Mixed Poisson Regression Models Using GEE and Gibbs Sampling Estimation Techniques: An Application to a Longitudinal Study of Alcohol Use Among Youth

Chin-Chien Yang

Social Research Methodology Division
Department of Education
Graduate School of Education & Information Studies
University of California, Los Angeles
USA

Hsiang-Chuan Liu

National Taichung Teachers' College
Taiwan, R.O.C.

293
1. Introduction

Consider a longitudinal data set consisting of a response variable $Y_{it}$ of count and a $p \times l$ vector $X_{it}$ of covariates observed at times $t = 1, \ldots, n_i$, for independent subjects $i = 1, \ldots, M$. Such data frequently arise in psychological intervention studies of behavioral disorders, as well as in other application areas. The methodological objective of this project is to study the characteristics of several statistical techniques designed specifically for longitudinal data.

To further describe the research question, a substantive example of such longitudinal data set is illustrated and analyzed through this project. A series of questions, asked during 1982-1985, 1988-1989, and 1992 by National Longitudinal Surveys of Youth (NLSY) surveys, elicited information on the development of drinking patterns, quantity of various alcoholic beverages consumed, frequency of use, impact of consumption schoolwork and/or job performance, and types of physiological and behavioral dependency symptoms experienced by respondents (OSU, 1994). NLSY is a subset sample from the National Longitudinal Surveys (NLS) which are sponsored by the Bureau of Labor Statistics, U.S. Department of Labor. NLSY contains 12,686 NLS respondents who were born between 1957 and 1964.

The substantive aspects of this project involve the study of alcohol problems among adolescents. This project focuses analyses on the three NLSY cohorts born in 1962, 1963, and 1964. The alcohol consumption related surveys had been started in NLSY from 1983. By the end of 1983, all these three cohorts were still under age 21. None of the 50 states in the United States legally allows people under age 21 to use alcohol. Therefore, analyses on the alcohol misuse and abuse of these age groups can reveal the actual alcohol problems among adolescents.

To reach the research goals, several variables of alcohol usage are chosen by suggestions from the alcohol literature and they are summarized.
as follows.

**Response:** Index of Alcohol Consumption (IAC)

IAC is obtained by the number of times of alcohol drinking in one month in 1983, 1984, and 1985. This variable was recorded for three years only. After 1985, it was no longer updated in NLSY. As a result, this analysis will contain three time points.

**Covariates:**

1. **Race:** The ethnicity of respondents defined in NLSY is Hispanic, African American, and Other.
2. **Gender**
3. **Family history:** The index for living with alcoholic relatives.
4. **Age start drinking:** Respondents’ age when started drinking.
5. **Alcohol usage information in 1982.**
6. **Status of school enrollment.**
7. **Poverty**

As shown in figures 1 and 2, for these young adults the variable of alcohol use is distributed as skew, non-normal, and more likely Poisson distributions. Moreover, table 1

![Graph](image)

**Figure 1.** Latino-American Young Adults (1983, 1984, 1985)

**Figure 2.** African-American Young Adults (1983, 1984, 1985)
also shows that these data have certain degrees of extra-Poisson variation, heteroscedasticity, and within-subject dependence. For example, these variances are far larger than their means—a contrast to regular Poisson distribution which has equal mean and variance. Both means and variances are increased as time increased—an instance of heteroscedasticity. Finally, the correlation matrix reveals these data are correlated cross years.

Table 1. Summary Statistics and Correlation Matrix for African American Alcohol Use

<table>
<thead>
<tr>
<th></th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1983</td>
</tr>
<tr>
<td>means variances</td>
<td></td>
</tr>
<tr>
<td>1983</td>
<td>18.06093</td>
</tr>
<tr>
<td></td>
<td>785.3093</td>
</tr>
<tr>
<td>1984</td>
<td>19.81004</td>
</tr>
<tr>
<td></td>
<td>715.5572</td>
</tr>
<tr>
<td>1985</td>
<td>21.73118</td>
</tr>
<tr>
<td></td>
<td>934.1110</td>
</tr>
</tbody>
</table>

To analyze such longitudinal data set, the statistics methods have to satisfy the criteria that the model shall (i) account for variability between subjects, (ii) provide an explicit representation of the dependence among the counts of each subject, (iii) allow the event rate to vary as a function of time, (iv) be sufficiently tractable and flexible to allow broad application, and (v) incorporate covariate data (Thall, 1988).

1.1 Overview of Organization of this Paper

Section 1 introduces the research question and gives the overview of this study. Section 2 reviews statistical aspects of several methodological literature for analyzing longitudinal count data sets. Section 3 describes more technical details of the selected methods for this project, especially, various analysis methods and procedures for NLSY alcoholic data are
included here. The estimation results by using standard and alternative methods provided in Section 3 are summarized and interpreted in Section 4. Finally, Section 5 concludes results discovered in previous sections and provides indications of further research.

1.2 Description of Models

The logistic model proposed originally for binomial outcomes by Breslow and Clayton (1993) is extended for mixed Poisson model and described as follows,

\[ \log(\mu_{it}) = a_i + b_i \times \text{TIME}_t + b0_{it} \]

where,

\[ \{y_{it}|b0_{it}\} \sim \text{Poisson}[\mu_{it}] \]

\[ a_i = \alpha_0 + \alpha_1 \times X_i + b1[1,i,1] \]

\[ b_i = \beta_0 + b1[1,i,2] \]

\[ b1[1,k] \sim \text{Bivariate Normal}[\mu_k, \sigma_k^2, \rho], \text{where } i = 1, \ldots, N; k = 1,2 \]

\[ b0_{it} \sim \text{Normal}[\mu_0, \sigma_0^2], \text{where } i = 1,\ldots, N; t = 0,1,2 \]

In this NLSY alcohol data analysis, Breslow and Clayton’s model is further extended and equation (1) now becomes

\[ b_i = \beta_0 + \beta_1 \times X_i + b1[1,i,2] \]

For practical explanation of these coefficients, \( \alpha_0 \) and \( \alpha_1 \) are the fixed effects of intercept and slope, respectively for initial status. Similarly, \( \beta_0 \) and \( \beta_1 \) are the intercept and slope for the growth factor in equation (2). Random effects \( b1[1,1] \) and \( b1[1,2] \) are those unobservable effects due to differences in subjects with respect to the growth factor. Finally, \( b0_{it} \) is the unexplained random effect due to overdispersions.

Several possible statistical estimation techniques for modeling the data pattern will be reviewed and compared in the study, for example, generalized linear model (GLM) (Nelder, 1988), general estimating equations (GEE) (Liang & Zeger, 1986), and Bayesian inference techniques using Gibbs sampling (e.g. Spiegelhalter, Thomas, & Best, 1995). Various aspects
of these estimation techniques will be discussed and studied by practical applications.

1.3 Statement of the Problem

For the data set with extra-Poisson variation or mixed-Poisson distribution, the ordinary least square (OLS) regression can not interpret the data adequately. Not only is OLS regression based on normality assumptions, but also OLS regression is not designed to account for heteroscedasticity, and within-subject dependence.

One possible method to model data with Possion distribution is Poisson regression with GLM procedures. Poisson regression is known for modeling "rare events" for relatively infrequent events; its distribution departs markedly from the symmetric normal distribution. If mean rate $\lambda$ is large, the Poisson distribution can be well-approximated by the normal distribution. Much research has been done in this area (e.g. Breslow & Clayton, 1993).

However, parameter estimation by Poisson regression imposes constraints that are not suitable for mixed Poisson distribution. First, it is based on the assumption that events occur independently over time. Secon, the conditional mean and variance of

Poisson regression are equal. Though Poisson regression can accommodate the non-normality characteristic of the dependent variable, the two constrains of Poisson regression prevent itself to fit data with extra Poisson variation.

To fit the extra variation, and adjusted Poisson regression, so called mixed Poisson regression or Poisson regression with extra-variation, is proposed by several researchers (Breslow & Clayton, 1993). The mixed Poisson regression works by adding an error term $\epsilon$ to the regular Poisson regression model. Suppose the response rate is denoted by $\lambda$, and the unexplained randomness is $\epsilon$, then, mixed Possion regression can be described
\[ \ln \lambda_i = X_i \beta + \varepsilon_i \]

where \( X_i \) is the predictor or covariate matrix.

However, a specification of the distribution of \( \varepsilon_i \) is necessary for completeness of the Mixed Poisson model. An example distribution of \( \varepsilon_i \) with a close-form solution of the parameter estimations is provided here.

When \( \varepsilon_i \) is assumed to have Gamma distribution with mean equal to 1 and variance equal to \( \alpha \), it is denoted as \( \text{Gamma}(\alpha,1) \). This assumption leads to

\[
E(Y_i) = \exp(X_i \beta) = \lambda_i
\]

\[
\text{var}(Y_i) = \lambda_i(1 + \alpha \lambda_i)
\]

where \( Y_i \) is the response count distributed as Poisson (\( \lambda_i \)) with extra variation \( \varepsilon_i \) distributed as Gamma (\( \alpha,1 \)). After some mathematical simplifications, the expected mean of this model is still the same \( \lambda_i \) as in regular Poisson model; however, the variance now is always larger than or equal to the mean (\( \lambda_i \)) of the regular Poisson model. The mixed Poisson model is equivalent to a regular Poisson model only when \( \alpha = 0 \).

The including of extra variations \( \varepsilon_i \) is a practical adjustment for more substantive research situations to cover the unexplained randomness in \( \lambda_i \). However, the distribution of \( \varepsilon_i \) may not be identified. Moreover, many other possible distributions of \( \varepsilon_i \) can even result in some parameter estimation procedures without close-form solutions.

Recent developments of statistical techniques have established several procedures to solve the problems. For example, GEE procedure suggested by Liang & Zeger (1986) is an iterative procedure to model longitudinal data for a general class of outcome variables including Gaussian, Poisson, binary and gamma outcomes. Numerous extended or moderated GEE procedures developed to handle a wider class of applications have been seen in literature in recent years. Other techniques are generalized linear mixed models (GLMM's) (Williams, 1992), HLM (Bryk & Raudenbush, 1992), etc. This is one of the main class of methods to deal with mixed Poisson prob-
lems. The more technically detailed reviews of these methods are furnished in literature review section.

The other class of methods is using Bayesian inference approaches. Bayesian inference incorporates prior information to the estimation procedures.; therefore, it is also known as a full information model. Recent developments of computing facilities have made the calculation for complex Bayesian model, for example, longitudinal data, no longer a unreachable task. A Markov chain Monte Carlo (MCMC) approach known as Gibbs sampling is developed to make the complicated numerical integration of Bayesian inference more accessible.

1.4 Questions to Address in this Study

For longitudinal study of NLSY alcohol data sets with extra-Poisson variation,

1. how different are the estimation results from the two statistical methods?

2. what are the substantive interpretations and explanations frome the estimation results of the methods?

2 Reviews of Literature

This section will review the related literature from both practical alcoholism research and statistical technique aspects.

2.1 Research on Alcoholism of Young Adults

2.1.1 Alcohol Risk Population

1. Age Groups

Excessive consumption of alcohol by young people is a challenging
problem for society; therefore, it is also an important research topic for many study areas (Martlatt, Bare, & Larimer, 1995). According to the law, it is illegal for Americans under the age of 21 to drink alcohol. Research on alcohol problems of young adults within these age ranges will constitute a unique area which is very necessary to be studied more closely (Martlatt, Baer, & Larimer, 1995).

2. Gender Differences

Historically, male young adults had more alcohol problems than their female peers. For example, considerable evidence suggests that men residing in college fraternities are at increased risk for alcohol-related problems as compared to their nonfraternity peers (Martlatt, Bear, & Larimer, 1995). Although women residing in sororities have been studied less frequently, anecdotal reports of heavy drinking by sorority women abound (Martlatt, Baer, & Larimer, 1995).

3. Ethnic Groups

African-Americans and Latino Americans are included in this project. Literature suggests that these two ethnic groups have their own specific characteristics of drinking patterns and/or alcoholic behaviors (Peterson, Hawkins, Abbott, & Catalano, 1995).

2.1.2 Risk Factors

1. Social/Family factors

Various research aspects and surveys suggest that influence from family, including genetic factors, parental drinking problems, parents, attitudes affect youth drinking behaviors significantly. Although its observed effects have often been small or indirect, parental alcohol consumption has been associated with adolescent alcohol initiation, current user, and anticipated future use (Peterson, Hawkins, Abbott & Catalano, 1995). Direct effects are explained primarily by social learning theory’s behavioral model-
ing (Bandura, 1977), in which adolescents learn to drink by observing their parents drink. Indirect effects of parental alcohol consumption may operate on attitudes and normative standards or on family management practices (Peterson, Hawkins, Abbott & Catalano, 1995).

2. School factors

Status of school enrollments is also an important factors for youth drinking problems. National Surveys reveal that U.S. college students have a slightly higher annual prevalence of any alcohol use (88%) compared to their age peers who do not attend college (85%); 43% of college students report at least one episode of binge drinking in the last 2 weeks, compared to 34% of their age peers (Johnston, O'Malley, & Bachman, 1992). Other research and literature also indicate that heavy alcohol use is associated with a wide range of adjustment problems for college students, including school failure, relationship difficulties, vandalism, etc (Berkowitz & Perkins, 1986; Engs & Hanson, 1985; Marlatt, Baer, & Larrimer, 1995).

### 2.2 Recent Developed Algorithms for Analysis of Longitudinal Data

The statistical methodologies to analyze extra variation Poisson or mixed Poisson regression problems can be found mainly in literature of biometrics. In fact, several estimation techniques have been proposed during the past decades. For example, when there is a single response for each subject, i.e., all $n_i = 1$, generalized linear models (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1983) are broadly applicable. Nevertheless, more recent developed statistics algorithms with abilities for accommodating extra-variation and correlation problems are reviewed and utilized in this project. Two of the major streams of these new methods are summarized as follows,

#### 2.2.1 Generalized Estimating Equations (GEE)
GEE approach was invented by Liang and Zeger (1986) to model longitudinal data for a general class of outcome variables including Gaussian, Poisson, binary and gamma outcomes. The algorithm uses an iterative procedure to estimate regression coefficients, treating the correlation among observation on the same individual as a nuisance. The basic principle behind the method is a generalization of the face that weighted least squares analyses give unbiased parameter estimates no matter what weights are used. Generalized linear models, such as logistic, Poisson regression, have similar robustness properties, giving asymptotically correct parameter estimates even when the data are correlated. This means that it is possible to estimate regression parameters using any convenient or plausible assumptions about the true correlation between observations and get the right answer even when the assumptions are not correct. It is only necessary to use a "model-robust" or "agnostic" estimate of the standard errors. However, it turns out that there is a moderate gain in efficiency resulting from choosing a working correlation structure close to the true one (Lumley, 1996). Because GEE was derived from generalized linear model, it is naturally to expect that when the observations are all independent, i.e. only one observation left in each group, the GEE method reduces to a regular GLM with the robust estimates of standard errors.

2.2.2 Gibbs Sampling

Bayesian inference on statistical problems using a Markov chain Monte Carlo approach to numerical integration known as Gibbs sampling has drawn many researchers' attention in recent years. Gibbs sampling is a numerical computer intensive statistical tool which is used to analytical derive univariate distributions from complex multivariate distributions. The Gibbs sampler is designed to determine the marginal density.

\[ f(x) = \int f(x, y_1, y_2, ..., y_p) \, dy_1 \ldots dy_p \]
from the multivariate distribution \( f(x, y_1, y_2, \ldots, y_p) \). The natural approach
to take would be to perform the above integration. However, there are
cases where the integration is extremely difficult to perform and this is
when the Gibbs sampler may be a good candidate.

The Gibbs sampler does not calculate \( f(x) \) directly. In fact the Gibbs
sampler generates a sample \( X_1, X_2, \ldots, X_n \) from a distribution whose limit is \( f(x) \) and uses this sample as the approximation to \( f(x) \).

Consider the two variable case and that the conditional distributions \( f\left(X \mid Y\right) \) and \( f(Y \mid X) \) are known. Let \( y_0 \) be an initial starting value for the
Gibbs sequence. The Gibbs sequence is generated as follows.

Step 1. Generate an initial value in the sequence, \( y_0 \).

Step 2. From the conditional distribution \( f(x \mid y_0) \), generate an \( x_0 \) value.

Step 3. From \( f(y \mid x_0) \), generate a \( y_1 \) value.

Step 4. Repeat the process of Step 2-3 until convergence is reached.

It turns out that for large samples and under certain conditions \( x_k \) is
virtually independent of the initial starting value. Moreover, in this case
\( f_k(x) \to f(x) \). This is are generated then this sample should be approxi-
mately equivalent to a sample from the target distribution. However, there
are a number of factors that must be considered before Gibbs sampling can
reliably be performed.

One of the factors should be considered is the cholic of initial values.
The Gibbs sampling techniques requires an initial value to set the
sequence in motion. However, to accurately produce such a value would
require a single univariate distribution. In most cases such univariate distri-
butions can not be found, fortunately the technique is sufficiently robust
that the starting value will not influence the results (for example, given a
large enough sample size).

Perhaps, the most important part of the Gibbs sampling technique is
the fact that the sample sequences converge, i.e. \( f_k(x) \to f(x) \). But what
are the conditions ensure that the sequences converge? The basis is that the
transition from \( x_t \) to \( x_{t+1} \) can be written as a transition matrix. This matrix
means that a limiting stationary distribution exists and moreover the sequence will converge. Various methods for detecting the convergence of Gibbs sampler have been proposed by many researchers, for example, Geweke (1992), Gelman and Rubin (1992), Raftery & Lewis (1992), and Heidelberger & Welch (1983).

Although the sequence may converge it is not guaranteed that convergence will be rapid. Thus, it would be good if a condition could be developed that set a burn-in time for which sample values could be taken after such a point.

Bayesian Inference Using Gibbs Sampling (BUGS) used by this project is a program (Spiegelhalter, Thomas, Best, & Gilks, 1995) that carries out the Gibbs sampling algorithms. BUGS is intended particularly for complex models for which there may be no exact analytic solution, and even standard approximation techniques hierarchical or multi-level random effects; latent variable models, frailty models; measurement errors in responses and covariates; censored data, constrained estimation and missing data (Spiegelhalter, Thomas, Best, & Gilks, 1995).

BUGS provides a declarative language for straightforward specification of statistical models. A compiler then processes the model and data and generates the sampling distributions. Finally, appropriate sampling algorithms are implemented to simulate values of the unknown quantities in the model. A menu-driven set of S-Plus functions (CODA) are supplied with BUGS to calculate convergence diagnostics and graphical and statistical summaries of the simulated samples if desired (Spiegelhalter, Thomas, Best, & Gilks, 1995).

3 Methodology

The detailed analysis procedures utilized by this project are described in this section. The selection of data set, definition of covariates for analyzing
NLSY alcohol data are summarized here. The specific model assumptions for each estimation method are also outlined.

3.1 Analysis Procedures for NLSY Alcohol Data

3.1.1 The Data

1. Selections of Population

The targeted age cohorts in this study are respondents who were born in 1962, 1963, and 1964. These respondents were under the minimum legal drinking age 21 when they were surveyed in 1983. Those non-drinkers who answered "no" to the 1982 survey question "Ever had a drink" are deleted from this study. Only two ethnic groups which are Latino Americans and African Americans are included in this study. There are two major reasons of selection of these two groups. First, NLSY used a vague definition of ethnic groups: Hispanic, black, and non-black, non-Hispanic. Therefore, in the third groups, it included not only Anglo-American but also the other races. To avoid including a less well defined groups, the third groups are deleted. Second, literature (Peterson, Hawkins, Abbott, & Catalano, 1995) also suggests that Latino and African Americans have specific alcoholic characteristics and worth to be studied separately.

2. Responses

The index for alcohol consumption is calculated by the following formula (Muthen, 1995).

Alcohol consumption = (number of days had 1 drink in last month × 1) + (number of days had 2 drinks in last month × 2) + (number of days had 3 drinks in last month × 3) + ... + (number of days had 6 drinks in last month × 6).

This measurement is a count process for frequencies of alcohol use in one month and it had been repeated for three survey years of 1983, 1984, and 1985.
3. covariates

The covariates for mixed Poisson regression model are selected from the suggestions of alcoholism literature to the relationship between risk factor and alcohol abuse. Coding system of these covariates are described as follows,

A) sex: Gender of respondents re-coded as 0 for male and 1 for female
B) race: Race of respondents re-coded as 0 for African Americans and 1 for Hispanic Americans
C) age of onset: Age when first began drinking alcoholic beverages on a regular basis, that is at least once or twice a month.
D) alcoholic history records: alcohol usage in 1982; i.e., frequency of six or more drinks in last month coded as -0:never, 1:once; 2:2 or 3 times; 3:4 or 5 times; 4:6 or 7 times; 5:8 or 9 times; and 6:10 or more times.
E) family history: This variable is created by combining information from two separated variable. Frist, respondents had to have relatives at any time in their lives. Second, the alcoholic relatives also had to live with the respondents during the survey year of 1983. If the conditions are both satisfied, the respondent is coded as 1 for this variable, otherwise it is coded as 0.
F) poverty: family poverty status in the survey year of 1983; it is coded as 1 if in poverty and 0 if not in poverty.
G) enrollment status: school enrollment status as of May 1 in the survey year of 1983. This variable is re-coded as 0:not enrolled in any school, and as 1:enrolled in either high school or in college.

3.1.2 Analysis Tools

1. Ordinary Least Square

This estimation procedure is carried out by using ordinary least square (OLS) methods. The model to be estimated by OLS is
\[ E(y_{it}) = \alpha_0 + \alpha_1 TIME_i + \alpha_2 X \]

where \( y_{it} \) is the response.

2. Generalized Linear Model

The approach assumes that the responses are distributed as normal distribution. Therefore, generalized linear model (GLM) with identity link using weight least square (WLS) approaches are performed in this estimation procedure. The estimated model is

\[ E(y_{it}) = a_i + b_i \cdot TIME_i + b_0_{it} \]

where,

\[ \{ y_{it} | b_{0_{it}} \} \sim Normal(\mu_{it}) \]
\[ a_i = \text{alpha0} + \text{alpha.x} \cdot X_i + b_1[i,1] \]
\[ b_i = \text{beta0} + \text{beta.x} \cdot X_i + b_2[i,2] \]
\[ b1[i,k] \sim \text{Bivariate Normal}(\mu_k, \sigma_k^2, \rho), \text{ where } i = 1, ..., N; k = 1, 2 \]
\[ b0_{jk} \sim \text{Normal}(\mu_0, \sigma_0^2), \text{ where } j = 1, ..., N; k = 0, 1, 2 \]

where \( E(y_{it}) \) is the expected mean value for \( y_{it} \) and \( y_{it} \) is distributed as normal.

3. General Estimating Equations

The method is a deviation from generalized linear model by accounting the specific characteristics of longitudinal data set. Therefore, GEE procedure can estimate the model accurately, even the model with potentially correlated problems. The model to be estimated is similar to GLM's but with the following extra assumptions,

\[ \log(\mu_{it}) = a_i + b_i \cdot TIME_i + b0_{it} \]

where,

\[ \{ y_{it} | b0_{it} \} \sim \text{Poisson}(\mu_{it}) \]
\[ a_i = \alpha_0 + \alpha_x X_i + b_i \cdot [i, 1] \]

\[ b_i = \beta_0 + \beta_x X_i + b_i \cdot [i, 2] \]

\[ b_1[i, k] \sim \text{Bivariate Normal} [\mu_k, \sigma_k^2, \rho], \text{ where } i = 1, \ldots, N; k = 1, 2 \]

\[ b_{0,j} \sim \text{Normal} [\mu_0, \sigma_0^2], \text{ where } j = 1, \ldots, N; t = 0, 1, 2 \]

4. Gibbs Sampling

Gibbs sampler is an estimation technique for complicated Bayesian inference. Therefore, the estimation procedure has all the characteristics of Bayesian statistics. In a short summary, this method can handle relatively complicated models. Moreover, it can be a superior method when the sample size is fairly small. The model specified in this project for Gibbs Sampler is identical to GEE's but with the following "non-informative" priors.

1. \( \alpha_0 \sim \text{Normal}(0.0, 1.0E^{-4}); \alpha_x \sim \text{Normal}(0.0, 1.0E^{-4}); \) *1.0E-4 is precision

2. \( \beta_0 \sim \text{Normal}(0.0, 1.0E^{-4}); \beta_x \sim \text{Normal}(0.0, 1.0E^{-4}); \) *1.0E-4 is precision.

3. \( b_{0,t} \sim \text{Normal}(0.0, 1.0E^{-3}); \text{ where } \tau b \sim \text{Gamma}(1.0E^{-3}, 1.0E^{-3}) \) is the prior on precision

4. \( b_1[i, k] \sim \text{Bivariate Normal}(\mu b_{ik}, \Sigma_{kk}); \text{ where } \mu b_{ik} \sim \text{Normal}(0.1.0E^{-4}) \) and \( \Sigma_{kk} \sim \text{Wishart}(R, k, k); \) *Wishart distribution prior on precision matrix

An approximation to Gewek's (1992) diagnosis method included in BUGS 0.51 for detecting convergence sequences is used to see if the analyses of NLSY data by Gibbs sampler were roughly converged.

A summary for abilities and assumptions of each method used in the project is made in Table 2

309
Table 2. Comparison of Model Assumptions

<table>
<thead>
<tr>
<th>methods</th>
<th>normality</th>
<th>Autocorrelation</th>
<th>Mix − Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLM/ identity link</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>GEE</td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Gibbs Sampler</td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

4 Analysis

In this section, descriptive statistics of the NLSY alcohol data sets are presented here. Results after performing the procedures described in Section three are also summarized in this section.

4.1 Summary Statistics

The following summary statistics table is generated to reveal the brief characteristics of NLSY alcohol data. As expected, the descriptive statistics show that this counted data set is bearing with non-normal curved, extra-varied and correlated in survey years.

Table 3. Summary Statistics

<table>
<thead>
<tr>
<th>African &amp; Hispanic Americans</th>
<th>correlation (P-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1983</td>
</tr>
<tr>
<td></td>
<td>mean</td>
</tr>
<tr>
<td></td>
<td>(variance)</td>
</tr>
<tr>
<td>1983 (N=537)</td>
<td>19.5345 (725.7008)</td>
</tr>
<tr>
<td>1987 (N=537)</td>
<td>20.9143 (763.0225)</td>
</tr>
<tr>
<td>1985 (N=537)</td>
<td>21.8603 (951.7958)</td>
</tr>
</tbody>
</table>

4.2 Normal Multiple Linear Regression (Ordinary Least Square) & GLM (Weighted Least Square) using identity Link
Normal multiple linear regression using the following model with regular OLS assumptions is calculated for comparison purpose.

\[ E\{y_i\} = \alpha_0 + \alpha_1 \text{TIME}_i + \alpha_2 X \]

Generalized linear model using weighted least square algorithms with identity link, i.e. assumption of response is normal distributed, is also applied to the NLSY alcohol data. The model specified for GLM is identical to the model for GEE. Those coefficients marked as * are non-significant at 95% confidence levels and with P values larger than 0.05.

| Table 4. Result Summary for African & Hispanic Americans Alcoholic Youth |
|-----------------------------|-------------|-------------|
| factor | coefficient | OLS (std. err.) | GLM (std. err.) |
| initial status | intercept | 19.606815 (0.880173) | 19.606769 (1.006434) |
| | race | -0.015194 (1.803210) | -1.330140 (2.063121) |
| | sex | -10.051792 (1.922123) | -8.424429 (2.199146) |
| | age of onset | -1.888210 (0.526554) | -2.644643 (0.602450) |
| | family history | -0.046134 (1.830506) | -1.150569 (2.094376) |
| | baseline/1982 | 4.858868 (0.571282) | 4.635645 (0.653613) |
| | enrollment | -3.136395 (1.996161) | -3.204478 (2.283800) |
| | poverty | -2.120493 (1.952482) | -1.827441 (2.233855) |
| | random effect1(var) | 415.18289 | 526.62385 |
| growth factor | intercept | 1.162942 (0.030514) | 1.462940 (0.646772) |
| | race | N/A | 1.314954 (1.325839) |
| | sex | N/A | -1.627248 (1.413251) |
| | age of onset | N/A | 0.456138 (0.387157) |
| | family history | N/A | 1.104435 (1.345924) |
| | baseline/1982 | N/A | 0.221357 (0.420036) |
| | enrollment | N/A | 0.069225 (1.467651) |
| | poverty | N/A | -0.293005 (1.435557) |
| | random effect2(var) | N/A | 214.24967 |

correlation: random 1 & random 2 N/A -0.495
Note:* non-significant at 95% level.

### 4.3 Analyses of NLSY Alcohol Data by GEE and Gibbs Sampling

The following table summarizes the estimation results obtained from GEE and Gibbs sampler.

Table 5. Result Summary for African & Hispanic Americans Alcoholic Youth

<table>
<thead>
<tr>
<th>factor</th>
<th>GEE (std. err)</th>
<th>Gibbs Sampler (std. dev)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>model 1</td>
<td>model 2</td>
</tr>
<tr>
<td>intercept</td>
<td>2.858</td>
<td>2.8289</td>
</tr>
<tr>
<td></td>
<td>(0.0457)</td>
<td>(0.0477)</td>
</tr>
<tr>
<td>race</td>
<td>-0.149*</td>
<td>-0.1220*</td>
</tr>
<tr>
<td></td>
<td>(0.1047)</td>
<td>(0.0923)</td>
</tr>
<tr>
<td>sex</td>
<td>-0.5765</td>
<td>-0.5884</td>
</tr>
<tr>
<td></td>
<td>(0.1047)</td>
<td>(0.1036)</td>
</tr>
<tr>
<td>age of onset</td>
<td>-0.1160</td>
<td>-0.1150</td>
</tr>
<tr>
<td></td>
<td>(0.0245)</td>
<td>(0.0265)</td>
</tr>
<tr>
<td>family history</td>
<td>0.0426*</td>
<td>-0.0195*</td>
</tr>
<tr>
<td></td>
<td>(0.0996)</td>
<td>(0.0977)</td>
</tr>
<tr>
<td>baseline/1982</td>
<td>0.1916</td>
<td>0.1942</td>
</tr>
<tr>
<td></td>
<td>(0.0248)</td>
<td>(0.0267)</td>
</tr>
<tr>
<td>enrollment</td>
<td>-0.1373*</td>
<td>-0.1444*</td>
</tr>
<tr>
<td></td>
<td>(0.1267)</td>
<td>(0.1159)</td>
</tr>
<tr>
<td>poverty</td>
<td>-0.0374*</td>
<td>-0.0487*</td>
</tr>
<tr>
<td></td>
<td>(0.1287)</td>
<td>(0.1118)</td>
</tr>
<tr>
<td>random effect1:SD</td>
<td>1.2943</td>
<td>0.8094</td>
</tr>
<tr>
<td></td>
<td>(0.0592)</td>
<td>(0.1102)</td>
</tr>
<tr>
<td>growth factor</td>
<td>0.0517*</td>
<td>0.0588*</td>
</tr>
<tr>
<td></td>
<td>(0.0259)</td>
<td>(0.0289)</td>
</tr>
<tr>
<td>race</td>
<td>0.1020*</td>
<td>0.0890*</td>
</tr>
<tr>
<td></td>
<td>(0.0518)</td>
<td>(0.0566)</td>
</tr>
<tr>
<td>sex</td>
<td>-0.1186*</td>
<td>-0.0822*</td>
</tr>
<tr>
<td></td>
<td>(0.0628)</td>
<td>(0.0680)</td>
</tr>
<tr>
<td>age of onset</td>
<td>0.0324</td>
<td>0.0275</td>
</tr>
<tr>
<td></td>
<td>(0.0136)</td>
<td>(0.0141)</td>
</tr>
<tr>
<td>family history</td>
<td>0.0068*</td>
<td>0.0465*</td>
</tr>
<tr>
<td></td>
<td>(0.0558)</td>
<td>(0.0586)</td>
</tr>
<tr>
<td>baseline/1982</td>
<td>0.0135*</td>
<td>0.0032*</td>
</tr>
<tr>
<td></td>
<td>(0.0152)</td>
<td>(0.0156)</td>
</tr>
<tr>
<td>enrollment</td>
<td>0.0203*</td>
<td>0.0285*</td>
</tr>
<tr>
<td></td>
<td>(0.0677)</td>
<td>(0.0665)</td>
</tr>
<tr>
<td>poverty</td>
<td>0.0012*</td>
<td>-0.0004*</td>
</tr>
<tr>
<td></td>
<td>(0.0670)</td>
<td>(0.0635)</td>
</tr>
<tr>
<td>random effect2:SD</td>
<td>0.6717</td>
<td>0.3811</td>
</tr>
<tr>
<td></td>
<td>(0.0282)</td>
<td>(0.0429)</td>
</tr>
<tr>
<td>correlation:random 1 &amp; random 2</td>
<td>-0.4209</td>
<td>-0.337</td>
</tr>
<tr>
<td></td>
<td>(0.0494)</td>
<td>(0.1316)</td>
</tr>
<tr>
<td>overdispersion:SD</td>
<td>N/A</td>
<td>2.9276</td>
</tr>
<tr>
<td></td>
<td>(0.0306)</td>
<td></td>
</tr>
</tbody>
</table>

Note:1. Model 1: without overdispersion; Model 2: with overdispersion.
2. *: non-significant at 95% level.

312
4.4 Computer Resources

All the computation results presented here were obtained from an IBM compatible PC with Intel Pentium-100 MHZ CPU and 16 MB RAM. The following sets of software are used for the specific estimation procedures. For OLS, GLM and GEE procedure, the Hierarchical Linear Model (HLM) for Microsoft Windows 95 version 4.0 by Bryk, Raudenbush, and Congdon (19960 is used. For Gibbs Sampling estimation, BUGS 0.51 with CODA 0.30 (Best, cowles, & Vines, 1995) which is a convergence diagnosos and output analysis software for Gibbs sampling output is used.

5 Conclusions

Conclusions are divided into two section, one is for the substantive interpretation and explanation of NLSY alcoholic data; the other is for the comparisons of statistics methodology utilized in this project.

5.1 Substantive Alcoholism Aspects

The results in summary Table 5 are re-calculated without the nature logarithm transformations, a table with interpretable results is in Table 4. Histograms of these coefficients are plotted in Appendix for better understanding.
Table 6. Re-calculated Coefficients.

<table>
<thead>
<tr>
<th></th>
<th>GEE</th>
<th>Gibbs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>model 1</td>
<td>model 2</td>
</tr>
<tr>
<td>initial factor</td>
<td>intercept</td>
<td>1742.94%</td>
</tr>
<tr>
<td></td>
<td>alpha.race</td>
<td>*86.16%</td>
</tr>
<tr>
<td></td>
<td>alpha.sex</td>
<td>56.19%</td>
</tr>
<tr>
<td></td>
<td>alpha.age</td>
<td>89.05%</td>
</tr>
<tr>
<td></td>
<td>alpha.hist</td>
<td>*104.35%</td>
</tr>
<tr>
<td></td>
<td>alpha.base</td>
<td>121.12%</td>
</tr>
<tr>
<td></td>
<td>alphas.enrol</td>
<td>*87.17%</td>
</tr>
<tr>
<td></td>
<td>alphas.pov</td>
<td>*96.33%</td>
</tr>
</tbody>
</table>

Growth factor

<table>
<thead>
<tr>
<th></th>
<th>GEE</th>
<th>Gibbs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>model 1</td>
<td>model 2</td>
</tr>
<tr>
<td>initial factor</td>
<td>intercept</td>
<td>*105.30%</td>
</tr>
<tr>
<td></td>
<td>beta.race</td>
<td>*110.74%</td>
</tr>
<tr>
<td></td>
<td>alpha.sex</td>
<td>*88.82%</td>
</tr>
<tr>
<td></td>
<td>alpha.age</td>
<td>103.29%</td>
</tr>
<tr>
<td></td>
<td>alpha.hist</td>
<td>*100.68%</td>
</tr>
<tr>
<td></td>
<td>alpha.base</td>
<td>*101.35%</td>
</tr>
<tr>
<td></td>
<td>alphas.enrol</td>
<td>*102.05%</td>
</tr>
<tr>
<td></td>
<td>alphas.pov</td>
<td>*100.12%</td>
</tr>
</tbody>
</table>

Note: 1. Model 1: without overdispersions; Model 2: with overdispersions.
2. *: non-significant at 95% level.

5.1.1 Coefficient Interpretations

The practical interpretations of coefficient now become intuitively after the re-calculation. For example, the estimated coefficient (alpha.sex) of GEE model 1 for variable sex is 59.16%. Because female is coded as 1 and male is 0, female drinks only 56.19% of the frequencies that male drinks in the same month. In another work, if a male drinks 100 times of alcohol, female will drink only approximate 56 times in the same month.
Moreover, this coefficient associates to the initial status. This is saying that a female drinks 56.19% of the frequencies that male drinks in their initial drinking.

Similarly, the estimate of beta.pov (coefficient of poverty) of Gibbs model 1 is 114.33%, this shows that those young adults who lived in poverty consumed more alcohol by 114.33% of frequencies than those who did not live in poverty. In another word, if the respondents who did not live in poverty had 100 times of drinks, those who lived in poverty consumed approximate 114 times of alcohol in the same month. Moreover, this coefficient (beta.pov) related to their growth factor; therefore, those respondents who lived in poverty would consume 114.33% more alcohol in each year. For example, if a respondent who lived in poverty had approximate 114 times of drinks in one month of the first year, he/she would have 228 times of drinks in one month of the second year. 342 times for the third year, and so on.

5.2 Statistics Methodological Aspects

As we can see from the coefficient tables, OLS and GLM generated much different results from GEE and Gibbs sampler, Kernel density plots (see Appendix) of OLS and GLM fitted values also show that neither of the two methods fit the NLSY data well.

In general, GEE and the Gibbs sampler give fairly similar estimates in this analysis. For model 1 both methods give almost identical results on estimation of coefficients and standard errors. However, there are small differences for model 2 on some parameters, for example, school enrollment and poverty level have more significant effects on initial status estimated by Gibbs sampler than by GEE. This might be caused by the different algorithms utilized by the two methods for calculating overdispersions. Therefore, the estimations of overdispersion are also very different by GEE and Gibbs sampler. A complete simulation study on this issue may be
engaged to reveal how and why is difference.

In terms of computing efficiency, GEE is more favorable than Gibbs sampler because it uses a simpler algorithm and computation procedure which makes itself faster and clearer. As a result, GEE needs far less computing time than Gibbs sampler using BUGS will need for the same model. For example, in estimating the alcohol models, GEE only requires few iterations to reach its convergence and take no more than 10 minutes to finish each computation.

Gibbs sampling by BUGS is less favored because of its relative much longer computing time. Moreover, there is still no robust convergence diagnostic methods able to perform consistently and reliably for the results by Gibbs sampling estimation. Nevertheless, the capabilities of incorporating of prior information and estimating accurately of smaller sample sizes of Gibbs sampling have made itself a very attractive statistical software and research tool.

5.3 NEED for Further Research

Both methodological and substantive aspects of research are of interest for further studies. For example, more detailed Monte Carlo studies on characteristics of Giggs sampling and GEE can be an attractive and interesting topic for methodologists. For example, effects of different pattern of correlation in time (highly correlated vs. non-highly correlated), effects of different sample sizes, and effects on standard errors because the two methods can be an unique research topic.

Another interesting research topic could be the study of mixed and mixture models. Currently, methods of quasi-likelihood assume a mean variance relationship for the response variable and include an additional dispersion parameter to account for overdispersion (Breslow, 1984; McCullagh & Nelder, 1989) In contrast, mixture model approaches assume that the Poisson mean is random. The distribution of this mean is referred to as a
mixing distribution. e.g. assuming the distribution is a discrete random variable yields a finite mixture model, as in Puterman, et al., 1996; and Wang, 1994.

Meanwhile, more studies on potential risk factors of young adults alcoholism can also be the interesting area for alcohol studies. For example, developing complex models for real data can substantively improve the explanation of alcohol abuse behaviros. Research on mixed or mixture Poisson were mainly done by multiple regression type of model (Breslow, 1984; Clayton, 1993; Wng, 1996) Including latent variables and mean stuctures into the model will be a challenging task, e.g. the models of Muthen’s general growth mixture modeling (GGMM) methods for normal outcomes can be extended to non-normal outcomes, say, mixture Poisson outcomes.

The including of other estimation algorithms could be one of the research areas need to be further studies. For example, Mixture Poission by EM algorithms (Wang. et a,1996; Wang,1994), or quasi-likelihood by marginal quasi-likelihood(MQL) as in Hedeke and Gibbons’(1994) Mixor and penalized quasi-likelihood(PQL) as in David Clayton’s(1993,1994) generalized linear mixed model (GLMM) may be excellent candidates.
References


Lumley, T (1996), GEE-Lispstat objects for generalized estimating
equation models, unpublished manual, Department of Biostatistics, University of Washington.


Appendix I

Kernel density plots for fitted values

Appendix II

Chart 1:GEE and Gibbs sampler estimations of alphas of model 1
Chart 2:GEE and Gibbs sampler estimations of alphas of model 2
Chart 3:GEE and Gibbs sampler estimations of betas of model 1
Chart 4:GEE and Gibbs sampler estimations of betas of model 2
Chart 5:GEE and Gibbs sampler estimations of all coefficients of model 1 without plotting of interceps.

Chart 6:GEE and Gibbs sampler estimations of all coefficients of model 2 without plotting of interceps.
Abstract

This project conducted using generalized estimating equations (Liang & Zeger, 1986) and Bayesian inference by Gibbs sampling approaches to investigate the relationship between potential risk factors and minority youth alcohol abuse problems. The National Longitudinal Study of Youth (NLSY) alcohol data set has specific extra-Poisson variation, heteroscedasticity, and within-subject dependence characteristics that make it difficult to be analyzed by regular statistical procedures. Generalized linear models (GLM), GEE, and Bayesian inference by Gibbs sampling with various model assumptions, (e.g. Poisson regression with extra variation) provide important alternatives to ordinary least square regression. Results show that estimation by the techniques that accommodate autocorrelation problems often seen in longitudinal studies perform better than those without such characteristics. This project utilizes GLM, GEE and Gibbs sampling techniques to reveal the specific patterns of longitudinal data with mixed Poisson distributions.

Analyses in this study also verify earlier findings in the alcohol literature that suggests that differences between gender, race, age of onset, family history, status of school enrollment, and poverty play significant roles in the disorder behavior of alcohol abuse among young adults.

Key words: Mixed Poisson Regression, Overdispersion, Gibbs Sampling, GEE, Longitudinal data, Alcoholism
Histogram and Kernel Density
Smooth Line for Raw Data

Raw Data Kernel Density Plot

Multiple Linear Regression (OLS)

GLM with Identity Link (WLS)

Generalized Estimating Equations

Bayesian Inference (Gibbs Sampling)