y^K=y^K,R=r\right]$$

Because units in Y and R are independent, $I(Y_i^H<u)$ is independent of Y_j^K and R_j^K for $i\neq j$. This implies $\mathrm{E}[I(Y_i^H<u)|Y^K=y^K,R=r]=\mathrm{E}[I(Y_i^H<u)|Y_i^K=y_i^k,R_i=r_i]$.

$$=\frac{1}{n_{H,K,\tilde{y}^K}}\sum_{i:y_i^K=\tilde{y}^K}T_{H,K,i}\mathrm{E}\left[I(Y_i^H<u)|Y_i^K=y_i^k,R_i=r_i\right]$$

Since the summation is limited to units such that $Y_i^K=\tilde{y}^K$, the conditioning argument $Y_i^K=y_i^k$ is re-written as $Y_i^K=\tilde{y}^K$.

$$=\frac{1}{n_{H,K,\tilde{y}^K}}\sum_{i:y_i^K=\tilde{y}^K}T_{H,K,i}\mathrm{E}\left[I(Y_i^H<u)|Y_i^K=\tilde{y}^K,R_i=r_i\right]$$

The component $T_{H,K,i}$ is non-zero only when $R_i^H=1^H$ and $R_i^M=1^M$, or equivalently, $T_{H,K,i}=1$. Hence, the conditioning argument $R_i=r_i$ is re-written as $T_{H,K,i}=1$.

$$=\frac{1}{n_{H,K,\tilde{y}^K}}\sum_{i:y_i^K=\tilde{y}^K}T_{H,K,i}\cdot\mathrm{E}\left[I(Y_i^H<u)|Y_i^K=\tilde{y}^K,T_{H,K,i}=1\right]$$
$$=\frac{1}{n_{H,K,\tilde{y}^K}}\sum_{i:y_i^K=\tilde{y}^K}T_{H,K,i}\cdot F_{H|K}(u|\tilde{y}^K)$$
$$=F_{H|K}(u|\tilde{y}^K),$$

which is the desired quantity.

A.5. Theorem

Given $\tilde{y}^K$ and suppose $n_{H,K,\tilde{y}^K}>0$. Suppose further that for each unit i in the set $\{i:r_i^H=1^H,r_i^K=1^K,Y_i^K=\tilde{y}^K\}$ one of the following sets of conditions holds for all u and y_i^M:

1. (a) $\Pr(T_{H,K,i}=1|Y_i^H=u,Y_i^K=\tilde{y}^K,Y_i^M=y_i^M)=\Pr(T_{H,K,i}=1|Y_i^K=\tilde{y}^K)$, and

(b) $\Pr(T_{H,K,i} = 1 | Y_i^H = u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$; or

2. (a) $\Pr(T_{H,K,i} = 1 | Y_i^H = u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(T_{H,K,i} = 1 | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M)$,

   (b) $\Pr(T_{H,K,i} = 1 | Y_i^H = u, Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) > 0$, and

   (c) $\Pr(Y_i^H < u | Y_i^K = \tilde{y}^K, Y_i^M = y_i^M) = \Pr(Y_i^H < u | Y_i^K = \tilde{y}^K)$.

Then $\mathrm{E}\left(\bar{I}_{H,K,\tilde{y}^K,u} \middle| Y^K = y^K, R = r\right) = F_{H|K}(u|\tilde{y}^K)$.

Proof: By Lemma 3, the conditioning argument $Y_i^H = u$ is interchangeable with $Y_i^H < u$ in the probabilities involving $T_{H,K,i} = 1$. With this change, every unit in the set $\{i : r_i^H = 1^H, r_i^K = 1^K, Y_i^K = \tilde{y}^K\}$ satisfies the conditions of either Lemma 1 or Lemma 2. This implies that, for each unit i in that set, $\mathrm{E}\left(Y_i^H < u | Y_i^K = \tilde{y}^K, T_{H,K,i} = 1\right) = F_{H|K}(u|\tilde{y}^K)$. By Lemma 4,

$$\mathrm{E}\left(\bar{I}_{H,K,\tilde{y}^K,u} \middle| Y^K = y^K, R = r\right) = F_{H|K}(u|\tilde{y}^K).$$

APPENDIX B

CHAPTER 3 DERIVATIONS

B.1. Derivation of β Score Equations

Below are formulas used in the derivation of the score equation with respect to β. We have

$$\begin{aligned}\frac{\partial \mu_{ij}^*}{\partial \beta} &= \frac{\partial \mu_{ij}}{\partial \beta} + C_{ijj-1}\left(\frac{\sigma_{ij}}{\sigma_{ij-1}}\frac{-\partial \mu_{ij-1}}{\partial \beta} + (Y_{ij-1}-\mu_{ij-1})\frac{\sigma_{ij-1}\frac{\partial \sigma_{ij}}{\partial \beta} - \sigma_{ij}\frac{\partial \sigma_{ij-1}}{\partial \beta}}{\sigma_{ij-1}^2}\right) \\ &= \frac{\partial \mu_{ij}}{\partial \beta} + C_{ijj-1}\left((Y_{ij-1}-\mu_{ij-1})\left(\frac{1}{\sigma_{ij-1}}\frac{\partial \sigma_{ij}}{\partial \beta} - \frac{\sigma_{ij}}{\sigma_{ij-1}^2}\frac{\partial \sigma_{ij-1}}{\partial \beta}\right) - \frac{\sigma_{ij}}{\sigma_{ij-1}}\frac{\partial \mu_{ij-1}}{\partial \beta}\right) \\ &= \frac{\partial \mu_{ij}}{\partial \beta} + C_{ijj-1}\frac{\sigma_{ij}}{\sigma_{ij-1}}\left((Y_{ij-1}-\mu_{ij-1})\left(\frac{1}{\sigma_{ij}}\frac{\partial \sigma_{ij}}{\partial \beta} - \frac{1}{\sigma_{ij-1}}\frac{\partial \sigma_{ij-1}}{\partial \beta}\right) - \frac{\partial \mu_{ij-1}}{\partial \beta}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial \sigma_{i1}}{\partial \beta} &= \frac{\partial \sqrt{Var(Y_{i1})}}{\partial \beta} = \frac{1}{2}(Var(Y_{i1}))^{-1/2}\frac{\partial Var(Y_{i1})}{\partial \beta} \\ &= \frac{1}{2}\frac{1}{\sqrt{Var(Y_{i1})}}\frac{\partial Var(Y_{i1})}{\partial \beta}\end{aligned}$$

And, for $j > 1$,

$$\begin{aligned}\frac{\partial \sigma_{ij}}{\partial \beta} &= \frac{\partial \sqrt{\frac{1}{1-C_{ijj-1}^2}E(Var(Y_{ij}|Y_{ij-1}))}}{\partial \beta} \\ &= \frac{1}{\sqrt{1-C_{ijj-1}^2}}\frac{\partial \sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}{\partial \beta} \\ &= \frac{1}{\sqrt{1-C_{ijj-1}^2}}\frac{1}{2}\frac{1}{\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}\frac{\partial E(Var(Y_{ij}|Y_{ij-1}))}{\partial \beta} \\ &= \frac{1}{2}\frac{1}{\sqrt{(1-C_{ijj-1}^2)E(Var(Y_{ij}|Y_{ij-1}))}}\frac{\partial E(Var(Y_{ij}|Y_{ij-1}))}{\partial \beta}\end{aligned}$$

Hence, for $j = 2$, $\frac{\partial \mu_{ij}^*}{\partial \beta}$ is equal to

$$\frac{\partial \mu_{i2}^*}{\partial \beta} = \frac{\partial \mu_{i2}}{\partial \beta} + C_{i21}\frac{\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{1-C_{i21}^2}}}{\sqrt{Var(Y_{i1})}}\left((Y_{i1}-\mu_{i1})\left(\frac{1}{\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{1-C_{i21}^2}}}\frac{1}{2}\frac{1}{\sqrt{(1-C_{i21}^2)E(Var(Y_{i2}|Y_{i1}))}}\right.\right.$$

$$\times \frac{\partial E(Var(Y_{i2}|Y_{i1}))}{\partial \beta} - \frac{1}{\sqrt{Var(Y_{i1})}}\frac{1}{2}\frac{1}{\sqrt{Var(Y_{i1})}}\frac{\partial Var(Y_{i1})}{\partial \beta}\Bigg) - \frac{\partial \mu_{i1}}{\partial \beta}\Bigg)$$

$$= \frac{\partial \mu_{i2}}{\partial \beta} + \frac{C_{i21}}{\sqrt{1-C_{i21}^2}}\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}}\Bigg(\frac{Y_{i1}-\mu_{i1}}{2}\Bigg(\frac{1}{E(Var(Y_{i2}|Y_{i1}))}\frac{\partial E(Var(Y_{i2}|Y_{i1}))}{\partial \beta}$$

$$- \frac{1}{Var(Y_{i1})}\frac{\partial Var(Y_{i1})}{\partial \beta}\Bigg) - \frac{\partial \mu_{i1}}{\partial \beta}\Bigg)$$

And, for $j > 2$, $\frac{\partial \mu_{ij}^*}{\partial \beta}$ is equal to

$$\frac{\partial \mu_{ij}^*}{\partial \beta} = \frac{\partial \mu_{ij}}{\partial \beta} + C_{ijj-1}\frac{\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{1-C_{ijj-1}^2}}}{\sqrt{\frac{E(Var(Y_{ij-1}|Y_{ij-2}))}{1-C_{ij-1j-2}^2}}}\Bigg((Y_{ij-1}-\mu_{ij-1})\Bigg(\frac{1}{\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{1-C_{ijj-1}^2}}}\frac{1}{2}$$

$$\times \frac{1}{\sqrt{(1-C_{ijj-1}^2)E(Var(Y_{ij}|Y_{ij-1}))}}\frac{\partial E(Var(Y_{ij}|Y_{ij-1}))}{\partial \beta} - \frac{1}{\sqrt{\frac{E(Var(Y_{ij-1}|Y_{ij-2}))}{1-C_{ij-1j-2}^2}}}\frac{1}{2}$$

$$\times \frac{1}{\sqrt{(1-C_{ij-1j-2}^2)E(Var(Y_{ij-1}|Y_{ij-2})}}\frac{\partial E(Var(Y_{ij-1}|Y_{ij-2})}{\partial \beta}\Bigg) - \frac{\partial \mu_{ij-1}}{\partial \beta}\Bigg)$$

$$= \frac{\partial \mu_{ij}}{\partial \beta} + C_{ijj-1}\frac{\sqrt{1-C_{ij-1j-2}^2}}{\sqrt{1-C_{ijj-1}^2}}\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2})}}\Bigg(\frac{Y_{ij-1}-\mu_{ij-1}}{2}\Bigg(\frac{1}{E(Var(Y_{ij}|Y_{ij-1}))}$$

$$\times \frac{\partial E(Var(Y_{ij}|Y_{ij-1}))}{\partial \beta} - \frac{1}{E(Var(Y_{ij-1}|Y_{ij-2}))}\frac{\partial E(Var(Y_{ij-1}|Y_{ij-2}))}{\partial \beta}\Bigg) - \frac{\partial \mu_{ij-1}}{\partial \beta}\Bigg)$$

B.2. Derivation of α Score Equations

Below are formulas used in the derivation of the score equation with respect to α. We have

$$\frac{\partial \mu_{ij}^*}{\partial \alpha} = (Y_{ij-1}-\mu_{ij-1})\left(C_{ijj-1}\frac{\sigma_{ij-1}\frac{\partial \sigma_{ij}}{\partial \alpha} - \sigma_{ij}\frac{\partial \sigma_{ij-1}}{\partial \alpha}}{\sigma_{ij-1}^2} + \frac{\sigma_{ij}}{\sigma_{ij-1}}\frac{\partial C_{ijj-1}}{\partial \alpha}\right)$$

$$= (Y_{ij-1}-\mu_{ij-1})\left(\frac{\sigma_{ij}}{\sigma_{ij-1}}\frac{\partial C_{ijj-1}}{\partial \alpha} + C_{ijj-1}\left(\frac{1}{\sigma_{ij-1}}\frac{\partial \sigma_{ij}}{\partial \alpha} - \frac{\sigma_{ij}}{\sigma_{ij-1}^2}\frac{\partial \sigma_{ij-1}}{\partial \alpha}\right)\right)$$

$$= (Y_{ij-1}-\mu_{ij-1})\frac{\sigma_{ij}}{\sigma_{ij-1}}\left(\frac{\partial C_{ijj-1}}{\partial \alpha} + C_{ijj-1}\left(\frac{1}{\sigma_{ij}}\frac{\partial \sigma_{ij}}{\partial \alpha} - \frac{1}{\sigma_{ij-1}}\frac{\partial \sigma_{ij-1}}{\partial \alpha}\right)\right)$$

$$\frac{\partial \sigma_{i1}}{\partial \alpha} = 0$$

And, for $j > 1$,

$$\begin{aligned}\frac{\partial \sigma_{ij}}{\partial \alpha} &= \sqrt{E(Var(Y_{ij}|Y_{ij-1}))}\frac{\partial (1-C_{ijj-1}^2)^{-1/2}}{\partial \alpha} \\ &= \sqrt{E(Var(Y_{ij}|Y_{ij-1}))}\frac{-1}{2}(1-C_{ijj-1}^2)^{-3/2}\cdot -2C_{ijj-1}\frac{\partial C_{ijj-1}}{\partial \alpha} \\ &= \frac{\partial C_{ijj-1}}{\partial \alpha}\frac{C_{ijj-1}}{(1-C_{ijj-1}^2)^{3/2}}\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}\end{aligned}$$

B.2.1. AR(1) Correlation Structure

For an AR(1) correlation structure, $\frac{\partial \sigma_{ij}}{\partial \alpha}$ simplifies to, for $j > 1$,

$$\frac{\partial \sigma_{ij}}{\partial \alpha} = \frac{\alpha}{(1-\alpha^2)^{3/2}}\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}$$

Hence, for $j = 2$, $\frac{\partial \mu_{ij}^*}{\partial \alpha}$ simplifies to

$$\begin{aligned}\frac{\partial \mu_{i2}^*}{\partial \alpha} &= (Y_{i1}-\mu_{i1})\frac{\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{1-\alpha^2}}}{\sqrt{Var(Y_{i1})}}\left(1+\alpha\left(\frac{1}{\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{1-\alpha^2}}}\frac{\alpha}{(1-\alpha^2)^{3/2}}\sqrt{E(Var(Y_{i2}|Y_{i1}))}-0\right)\right) \\ &= \frac{Y_{i1}-\mu_{i1}}{\sqrt{1-\alpha^2}}\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}}\left(1+\alpha\frac{\alpha}{1-\alpha^2}\right) \\ &= \frac{Y_{i1}-\mu_{i1}}{(1-\alpha^2)^{3/2}}\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}}\end{aligned}$$

And, for $j > 2$,

$$\begin{aligned}\frac{\partial \mu_{ij}^*}{\partial \alpha} &= (Y_{ij-1}-\mu_{ij-1})\frac{\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{1-\alpha^2}}}{\sqrt{\frac{E(Var(Y_{ij-1}|Y_{ij-2}))}{1-\alpha^2}}}\left(1+\alpha\left(\frac{1}{\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{1-\alpha^2}}}\frac{\alpha}{(1-\alpha^2)^{3/2}}\right.\right. \\ &\quad \times\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}-\frac{1}{\sqrt{\frac{E(Var(Y_{ij-1}|Y_{ij-2}))}{1-\alpha^2}}}\frac{\alpha}{(1-\alpha^2)^{3/2}} \\ &\quad \left.\left.\times\sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))}\right)\right) \\ &= (Y_{ij-1}-\mu_{ij-1})\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}}\end{aligned}$$

B.2.2. Markov Correlation Structure

For a Markov correlation structure, $\frac{\partial \sigma_{ij}}{\partial \alpha}$ simplifies to, for $j > 1$,

$$\frac{\partial \sigma_{ij}}{\partial \alpha} = (t_{ij} - t_{ij-1})\alpha^{t_{ij}-t_{ij-1}-1}\frac{\alpha^{t_{ij}-t_{ij-1}}}{(1-\alpha^{2t_{ij}-2t_{ij-1}})^{3/2}}\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}$$

Hence, for $j = 2$, $\frac{\partial \mu^*_{ij}}{\partial \alpha}$ simplifies to

$$\begin{aligned}
\frac{\partial \mu^*_{i2}}{\partial \alpha} &= (Y_{i1} - \mu_{i1})\frac{\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{1-\alpha^{2t_{i2}-2t_{i1}}}}}{\sqrt{Var(Y_{i1})}}\Bigg((t_{i2} - t_{i1})\alpha^{t_{i2}-t_{i1}-1} + \alpha^{t_{i2}-t_{i1}}\Bigg(\frac{1}{\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{1-\alpha^{2t_{i2}-2t_{i1}}}}}(t_{i2} - t_{i1}) \\
&\quad \times \alpha^{t_{i2}-t_{i1}-1}\frac{\alpha^{t_{i2}-t_{i1}}}{(1-\alpha^{2t_{i2}-2t_{i1}})^{3/2}}\sqrt{E(Var(Y_{i2}|Y_{i1}))} - 0\Bigg)\Bigg) \\
&= \frac{Y_{i1} - \mu_{i1}}{\sqrt{1-\alpha^{2t_{i2}-2t_{i1}}}}\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}}\Bigg((t_{i2} - t_{i1})\alpha^{t_{i2}-t_{i1}-1} + \frac{(t_{i2} - t_{i1})\alpha^{t_{i2}-t_{i1}-1}\alpha^{t_{i2}-t_{i1}}}{1-\alpha^{2t_{i2}-2t_{i1}}}\Bigg) \\
&= \frac{Y_{i1} - \mu_{i1}}{\sqrt{1-\alpha^{2t_{i2}-2t_{i1}}}}\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}}(t_{i2} - t_{i1})\alpha^{t_{i2}-t_{i1}-1}\Bigg(1 + \frac{\alpha^{t_{i2}-t_{i1}}}{1-\alpha^{2t_{i2}-2t_{i1}}}\Bigg)
\end{aligned}$$

And, for $j > 2$,

$$\begin{aligned}
\frac{\partial \mu^*_{ij}}{\partial \alpha} &= (Y_{ij-1} - \mu_{ij-1})\frac{\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{1-\alpha^{2t_{ij}-2t_{ij-1}}}}}{\sqrt{\frac{E(Var(Y_{ij-1}|Y_{ij-2}))}{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}}}\Bigg((t_{ij} - t_{ij-1})\alpha^{t_{ij}-t_{ij-1}-1} + \alpha^{t_{ij}-t_{ij-1}} \\
&\quad \times \Bigg(\frac{1}{\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{1-\alpha^{2t_{ij}-2t_{ij-1}}}}} \\
&\quad \times (t_{ij} - t_{ij-1})\alpha^{t_{ij}-t_{ij-1}-1}\frac{\alpha^{t_{ij}-t_{ij-1}}}{(1-\alpha^{2t_{ij}-2t_{ij-1}})^{3/2}}\sqrt{E(Var(Y_{ij}|Y_{ij-1}))} \\
&\quad - \frac{1}{\sqrt{\frac{E(Var(Y_{ij-1}|Y_{ij-2}))}{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}}} \\
&\quad \times (t_{ij-1} - t_{ij-2})\alpha^{t_{ij-1}-t_{ij-2}-1}\frac{\alpha^{t_{ij-1}-t_{ij-2}}}{(1-\alpha^{2t_{ij-1}-2t_{ij-2}})^{3/2}}\sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))}\Bigg)\Bigg) \\
&= (Y_{ij-1} - \mu_{ij-1})\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}}\sqrt{\frac{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}{1-\alpha^{2t_{ij}-2t_{ij-1}}}}\Bigg((t_{ij} - t_{ij-1})\alpha^{t_{ij}-t_{ij-1}-1} \\
&\quad + \alpha^{t_{ij}-t_{ij-1}}\Bigg(\frac{(t_{ij} - t_{ij-1})\alpha^{t_{ij}-t_{ij-1}-1}\alpha^{t_{ij}-t_{ij-1}}}{1-\alpha^{2t_{ij}-2t_{ij-1}}} \\
&\quad - \frac{(t_{ij-1} - t_{ij-2})\alpha^{t_{ij-1}-t_{ij-2}-1}\alpha^{t_{ij-1}-t_{ij-2}}}{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}\Bigg)\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= (Y_{ij-1} - \mu_{ij-1})\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}}\sqrt{\frac{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}{1-\alpha^{2t_{ij}-2t_{ij-1}}}}\alpha^{t_{ij}-t_{ij-1}-1}\Bigg((t_{ij}-t_{ij-1}) \\
&\quad + \frac{(t_{ij}-t_{ij-1})\alpha^{2t_{ij}-2t_{ij-1}}}{1-\alpha^{2t_{ij}-2t_{ij-1}}} - \frac{(t_{ij-1}-t_{ij-2})\alpha^{2t_{ij-1}-2t_{ij-2}}}{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}\Bigg) \\
&= (Y_{ij-1} - \mu_{ij-1})\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}}\sqrt{\frac{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}{1-\alpha^{2t_{ij}-2t_{ij-1}}}}\alpha^{t_{ij}-t_{ij-1}-1}\Bigg((t_{ij}-t_{ij-1}) \\
&\quad \times \left(1 + \frac{\alpha^{2t_{ij}-2t_{ij-1}}}{1-\alpha^{2t_{ij}-2t_{ij-1}}}\right) - \frac{(t_{ij-1}-t_{ij-2})\alpha^{2t_{ij-1}-2t_{ij-2}}}{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}\Bigg) \\
&= (Y_{ij-1} - \mu_{ij-1})\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}}\sqrt{\frac{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}{1-\alpha^{2t_{ij}-2t_{ij-1}}}}\alpha^{t_{ij}-t_{ij-1}-1} \\
&\quad \times \left(\frac{t_{ij}-t_{ij-1}}{1-\alpha^{2t_{ij}-2t_{ij-1}}} - \frac{(t_{ij-1}-t_{ij-2})\alpha^{2t_{ij-1}-2t_{ij-2}}}{1-\alpha^{2t_{ij-1}-2t_{ij-2}}}\right)
\end{aligned}
$$

B.2.3. AD(1) Correlation Structure

Let $\hat{I}_j$ denote a vector containing a 1 in the jth element and 0 elsewhere. For an AD(1) correlation structure, $\frac{\partial \sigma_{ij}}{\partial \alpha}$ simplifies to (with $\alpha = (\alpha_1, \cdots, \alpha_{n-1})$)

$$
\frac{\partial \sigma_{ij}}{\partial \alpha} = \hat{I}_{j-1}\frac{\alpha_{j-1}}{(1-\alpha_{j-1}^2)^{3/2}}\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}
$$

Hence, $\frac{\partial \mu_{ij}^*}{\partial \alpha}$ simplifies to, for $j = 2$,

$$
\begin{aligned}
\frac{\partial \mu_{i2}^*}{\partial \alpha} &= (Y_{i1} - \mu_{i1})\frac{\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{1-\alpha_1^2}}}{\sqrt{Var(Y_{i1})}}\Bigg(\hat{I}_1 \\
&\qquad + \alpha_1\left(\frac{1}{\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{1-\alpha^2}}}\hat{I}_1\frac{\alpha_1}{(1-\alpha_1^2)^{3/2}}\sqrt{E(Var(Y_{i2}|Y_{i1}))} - 0\right)\Bigg) \\
&= \frac{Y_{i1}-\mu_{i1}}{\sqrt{1-\alpha_1^2}}\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}}\left(\hat{I}_1 + \frac{\hat{I}_1\alpha_1^2}{1-\alpha_1^2}\right) \\
&= \frac{Y_{i1}-\mu_{i1}}{\sqrt{1-\alpha_1^2}}\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}}\hat{I}_1\left(1 + \frac{\alpha_1^2}{1-\alpha_1^2}\right) \\
&= \frac{Y_{i1}-\mu_{i1}}{\sqrt{1-\alpha_1^2}}\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}}\hat{I}_1\frac{1}{1-\alpha_1^2} \\
&= (Y_{i1}-\mu_{i1})\sqrt{\frac{E(Var(Y_{i2}|Y_{i1}))}{Var(Y_{i1})}}\frac{\hat{I}_1}{(1-\alpha_1^2)^{3/2}}
\end{aligned}
$$

And, for $j > 2$,

$$\begin{aligned}
\frac{\partial \mu_{ij}^*}{\partial \alpha} &= (Y_{ij-1} - \mu_{ij-1}) \frac{\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{1-\alpha_{j-1}^2}}}{\sqrt{\frac{E(Var(Y_{ij-1}|Y_{ij-2}))}{1-\alpha_{j-2}^2}}} \Bigg(\hat{I}_{j-1} + \alpha_{j-1} \Bigg(\frac{1}{\sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{1-\alpha_{j-1}^2}}} \hat{I}_{j-1} \frac{\alpha_{j-1}}{(1-\alpha_{j-1}^2)^{3/2}} \\
&\quad \times \sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))} - \frac{1}{\sqrt{\frac{E(Var(Y_{ij-1}|Y_{ij-2}))}{1-\alpha_{j-2}^2}}} \hat{I}_{j-2} \frac{\alpha_{j-2}}{(1-\alpha_{j-2}^2)^{3/2}} \\
&\quad \times \sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))} \Bigg) \Bigg) \\
&= (Y_{ij-1} - \mu_{ij-1}) \sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}} \sqrt{\frac{1-\alpha_{j-2}^2}{1-\alpha_{j-1}^2}} \Bigg(\hat{I}_{j-1} + \alpha_{j-1} \Bigg(\frac{\hat{I}_{j-1}\alpha_{j-1}}{1-\alpha_{j-1}^2} \\
&\quad - \frac{\hat{I}_{j-2}\alpha_{j-2}}{1-\alpha_{j-2}^2} \Bigg) \Bigg) \\
&= (Y_{ij-1} - \mu_{ij-1}) \sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}} \sqrt{\frac{1-\alpha_{j-2}^2}{1-\alpha_{j-1}^2}} \left(\hat{I}_{j-1} + \frac{\hat{I}_{j-1}\alpha_{j-1}^2}{1-\alpha_{j-1}^2} - \frac{\hat{I}_{j-2}\alpha_{j-1}\alpha_{j-2}}{1-\alpha_{j-2}^2} \right) \\
&= (Y_{ij-1} - \mu_{ij-1}) \sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}} \sqrt{\frac{1-\alpha_{j-2}^2}{1-\alpha_{j-1}^2}} \left(\hat{I}_{j-1} \left(1 + \frac{\alpha_{j-1}^2}{1-\alpha_{j-1}^2} \right) \right. \\
&\quad \left. - \frac{\hat{I}_{j-2}\alpha_{j-1}\alpha_{j-2}}{1-\alpha_{j-2}^2} \right) \\
&= (Y_{ij-1} - \mu_{ij-1}) \sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}} \sqrt{\frac{1-\alpha_{j-2}^2}{1-\alpha_{j-1}^2}} \left(\hat{I}_{j-1} \frac{1}{1-\alpha_{j-1}^2} - \frac{\hat{I}_{j-2}\alpha_{j-1}\alpha_{j-2}}{1-\alpha_{j-2}^2} \right) \\
&= (Y_{ij-1} - \mu_{ij-1}) \sqrt{\frac{E(Var(Y_{ij}|Y_{ij-1}))}{E(Var(Y_{ij-1}|Y_{ij-2}))}} \left(\hat{I}_{j-1} \frac{\sqrt{1-\alpha_{j-2}^2}}{(1-\alpha_{j-1}^2)^{3/2}} \right. \\
&\quad \left. - \hat{I}_{j-2} \frac{\alpha_{j-1}\alpha_{j-2}}{\sqrt{(1-\alpha_{j-1}^2)(1-\alpha_{j-2}^2)}} \right)
\end{aligned}$$

B.3. Derivation of r Score Equations for Negative-Binomial Case

$$\frac{\partial \theta_{i1}}{\partial r} = \frac{\frac{(1+r\mu_{i1})\mu_{i1} - r\mu_{i1}^2}{(1+r\mu_{i1})^2}}{\frac{r\mu_{i1}}{1+r\mu_{i1}}}$$

$$= \frac{\frac{\mu_{i1}}{(1+r\mu_{i1})^2}}{\frac{r\mu_{i1}}{1+r\mu_{i1}}}$$
$$= \frac{1}{r(1+r\mu_{i1})}$$

Let $\psi(x) = \frac{\partial \ln \Gamma(x)}{\partial x}$ be the digamma function. The digamma function has the following difference equation, for integer M:

$$\psi(x+M) - \psi(x) = \sum_{k=0}^{M-1} \frac{1}{x+k}$$

Hence,

$$\begin{aligned}
\frac{\partial C(y_{ij}, \phi)}{\partial r} &= -\frac{\partial \ln \Gamma(\gamma)}{\partial \gamma}\Big|_{\gamma=y_{ij}+1/r} \frac{\partial (y_{ij}+1/r)}{\partial r} + \frac{\partial \ln \Gamma(\gamma)}{\partial \gamma}\Big|_{\gamma=1/r} \frac{\partial (1/r)}{\partial r} \\
&= -\psi(y_{ij}+1/r) \cdot -r^{-2} + \psi(1/r) \cdot -r^{-2} \\
&= \frac{\psi(y_{ij}+1/r)}{r^2} - \frac{\psi(1/r)}{r^2} \\
&= \frac{1}{r^2} \sum_{k=0}^{y_{ij}-1} \frac{1}{\frac{1}{r}+k}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \theta^*_{ij}}{\partial r} &= \frac{\frac{(1+r\mu^*_{ij})(r\frac{\partial \mu^*_{ij}}{\partial r}+\mu^*_{ij}) - r\mu^*_{ij}(r\frac{\partial \mu^*_{ij}}{\partial r}+\mu^*_{ij})}{(1+r\mu^*_{ij})^2}}{\frac{r\mu^*_{ij}}{1+r\mu^*_{ij}}} \\
&= \frac{\frac{r\frac{\partial \mu^*_{ij}}{\partial r}+\mu^*_{ij}}{(1+r\mu^*_{ij})^2}}{\frac{r\mu^*_{ij}}{1+r\mu^*_{ij}}} \\
&= \frac{r\frac{\partial \mu^*_{ij}}{\partial r}+\mu^*_{ij}}{r\mu^*_{ij}(1+r\mu^*_{ij})}
\end{aligned}$$

For $j > 2$,

$$\begin{aligned}
\frac{\partial \mu^*_{ij}}{\partial r} &= \frac{\partial}{\partial r}\left(\mu_{ij} + C_{ijj-1}\frac{\sqrt{1-C^2_{ij-1j-2}}}{\sqrt{1-C^2_{ijj-1}}}\frac{\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}{\sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))}}(Y_{ij-1}-\mu_{ij-1})\right) \\
&= C_{ijj-1}\frac{\sqrt{1-C^2_{ij-1j-2}}}{\sqrt{1-C^2_{ijj-1}}}(Y_{ij-1}-\mu_{ij-1})\frac{\partial}{\partial r}\left(\frac{\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}{\sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))}}\right)
\end{aligned}$$

$$= C_{ijj-1}\frac{\sqrt{1-C^2_{ij-1j-2}}}{\sqrt{1-C^2_{ijj-1}}}(Y_{ij-1}-\mu_{ij-1})$$
$$\times \frac{\sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))}\frac{\partial\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}{\partial r} - \sqrt{E(Var(Y_{ij}|Y_{ij-1}))}\frac{\partial\sqrt{E(Var(Y_{ij-1}|Y_{ij-2}))}}{\partial r}}{E(Var(Y_{ij-1}|Y_{ij-2}))}$$

where

$$\begin{aligned}\frac{\partial\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}{\partial r} &= \frac{1}{2\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}\frac{\partial E(Var(Y_{ij}|Y_{ij-1}))}{\partial r}\\ &= \frac{1}{2\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}\frac{\left(1-r\frac{C^2_{ijj-1}}{1-C^2_{ijj-2}}\right)\mu^2_{ij}+(\mu_{ij}+r\mu^2_{ij})\frac{C^2_{ijj-1}}{1-C^2_{ijj-1}}}{\left(1-r\frac{C^2_{ijj-1}}{1-C^2_{ijj-1}}\right)^2}\\ &= \frac{1}{2\sqrt{E(Var(Y_{ij}|Y_{ij-1}))}}\left(\mu^2_{ij}+\mu_{ij}\frac{C^2_{ijj-1}}{1-C^2_{ijj-1}}\right)\left(1-r\frac{C^2_{ijj-1}}{1-C^2_{ijj-1}}\right)^{-2}\end{aligned}$$

For $j=2$,

$$\begin{aligned}\frac{\mu^*_{i2}}{\partial r} &= \frac{\partial}{\partial r}\left(\mu_{i1}+\frac{C_{i21}}{\sqrt{1-C^2_{i21}}}\frac{\sqrt{E(Var(Y_{i2}|Y_{i1}))}}{\sqrt{Var(Y_{i1})}}(Y_{i1}-\mu_{i1})\right)\\ &= \frac{C_{i21}}{\sqrt{1-C^2_{i21}}}(Y_{i1}-\mu_{i1})\frac{\partial}{\partial r}\left(\frac{\sqrt{E(Var(Y_{i2}|Y_{i1}))}}{\sqrt{Var(Y_{i1})}}\right)\\ &= \frac{C_{i21}}{\sqrt{1-C^2_{i21}}}(Y_{i1}-\mu_{i1})\frac{\sqrt{Var(Y_{i1})}\frac{\partial\sqrt{E(Var(Y_{i2}|Y_{i1}))}}{\partial r}-\sqrt{E(Var(Y_{i2}|Y_{i1}))}\frac{\partial\sqrt{Var(Y_{i1})}}{\partial r}}{Var(Y_{i1})}\end{aligned}$$

where

$$\begin{aligned}\frac{\partial\sqrt{Var(Y_{i1})}}{\partial r} &= \frac{\partial(\mu_{i1}+r\mu^2_{i1})^{1/2}}{\partial r}\\ &= \frac{1}{2}(\mu_{i1}+r\mu^2_{i1})^{-1/2}\cdot\mu^2_{i1}\\ &= \frac{\mu^2_{i1}}{2\sqrt{\mu_{i1}+r\mu^2_{i1}}}\end{aligned}$$

B.4. Expectations

$E((\mu^*_{ij})^2)$ can be expanded to

$$E((\mu^*_{ij})^2) = E\left(\left(\mu_{ij}+C_{ijj-1}\frac{\sigma_{ij}}{\sigma_{ij-1}}(Y_{ij-1}-\mu_{ij-1})\right)^2\right)$$

$$= E\left(\mu_{ij}^2 + 2\mu_{ij}C_{ijj-1}\frac{\sigma_{ij}}{\sigma_{ij-1}}(Y_{ij-1} - \mu_{ij-1}) + C_{ijj-1}^2\frac{\sigma_{ij}^2}{\sigma_{ij-1}^2}(Y_{ij-1} - \mu_{ij-1})^2\right)$$

$$= \mu_{ij}^2 + 2\mu_{ij}C_{ijj-1}\frac{\sigma_{ij}}{\sigma_{ij-1}}E(Y_{ij-1} - \mu_{ij-1}) + C_{ijj-1}^2\frac{\sigma_{ij}^2}{\sigma_{ij-1}^2}E((Y_{ij-1} - \mu_{ij-1})^2)$$

$$= \mu_{ij}^2 + 0 + C_{ijj-1}\frac{\sigma_{ij}^2}{\sigma_{ij-1}^2}\sigma_{ij-1}^2$$

$$= \mu_{ij}^2 + C_{ijj-1}^2\sigma_{ij}^2$$

$$= \mu_{ij}^2 + \frac{C_{ijj-1}^2}{1 - C_{ijj-1}^2}E(Var(Y_{ij}|Y_{ij-1}))$$

APPENDIX C

CODE FOR CHAPTERS 3 AND 4

Provided in this appendix is the R software for the code described in Chapters 3 and 4 on the methodology of maximum-likelihood based analysis of longitudinal data with specified marginal means, first-order antedependence, and linear conditional expectations.

The software requires the use of the alabama package.

```
library(alabama)
```

The function below obtains the necessary information from the dataset provided. This function was written by Matt Guerra.

```
cluster.size = function(id)
 {
 clid = unique(id)
 m = length(unique(id))
 n = rep(0,m)
 autotime = rep(0,0)
 for(i in 1:m)
  {
  n[i] = length(which(id == clid[i]))
  autotime = c(autotime, 1:n[i])
  }
 id = rep(1:m, n)
 return(list(m = m, n = n, id = id, autotime = autotime))
 }
```

The function below formats the dataset and deletes unnecessary information. This function was written by Matt Guerra with additions by Shaun Bender.

```
data.proc = function(data,formula,time=NULL,id,del.n, binom = NULL)
 {
```

```
dat = data.frame(data)
col.name = names(dat)
cluster = cluster.size(id)
m = cluster$m
n = cluster$n
id = cluster$id
if(length(time)==0)
 {
 time = cluster$autotime
 }
autotime = cluster$autotime
index = order(id,time)
if(ncol(dat) == 1)
 {
 dat = dat[index,]
 } else
 {
 dat = dat[index,]
 }
dat = data.frame(dat)
names(dat) = col.name
if(Dist == "Binomial"){binomN = binom[index]} else
{binomN = NULL}
del = which(n <= del.n)
if(length(del) > 0)
 {
 n = n[-del]
 m = length(n)
 mtch = match(id, del)
 del.id = which(mtch != "NA")
 dat = dat[-del.id,]
```

```
  dat = data.frame(dat)
  names(dat) = col.name
  row.names(dat) = 1:nrow(dat)
  time = time[-del.id]
  autotime = autotime[-del.id]
  id = rep(1:m, n)
  if(Dist == "Binomial"){binomN = binomN[-del.id]}
  }
 formula = as.formula(formula)
 fml = as.formula(paste("~", formula[3], "+", formula[2], sep=""))
 dat = model.matrix(fml, data=dat)
 return(list(data = dat, time = time, autotime = autotime, id = id, m = m, n = n,
             binomN = binomN))
}
```

The function returns the value of the log likelihood for the inputted parameter values and a given dataset. This function was written by Victoria Gamerman with additions by Shaun Bender.

```
drv.logl = function(start.values)
 {
 if(Dist == "Negative-Binomial"){Anc = 1}
 if(Dist != "Negative-Binomial"){Anc = 0}

 if(CorrStr == "AR(1)" | CorrStr == "Markov")
  {
  alpha = start.values[1]
  beta = start.values[2:(length(start.values)-Anc)]
  if(Dist == "Negative-Binomial"){r = start.values[length(start.values)]}
  }
 if(CorrStr == "AD(1)")
  {
  alpha = start.values[1:max(n)-1]
```

```
 beta = start.values[max(n):(length(start.values)-Anc)]
 if(Dist == "Negative-Binomial"){r = start.values[length(start.values)]}
 }
LogLik = 0
for (i in 1:m)
 {
 data_i = matrix(NA, nrow=n[i], ncol=dim(dataset$data)[2])
 data_i[1:n[i],1:dim(dataset$data)[2]] = dataset$data[which(id==i),]
 data.end = ncol(data_i)
 x_i = matrix(NA, nrow=n[i], ncol=k+1)
 x_i[1:n[i],1:(k+1)] = data_i[,-data.end]
 y_i = data_i[,data.end]
 n_i = nrow(data_i)
 time_i = dataset$time[which(id==i)]
 for (j in 1:n_i)
  {
  if (j == 1)
   {
   lam_ij = LinkInv(i, j, beta, x_i)
   if(Dist == "Negative-Binomial"){LogLik = LogLik +
                                    UnitLikelihood(i, j, y_i, lam_ij, r)}
   else {LogLik = LogLik + UnitLikelihood(i, j, y_i, lam_ij)}
   }
  if (j > 1)
   {
   lam = c(LinkInv(i, j, beta, x_i), LinkInv(i, j-1, beta, x_i))
   lamdot_i2 = MuSt_ij(i, j, y_i, time_i, alpha, lam, r)

   if(Dist == "Negative-Binomial"){LogLik = LogLik +
                                    UnitLikelihood(i, j, y_i, lamdot_i2, r)}
   else {LogLik = LogLik + UnitLikelihood(i, j, y_i, lamdot_i2)}
```

```
    }
   }
  }
 return(-LogLik)
 }
```

The function returns the value of the gradient of the log likelihood for the inputted parameter values and a given dataset. This function was written by Victoria Gamerman with additions by Shaun Bender.

```
drv.grad = function(start.values)
 {
 D_Beta = matrix(0, nrow = k+1, ncol = 1)
 D_R = matrix(0, nrow = 1, ncol = 1)
 if(Dist == "Negative-Binomial"){Anc = 1}
 if(Dist != "Negative-Binomial"){Anc = 0}
 if(CorrStr == "AR(1)" | CorrStr == "Markov")
  {
  alpha = start.values[1]
  beta = start.values[2:(length(start.values)-Anc)]
  if(Dist == "Negative-Binomial"){r = start.values[length(start.values)]}
  if(Dist != "Negative-Binomial"){r = NULL}
  D_Alpha = matrix(0, nrow = 1, ncol = 1)
  }
 if(CorrStr == "AD(1)")
  {
  alpha = start.values[1:max(n)-1]
  beta = start.values[max(n):(length(start.values)-Anc)]
  if(Dist == "Negative-Binomial"){r = start.values[length(start.values)]}
  if(Dist != "Negative-Binomial"){r = NULL}
  D_Alpha = matrix(0, nrow = max(n)-1, ncol = 1)
  }
```

```
for (i in 1:m)
 {
 data_i = matrix(NA, nrow = n[i], ncol = dim(dataset$data)[2])
 data_i[1:n[i],1:dim(dataset$data)[2]] = dataset$data[which(id==i),]
 data.end = ncol(data_i)
 x_i = matrix(NA, nrow = n[i], ncol = k+1)
 x_i[1:n[i],1:(k+1)] = data_i[,-data.end]
 y_i = data_i[,data.end]
 n_i = nrow(data_i)
 time_i = dataset$time[which(id==i)]
 if(n_i>=1)
  {
  for(j in 1:n_i)
   {
   if(j == 1)
    {
    lam_ij = LinkInv(i, j, beta, x_i)
    if(Dist == "Negative-Binomial")
     {
     D_Beta = D_Beta + functBeta(i, j, x_i, y_i, time_i, alpha, lam_ij, r)
     D_R = D_R + functR(i, j, r, y_i, time_i, lam_ij, alpha)
     }
    if(Dist != "Negative-Binomial")
     {
     D_Beta = D_Beta + functBeta(i, j, x_i, y_i, time_i, alpha, lam_ij)
     }
    }
   if(j > 1)
    {
    lam = c(LinkInv(i, j, beta, x_i), LinkInv(i, j-1, beta, x_i))
    D_Alpha = D_Alpha + functAlpha(i, j, y_i, time_i, alpha, lam, r)
```

```
    if(Dist == "Negative-Binomial")
     {
     D_Beta = D_Beta + functBeta(i, j, x_i, y_i, time_i, alpha, lam, r)
     D_R = D_R + functR(i, j, r, y_i, time_i, lam, alpha)
     }
    if(Dist != "Negative-Binomial")
     {
     D_Beta = D_Beta + functBeta(i, j, x_i, y_i, time_i, alpha, lam)
     }
    }
   }
  }
 }
 Output = t(t(c(-D_Alpha, -D_Beta)))
 if(Dist == "Negative-Binomial"){Output = t(t(c(-D_Alpha, -D_Beta, -D_R)))}
 return(-Output)
 }
```

This function calculates the inverse of the link function used. Note that for the Negative Binomial case, a non-canonical log inverse link is used. This function was written by Shaun Bender.

```
LinkInv = function(i, j, beta, x_i)
 {
 if(Dist == "Poisson")
  {
  lam_ij = exp(t(beta)%*%x_i[j,])
  }
 if(Dist == "Negative-Binomial")
  {
  lam_ij = exp(t(beta)%*%x_i[j,])
  if(lam_ij[1] == 0){lam_ij[1] = .001}
  }
```

```
 if(Dist == "Binomial")
  {
  N = binomN[which(id==i)][j]
  lam_ij = N * exp(t(beta)%*%x_i[j,]) / (1 + exp(t(beta)%*%x_i[j,]))
  }
 return(lam_ij[1])
 }
```

This function calculates the log likelihood for a single unit. This function was written by Shaun Bender.

```
UnitLikelihood = function(i, j, y_i, lam_ij, r)
 {
 if(Dist == "Poisson")
  {
  Theta = log(lam_ij)
  B_Theta = lam_ij
  if(y_i[j] == 0){Const = 0}
  if(y_i[j] > 0){Const = sum(log(seq(from = 1, to = y_i[j], by = 1)))}
  }
 if(Dist == "Negative-Binomial")
  {
  Theta = log(r * lam_ij / (1 + r * lam_ij))
  B_Theta = 1 / r * log(1 + r * lam_ij)
  Const = -lgamma(y_i[j] + 1/r) + lgamma(y_i[j] + 1) + lgamma(1/r)
  }
 if(Dist == "Binomial")
  {
  N = binomN[which(id==i)][j]
  Theta = log(lam_ij / (N - lam_ij))
  B_Theta = -N * log((N - lam_ij) / N)
  Const = -log(choose(N, y_i[j]))
```

```
 }
Unit = y_i[j]*Theta - B_Theta - Const
return(Unit)
}
```

This function calculates the derivative of the log likelihood with respect to r, under the assumption of a Negative-Binomial distribution. This function was written by Shaun Bender.

```
functR = function(i, j, r, y_i, time_i, lam, alpha)
 {
 lam_ij = lam[1]
 if(j == 1){DTheta = 1 / r / (1 + r * lam_ij)}
 if(j > 1)
  {
  lam_ij_1 = lam[2]
  Corr = FindCorr(j, time_i, alpha)
  C_ij = Corr[1]
  if(j>2){C_ij_1 = Corr[2]}
  E_V = FindE_V(i, j, lam, Corr, r)
  E_V_ij_1 = E_V[1]
  E_V_ij = E_V[2]
  if(j == 2){DE_V_ij_1 = lam_ij_1^2 / 2 / sqrt(lam_ij_1 + r * lam_ij_1^2)}
  if(j > 2){DE_V_ij_1 = 1/2/sqrt(E_V_ij_1) * (lam_ij_1^2+lam_ij_1*C_ij_1^2/
            (1 - C_ij_1^2)) / (1-r*C_ij_1^2/(1-C_ij_1^2))^2}
  DE_V_ij = 1/2/sqrt(E_V_ij) * (lam_ij^2+lam_ij*C_ij^2/(1 - C_ij^2)) /
            (1-r*C_ij^2/(1-C_ij^2))^2
  Num = sqrt(E_V_ij_1) * DE_V_ij - E_V_ij * DE_V_ij_1
  if(j == 2){DMuSt = C_ij / sqrt(1-C_ij^2) * (y_i[j-1]-lam_ij_1) * Num /
                     E_V_ij_1}
  if(j > 2){DMuSt = C_ij * sqrt(1-C_ij_1^2)/sqrt(1-C_ij^2)*(y_i[j-1]-lam_ij_1) *
            Num / E_V_ij_1}
  MuSt = MuSt_ij(i, j, y_i, time_i, alpha, lam, r)
```

```
 DTheta = (r * DMuSt + MuSt) / r / MuSt / (1 + r * MuSt)
 }
DC = 0
if(y_i[j] != 0){for(i in 0:(y_i[j]-1)){DC = DC + (1/r^2) * 1 / (1/r + i)}}
Output = (y_i[j]-lam_ij) * DTheta - DC
return(Output)
}
```

This function calculates the term to be added for the derivative of the log-likelihood. This function was written by Shaun Bender.

```
functAlpha = function(i, j, y_i, time_i, alpha, lam, r)
 {
 MuSt = MuSt_ij(i, j, y_i, time_i, alpha, lam, r)
 Added = (y_i[j]-MuSt)*DerivG(i, MuSt, r)*DMuStarAlpha(i, j, y_i, time_i, alpha,
                                                       lam, r)
 return(Added)
 }
```

This function calculates the value of μ_{ij}^*. This function was written by Shaun Bender.

```
MuSt_ij = function(i, j, y_i, time_i, alpha, lam, r)
 {
 Corr = FindCorr(j, time_i, alpha)
 Corr_ij = Corr[1]
 Corr_ij_1 = Corr[2]
 E_V = FindE_V(i, j, lam, Corr,r)
 E_V_ij = E_V[1]
 E_V_ij_1 = E_V[2]
 lam_ij = lam[1]
 lam_ij_1 = lam[2]
 if(j == 2)
  {
```

```
 MuSt = lam_ij + Corr_ij/sqrt(1-Corr_ij^2) * sqrt(E_V_ij / E_V_ij_1) *
        (y_i[j-1] - lam_ij_1)
 }
if(j > 2)
 {
 MuSt = lam_ij + Corr_ij * sqrt(E_V_ij / E_V_ij_1) * sqrt((1-Corr_ij_1^2) /
        (1-Corr_ij^2)) * (y_i[j-1] - lam_ij_1)
 }
constr = sqrt(lam_ij / (lam_ij_1 + lam_ij))
if(is.finite(constr) == FALSE){constr = 0.2}
if(is.finite(MuSt) == FALSE){MuSt = 0.5*constr}
if(MuSt < 0){MuSt = 0.5*constr}
return(MuSt)
}
```

This function calculates $\frac{\partial \mu_{ij}^*}{\partial \alpha}$, the derivative of μ_{ij}^* with respect to α. This function was written by Shaun Bender.

```
DMuStarAlpha = function(i, j, y_i, time_i, alpha, lam, r)
 {
 Corr = FindCorr(j, time_i, alpha)
 Corr_ij = Corr[1]
 Corr_ij_1 = Corr[2]
 E_V = FindE_V(i, j, lam, Corr,r)
 E_V_ij = E_V[1]
 E_V_ij_1 = E_V[2]
 lam_ij = lam[1]
 lam_ij_1 = lam[2]
 if(CorrStr == "AR(1)")
 {
 if(j == 2){Output = sqrt(E_V_ij/E_V_ij_1) * (y_i[1] - lam_ij_1) /
                     (1-alpha^2)^(3/2)}
```

```
 if(j > 2){Output = sqrt(E_V_ij/E_V_ij_1) * (y_i[j-1] - lam_ij_1)}
 }
if(CorrStr == "Markov")
 {
 if(j == 2)
  {
  Output = (y_i[1]-lam_ij_1) / sqrt(1-alpha^(2*time_i[2]-2*time_i[1])) *
           sqrt(E_V_ij/E_V_ij_1) * (time_i[2]-time_i[1]) *
           alpha^(time_i[2]-time_i[1]-1) * (1 + alpha^(2*time_i[2]-2*time_i[1])/
           (1-alpha^(2*time_i[2]-2*time_i[1])))
  }
 if(j > 2)
  {
  Output = (y_i[j-1]-lam_ij_1)*sqrt(1-alpha^(2*time_i[j-1]-2*time_i[j-2])) /
           sqrt(1-alpha^(2*time_i[j]-2*time_i[j-1])) *
           sqrt(E_V_ij/E_V_ij_1)*alpha^(time_i[j]-time_i[j-1]-1) *
           ((time_i[j]-time_i[j-1])/
           (1-alpha^(2*time_i[j]-2*time_i[j-1]))-(time_i[j-1]-time_i[j-2]) *
           alpha^(2*time_i[j-1]-2*time_i[j-2]) /
           (1-alpha^(2*time_i[j-1]-2*time_i[j-2])))
  }
 }
if(CorrStr == "AD(1)")
 {
 if(j == 2)
  {
  One = c(1, rep(0,(length(alpha)-1)))
  Output = sqrt(E_V_ij/E_V_ij_1) * (y_i[1] - lam_ij_1) * One /
           (1-alpha[1]^2)^(3/2)
  }
 if(j > 2)
```

```
   {
   J_1 = c(rep(0,j-2),1,rep(0,length(alpha)-j+1))
   J_2 = c(rep(0,j-3),1,rep(0,length(alpha)-j+2))
   Output = sqrt(E_V_ij/E_V_ij_1) * (y_i[1] - lam_ij_1) *
            (J_1 * sqrt(1-alpha[j-2]^2)/(1-alpha[j-1]^2)^(3/2) -
   J_2 * alpha[j-1] * alpha[j-2] / sqrt(1-alpha[j-1]^2) / sqrt(1-alpha[j-2]^2))
   }
  }
 return(Output)
 }
```

This function calculates the term to be added for the derivative of the log-likelihood. This function was written by Shaun Bender.

```
functBeta = function(i, j, x_i, y_i, time_i, alpha, lam, r)
 {
 if(j == 1){Added = (y_i[j]-lam) * DerivG(i, lam, r) * Dlam_ij(j, lam, x_i)}
 if(j > 1)
  {
  MuSt = MuSt_ij(i, j, y_i, time_i, alpha, lam, r)
  Added = (y_i[j]-MuSt)*DerivG(i, MuSt, r)*DMuStarBeta(i, j, x_i, y_i, time_i,
                                                       alpha, lam, r)
  }
 return(Added)
 }
```

This function calculates $\frac{\partial \mu_{ij}^*}{\partial \beta}$, the derivative of μ_{ij}^* with respect to β. This function was written by Shaun Bender.

```
DMuStarBeta = function(i, j, x_i, y_i, time_i, alpha, lam,r)
 {
 Corr = FindCorr(j, time_i, alpha)
 Corr_ij = Corr[1]
```

```
 Corr_ij_1 = Corr[2]
 E_V = FindE_V(i, j, lam, Corr, r)
 E_V_ij = E_V[1]
 E_V_ij_1 = E_V[2]
 lam_ij = lam[1]
 lam_ij_1 = lam[2]
 if(j == 2)
  {
  Output = Dlam_ij(j, lam_ij, x_i) + Corr_ij/sqrt(1-Corr_ij^2) *
           sqrt(E_V_ij/E_V_ij_1)*((y_i[1]-lam_ij_1)/2 *
           (DE_V(i,j,lam_ij,x_i,time_i,alpha,r)/E_V_ij -
           DE_V(i,j-1,lam_ij_1,x_i,time_i,alpha,r)/E_V_ij_1) -
           Dlam_ij(j-1, lam_ij_1, x_i))
  }
 if(j > 2)
  {
  Output = Dlam_ij(j, lam_ij, x_i) + Corr_ij*sqrt(1-Corr_ij_1^2) /
           sqrt(1-Corr_ij^2)*sqrt(E_V_ij/E_V_ij_1)*
           ((y_i[j-1]-lam_ij_1)/2*(DE_V(i,j,lam_ij,x_i,time_i,alpha,r)/E_V_ij -
           DE_V(i,j-1,lam_ij_1,x_i,time_i,alpha,r)/E_V_ij_1) -
           Dlam_ij(j-1, lam_ij_1, x_i))
  }
 return(Output)
 }
```

This function calculates the adjacent correlations at time j (and $j-1$ for $j > 2$). This function was written by Shaun Bender.

```
FindCorr = function(j, time_i, alpha)
 {
 if(CorrStr == "AR(1)")
  {
```

```
 Corr_ij = alpha
 if(j > 2){Corr_ij_1 = alpha}
 }
if(CorrStr == "Markov")
 {
 Corr_ij = alpha^(time_i[j]-time_i[j-1])
 if(j > 2){Corr_ij_1 = alpha^(time_i[j-1]-time_i[j-2])}
 }
if(CorrStr == "AD(1)")
 {
 Corr_ij = alpha[j-1]
 if(j > 2){Corr_ij_1 = alpha[j-2]}
 }
if(j == 2){Output = Corr_ij}
if(j > 2){Output = c(Corr_ij, Corr_ij_1)}
return(Output)
}
```

This function calculates $E(Var(Y_{ij}|Y_{ij-1}))$ (and if $j > 2$, $E(Var(Y_{ij_1}|Y_{ij_2}))$). This function was written by Shaun Bender.

```
FindE_V = function(i, j, lam, Corr, r)
 {
 lam_ij = lam[1]
 lam_ij_1 = lam[2]
 Corr_ij = Corr[1]
 if(j > 2){Corr_ij_1 = Corr[2]}
 if(Dist == "Poisson")
  {
  E_V_ij = lam_ij
  E_V_ij_1 = lam_ij_1
  }
```

```
 if(Dist == "Negative-Binomial")
  {
  C_ij = 1 - r * Corr_ij^2 / (1 - Corr_ij^2)
  if(j > 2){C_ij_1 = 1 - r * Corr_ij_1^2 / (1 - Corr_ij_1^2)}
  E_V_ij = (lam_ij + r * lam_ij^2) / C_ij
  if(j == 2){E_V_ij_1 = lam_ij_1 + r * lam_ij_1^2}
  if(j > 2){E_V_ij_1 = (lam_ij_1 + r * lam_ij_1^2) / C_ij_1}
  }
 if(Dist == "Binomial")
  {
  N = binomN[which(id==i)][j]
  if(j > 1){N_1 = binomN[which(id==i)][j-1]}
  C_ij = 1 + Corr_ij^2 / (1 - Corr_ij^2) / N
  if(j > 2){C_ij_1 = 1 + Corr_ij_1^2 / (1 - Corr_ij_1^2) / N_1}
  E_V_ij = lam_ij * (N - lam_ij) / N / C_ij
  if(j == 2){E_V_ij_1 = lam_ij_1 * (N_1 - lam_ij_1) / N_1}
  if(j > 2){E_V_ij_1 = lam_ij_1 * (N_1 - lam_ij_1) / N_1 / C_ij_1}
  }
 Output = c(E_V_ij, E_V_ij_1)
 return(Output)
 }
```

This function calculates $\frac{\partial E(Var(Y_{ij}|Y_{ij-1}))}{\partial \beta}$, the derivative of $E(Var(Y_{ij}|Y_{ij-1}))$ with respect to β. This function was written by Shaun Bender.

```
DE_V = function(i, j, lam_ij, x_i, time_i, alpha, r)
 {
 if(Dist == "Poisson"){Output = Dlam_ij(j, lam_ij, x_i)}
 if(Dist == "Negative-Binomial")
  {
  if(j == 1){Output = (2 * r * lam_ij +1) * Dlam_ij(j, lam_ij, x_i)}
  if(j > 1)
```

```
   {
   Corr_ij = FindCorr(j, time_i, alpha)[1]
   C_ij = 1 - r * Corr_ij^2 / (1 - Corr_ij^2)
   Output = (2 * r * lam_ij + 1) * Dlam_ij(j, lam_ij, x_i) / C_ij
   }
  }
 if(Dist == "Binomial")
  {
  N = binomN[which(id==i)][j]
  if(j == 1){Output = (K - 2 * lam_ij) * Dlam_ij(j, lam_ij, x_i) / N}
  if(j > 1)
   {
   Corr_ij = FindCorr(j, time_i, alpha)[1]
   C_ij = 1 + Corr_ij^2 / (1 - Corr_ij^2) / N
   Output = (N - 2 * lam_ij) * Dlam_ij(j, lam_ij, x_i) / N / C_ij
   }
  }
 return(Output)
 }
```

This function calculates $\frac{\partial \mu_{ij}}{\partial \beta}$, the derivative of μ_{ij} with respect to β. This function was written by Shaun Bender.

```
Dlam_ij = function(j, lam_ij, x_i)
 {
 if(Dist == "Poisson")
  {
  Dlam = x_i[j,] * lam_ij
  }
 if(Dist == "Negative-Binomial")
  {
  Dlam = x_i[j,] * lam_ij
```

```
 }
if(Dist == "Binomial")
 {
 Dlam = x_i[j,] * lam_ij / (1 + exp(t(beta)%*%x_i[j,]))
 }
return(Dlam)
}
```

This function calculates $\frac{\partial g(\gamma)}{\partial \gamma}\Big|_{\gamma=\text{Eval}}$, the derivative of the link function $g()$ evaluated at "Eval". This function was written by Shaun Bender.

```
DerivG = function(i, Eval, r)
 {
 if(Dist == "Poisson")
  {
  Output = 1 / Eval
  }
 if(Dist == "Negative-Binomial")
  {
  Output = 1 / Eval / (1 + r * Eval)
  }
 if(Dist == "Binomial")
  {
  N = binomN[which(id==i)][j]
  Output = N / Eval / (N - Eval)
  }
 return(Output)
 }
```

This function organizes the output and computes several statistics of interest. This function was written by Shaun Bender, based on code by Victoria Gamerman.

```
CompileResults = function(model, formula, N_Alp, N_Var, N_Subjects)
```

```
{
mle.alpha = model$par[1:N_Alp]
mle.beta = model$par[(N_Alp+1):(N_Alp+N_Var)]
if(Dist == "Negative-Binomial"){mle.r = model$par[N_Alp+N_Var+1]}
mle.full = -model$value
mle.cov = solve(model$hessian)
AIC = 2*(N_Var+1)-2*(mle.full)
BIC = log(N_Subjects)*(length(mle.beta)+1)-2*(mle.full)
Stderr = matrix(NA, nrow = N_Var, ncol = 1)
Wald = matrix(NA, nrow = N_Var, ncol = 1)
pval = matrix(NA, nrow = N_Var, ncol = 1)
for(p in (N_Alp+1):(N_Alp+N_Var))
 {
 Stderr[p-N_Alp,] = sqrt(mle.cov[(p),(p)])
 Wald[p-N_Alp,] = (mle.beta[p-N_Alp] / sqrt(mle.cov[(p),(p)]))^2
 pval[p-N_Alp,] = 1-pchisq(Wald[p-N_Alp,1], df = 1, lower.tail = TRUE,
                           log.p = FALSE)
 }
results = cbind(mle.beta, Stderr, Wald, pval)
Alpha_Cov = NULL
for(p in 1:N_Alp)
 {
 Alpha_Cov = c(Alpha_Cov, sqrt(mle.cov[p,p]))
 }
alpha_results = cbind(mle.alpha,Alpha_Cov)
fit_stats = rbind(mle.full, AIC, BIC)
if(Dist == "Negative-Binomial")
 {
 r_Cov = NULL
 r_Cov = sqrt(mle.cov[N_Alp+N_Var+1,N_Alp+N_Var+1])
 r_results = cbind(mle.r,r_Cov)
```

```
 colnames(r_results) = c("Estimate", "Std.err")
 rownames(r_results) = "r"
 }
#format output
rownames(fit_stats) = c("Log-Likelihood:", "AIC:", "BIC:")
colnames(fit_stats) = c("")
colnames(results) = c("Estimate", "Std.err", "Wald", "Pr(>|W|)")
rownames(results) = c("(Intercept)", all.vars(formula[[3]]))
colnames(alpha_results) = c("Estimate", "Std.err")
rownames(alpha_results) = c(rep("alpha",N_Alp))
if(Dist != "Negative-Binomial"){return(list(fit_stats, results, alpha_results))}
if(Dist == "Negative-Binomial"){return(list(fit_stats, results, alpha_results,
                                r_results))}
}
```

The function below calculates the value of the constraints. This function was written by Shaun Bender.

```
Constraints = function(Values)
 {
 Last = length(Values)
 Output = rep(NA, 1)
 if(CorrStr == "AR(1)")
  {
  Output[1] = 1 - Values[1]
  Output[2] = Values[1] + 1
  if(Dist == "Negative-Binomial"){Output[3] = - Values[1]^2 + 1/(Values[Last]+1)}
  }
 if(CorrStr == "Markov")
  {
  Output[1] = 1-Values[1]
  Output[2] = Values[1] + 1
```

```
 if(Dist == "Negative-Binomial"){Output[3] = -Values[1]^2 + 1/(Values[Last]+1)}
 }
if(CorrStr == "AD(1)")
 {
 for(i in 0:(N_Alp-1))
  {
  Output[2*i+1] = 1 - Values[i+1]
  Output[2*i+2] = Values[i+1] + 1
  }
 if(Dist == "Negative-Binomial")
  {
  for(i in 1:N_Alp)
   {
   Output[2*N_Alp+i] = - Values[i]^2 + 1/(Values[Last]+1)
   }
  }
 }
return(Output)
}
```

The function belew calculates the value of the Jacobian of the constraints. This function was written by Shaun Bender.

```
ConstraintsJacobian = function(Values)
 {
 Last = length(Values)
 if(CorrStr == "AR(1)")
  {
  if(Dist != "Negative-Binomial"){Output = matrix(0, 2, Last)}
  if(Dist == "Negative-Binomial")
   {
   Output = matrix(0, 3, length(Values))
```

```
  Output[3,1] = -2*Values[1]
  Output[3,Last] = 1 / (Values[Last] + 1)^2
  }
 Output[1,1] = -1
 Output[2,1] = 1
 }
if(CorrStr == "Markov")
 {
 if(Dist != "Negative-Binomial"){Output = matrix(0, 2, Last)}
 if(Dist == "Negative-Binomial")
  {
  Output = matrix(0, 3, length(Values))
  Output[3,1] = -2*Values[1]
  Output[3,Last] = 1 / (Values[Last] + 1)^2
  }
 Output[1,1] = -1
 Output[2,1] = 1
 }
if(CorrStr == "AD(1)")
 {
 if(Dist != "Negative-Binomial"){Output = matrix(0, 2*N_Alp, Last)}
 if(Dist == "Negative-Binomial")
  {
  Output = matrix(0, 3*N_Alp, Last)
  for(i in 1:N_Alp)
   {
   Output[2*N_Alp+i,i] = -2*Values[i]
   Output[2*N_Alp+i,Last] = 1 / (Values[Last] + 1)^2
   }
  }
 for(i in 0:(N_Alp-1))
```

```
  {
  Output[2*i+1,i+1] = -1
  Output[2*i+2,i+1] = 1
  }
 }
return(Output)
}
```

The function below is the function that is called by the user. It calls the above functions in order to organize the data, run the methodology using the alabama package, and organize the results for presentation.

```
EndResults = function(formula, CorrInput, Dist, DatasetInput, IDInput, TimeInput,
                      start.values, DistOpts)
 {
 id = IDInput
 t = TimeInput
 d = dim(DatasetInput)
 k <<- length(all.vars(formula))-1
 dt.fm = data.frame(DatasetInput)
 Dist <<- Dist
 if(Dist == "Binomial")
  {
  dataset <<- data.proc(data = dt.fm, formula = formula, time = t, id = id,
                        del.n = 0, binom = DistOpts)
  }
 if(Dist != "Binomial")
  {
  dataset <<- data.proc(data = dt.fm, formula = formula, time = t, id = id,
                        del.n = 0, binom = NULL)
  }
 m <<- dataset$m
```

```
n <<- dataset$n
id <<- dataset$id
.GlobalEnv$time <- dataset$time
autotime <<- dataset$autotime
binomN <<- dataset$binomN
CorrStr <<- CorrInput
N_Var = length(all.vars(formula[[3]]))+1
if(CorrStr == "AR(1)" | CorrStr == "Markov"){N_Alp <<- 1}
if(CorrStr == "AD(1)"){N_Alp <<- length(unique(time))-1}
lb = rep(-Inf, length(start.values))
ub = rep(Inf, length(start.values))
lb[1:N_Alp] = rep(-1, N_Alp)
ub[1:N_Alp] = rep(1, N_Alp)
full.ml = auglag(par = start.values, fn = drv.logl, hin = Constraints,
                 hin.jac = ConstraintsJacobian, control.outer = list("itmax" =
                 1000, "trace" = FALSE, ilack.max = 10, eps = 10^-8))
Output = CompileResults(full.ml, formula, N_Alp, N_Var, m)
rm(list = c('k','dataset','m','n','time','autotime','CorrStr','K'),
            pos = ".GlobalEnv")
return(Output)
}
```

BIBLIOGRAPHY

Bartlett, JW, Carpenter, JR, Tilling, K, and Vansteelandt, S (2014). Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics* 15.4, 719–730. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxu023. URL: http://www.ncbi.nlm.nih.gov/pubmed/24907708.

Bradley, RA and Gart, JJ (1962). The Asymptotic Properties of ML Estimators when Sampling from Associated Populations. *Biometrika* 49.1/2, 205–214. ISSN: 00063444. DOI: 10.2307/2333482. URL: http://www.jstor.org/stable/2333482.

Broyden, CG (1970a). The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics* 6.1, 76–90.

Broyden, CG (1970b). The Convergence of a Class of Double-rank Minimization Algorithms: 2. The New Algorithm. *IMA Journal of Applied Mathematics* 6.3, 222–231.

Chernoff, H (1954). On the Distribution of the Likelihood Ratio. *The Annals of Mathematical Statistics* 25.3, 573–578.

Crowder, M (1995). On the Use of a Working Correlation Matrix in Using Generalised Linear Models for Repeated Measures. *Biometrika1* 82.2, 407–410.

Dienemann, T, Bender, S, Wilson, FP, Reese, P, Long, J, and Leonard, MB. Improvements in muscle mass and function following renal transplantation.

Diggle, PJ, Heagerty, P, Liang, KY, and Zeger, SL (2002). *Analysis of Longitudinal Data*. 2nd. Oxford: Oxford University Press.

Donabedian, A (1966). Evaluating the Quality of Medical Care. *The Milbank Memorial Fund Quarterly* 44.3, 166–206.

Fitzmaurice, GM, Laird, NM, and Ware, JH (2011). *Applied Longitudinal Analysis*. 2nd. Hoboken, NJ: John Wiley & Sons, Inc.

Fletcher, R (1970). A New Approach to Variable Metric Algorithms. *The Computer Journal* 13.3, 317–322.

Frees, EW (2004). *Longitudinal and Panel Data: Analysis and Applications In the Social Sciences*. Cambridge, United Kingdom: Cambridge University press.

Gabriel, K (1962). Ante-dependence analysis of an ordered set of variables. *The Annals of Mathematical Statistics* 33.1, 201–212. URL: http://www.jstor.org/stable/10.2307/2237651.

Galati, JC and Seaton, KA (2013). MCAR is not necessary for the complete cases to constitute a simple random subsample of the target sample. *Statistical Methods in Medical Research*. ISSN: 1477-0334. DOI: 10.1177/0962280213490360. URL: http://www.ncbi.nlm.nih.gov/pubmed/23698868.

Gamerman, V, Guerra, M, and Shults, J (2016). Maximum-likelihood based analysis of equally spaced longitudinal count data with specified marginal means, first-order antedependence, and linear conditional expectations. URL: http://biostats.bepress.com/upennbiostat/art45.

Goldfarb, D (1970). A Family of Variable-Metric Methods Derived by Variational Means. *Mathematics of Computation* 24.109, 23–26.

Guerra, MW and Shults, J (2014). A Note on the Simulation of Overdispersed Random Variables With Specified Marginal Means and Product Correlations. *The American Statistician* 68.2, 104–107. ISSN: 0003-1305. DOI: 10.1080/00031305.2014.887592. URL: http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2014.887592{\#}.U5LFQXWSzmE.

Guerra, MW, Shults, J, Amsterdam, J, and Ten-Have, T (2012). The analysis of binary longitudinal data with time-dependent covariates. *Statistics in Medicine* 31.10, 931–948. ISSN: 02776715. DOI: 10.1002/sim.4465.

Heitjan, DF (1997). Ignorability, Sufficiency and Ancillarity. *Journal of the Royal Statistical Society* 59.2, 375–381.

Hilbe, JM (2011). *Negative Binomial Regression*. 2nd. Cambridge, United Kingdom: Cambridge University Press.

Horton, NJ and Kleinman, KP (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* 61.1, 79–90. ISSN: 0003-1305. DOI: 10.1198/000313007X172556. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1839993{\&}tool=pmcentrez{\&}rendertype=abstract.

Liang, KY and Zeger, SL (1986). Longitudinal Data Analysis Using Generalized Linear Models Author. *Biometrika* 73.1, 13–22.

Little, RJ (1992). Regression With Missing X ' s : A Review. *Journal of the American Statistical Association* 87.420, 1227–1237.

Little, RJ and Rubin, DB (2002). *Statistical Analysis with Missing Data*. 2nd. Hoboken, NJ: John Wiley & Sons, Inc.

Little, RJ and Zhang, N (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society Series C Applied Statistics* 60.4, 591–605.

McCullagh, P and Nelder, J (1989). *Generalized Linear Models*. 2nd. Boca Raton, Florida: Chapman \& Hall/CRC.

McDaniel, LS and Henderson, N (2015). *geeM: Fit Generalized Estimating Equations*. R package version 0.7.3. URL: http://CRAN.R-project.org/package=geeM.

Olmsted, MG, Geisen, E, Murphy, J, Bell, D, Morley, M, and Stanley, M (2015). *Methodology: U.S. News & World Report Best Hospitals 2015-16*. Tech. rep.

Pan, W (2001). Akaike's Information Criterion in Generalized Estimating Equations. *International Biometric Society* 57.1, 120–125.

Prentice, RL (1988). Correlated Binary Regression with Covariates Specific to Each Binary Observation. *Biometrics* 44.4, 1033–1048.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Robins, JM, Rotnitzky, A, and Zhao, LP (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association* 89.427, 846–866.

Rubin, DB (1976). Inference and missing data. *Biometrika* 63.3, 581–592.

Rubin, DB (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Seaman, S, Galati, J, Jackson, D, and Carlin, J (2013). What Is Meant by Missing at Random? *Statistical Science* 28.2, 257–268. ISSN: 0883-4237. DOI: 10.1214/13-STS415. arXiv:arXiv:1306.2812v1. URL: http://projecteuclid.org/euclid.ss/1369147915.

Shanno, D (1970). Conditioning of Quasi-Newton Methods for Function Minimization. *Mathematics of Computation* 24.111, 647–656.

Simonoff, JS (1988). Regression Diagnostics to Detect Nonrandom Missingness in Linear Regression. *Technometrics* 30.2, 205–214.

Sutradhar, BC and Das, K (1999). On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika* 86.2, 459–465.

Sutradhar, BC and Das, K (2000). On the Accuracy of Efficiency of Estimating Equation Approach. *Biometrics* 56.2, 622–625.

Thall, PF and Vail, SC (1990). Some Covariance Models for Longitudinal Count Data with Overdispersion. *Biometrics* 46.3, 657–671.

Varadhan, R (2015). *alabama: Constrained Nonlinear Optimization*. R package version 2015.3-1. URL: http://CRAN.R-project.org/package=alabama.

White, IR and Carlin, JB (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine* 29.28, 2920–2931. ISSN: 1097-0258. DOI: 10.1002/sim.3944. URL: http://www.ncbi.nlm.nih.gov/pubmed/20842622.

Zhang, H, Xia, Y, Chen, R, Gunzler, D, Tang, W, and Tu, X (2011). Modeling longitudinal binomial responses : implications from two dueling paradigms. *Journal of Applied Statistics* 38.11, 2373–2390.