Journal of Data Science **7**(2009), 469-485

Statistical Methods for the Analysis of Alcohol and Drug Uses for Young Adults

Liang Zhu, Jianguo Sun and Phillip Wood
University of Missouri

Abstract: Alcohol and drug uses are common in today's society and it is well-known that they can lead to serious consequences. Studies have been conducted in order, for example, to understand short- or long-term temporal processes of alcohol and drug uses. This paper discusses statistical modeling for joint analysis of alcohol and drug uses and several models and the corresponding estimation approaches are presented. The methods are applied to a prospective study of alcohol and drug uses on college freshmen, which motivated this investigation. The analysis results suggest that female subjects seem to have much less consequences of alcohol and drug uses than male subjects and the consequences of alcohol and drug uses decrease along with ages.

Key words: Marginal mean model, missing covariates, multivariate longitudinal outcomes, random effects model.

1. Introduction

Alcohol and drug uses are common and costly in the United States and other countries and the young adult and college years have been identified as a particular period of vulnerability for such uses (Jackson *et al.*, 2002; Sher *et al.*, 1991). For example, it is well-known that there exist many deaths, injuries and sexual assault cases each year due to college student drinking. Also the number of deaths directly attributable to drug use has more than doubled over the last two decades and is close to the number of deaths directly resulting from alcohol use. In addition to the deaths directly resulting from alcohol and drug uses, of course, there are many deaths indirectly related to their uses such as motor vehicle accidents and suicide. Among the alcohol and drug users, some suffer from alcohol or drug use disorder, abuse or dependence, but many simply misuse alcohol or drug.

A number of studies have been conducted to investigate various aspects of alcohol and drug uses. One example, which motivated this paper and will be discussed in more detail below, is a follow-up study of freshmen at the University of Missouri with the goal of examining the development and persistence of alcohol and drug uses in a high-risk sample. The focus of the study was about alcohol, health and behavior and it is often referred to as AHB study (Sher *et al.*, 1991; Jackson, Sher and Schulenberg, 2005). In this study, during the 11 years follow-up, each study subject was supposed to be examined six times and at each examination, the questions about their alcohol and drug uses such as the use frequencies were asked. For the alcohol use, for example, the focus was on heavy drinking and the questions asked include "In the past 30 days, how many times have you had five or more drinks at a single sitting, either of beer, wine, wine coolers, liquor, or some combination of these?". The collected data summarize these information and include the numbers of heavy drinking and drug use occasions per week along with many characteristics of the study subjects.

Before discussing the analysis of the data arising from the AHB study, we will first consider the analysis of general multivariate longitudinal data and present some novel statistical models and the corresponding inference procedures. These approaches will then be applied to the data from the AHB study with the focus on the assessment of temporal processes of alcohol and drug uses and the effects of covariates such as gender and personality on the processes. Note that for alcohol and drug use studies, response variables are often given in the form of counts such as the number of alcohol or drug uses and the number of negative consequences due to alcohol or drug use within a week or month. Also note that many methods have been developed specifically for the analysis of alcohol and drug use-related studies based on the Poisson distribution or process assumption for the count response variables representing alcohol and drug uses. In this paper, we present some alternatives to these Poisson-based methods. In addition to avoid the Poisson assumption, the proposed approaches also allow time-varying covariates, missing responses and covariates, and correlated response variables, which present in the data from the AHB study and the longitudinal data from many other studies or fields.

A number of statistical methods have been developed for the analysis of univariate longitudinal data and these include generalized estimating equation-based approaches and random effects model-based approaches (Diggle *et al.*, 1994; Laird and Ware, 1982; Liang and Zeger, 1986; Zeger and Liang, 1986). In comparison, only limited literature exists that deals with multivariate longitudinal data. In addition to taking into account the correlation among repeated measurements on the same subject, the analysis of multivariate longitudinal data needs to pay attention to the correlation among different response variables. Among others,

Shah, Laird and Schoenfeld (1997) extended linear mixed effects models of Laird and Ware (1982) to the cases where there exist multiple longitudinal response variables.

The remainder of this paper is organized as follows. In Section 2, we describe the AHB study and the observed data in more detail and Section 3 introduces notation and three models that will be used for the analysis. The first two are latent variable models that specifically model the relationship among different types of longitudinal response processes, while the third model is a marginal mean model that leaves the relationship arbitrary. Inference procedures for these models are discussed in Section 4. Specifically, for the two latent variable models, as most authors, we apply the maximum likelihood approach and rely estimation of parameters on EM algorithms. For the marginal mean model, the estimating equation approach is used for estimation of regression parameters and the proposed estimates are consistent and have asymptotically normal distributions. Section 5 applies the presented statistical approaches to the AHB study and some concluding remarks are given in Section 6.

2. Alcohol, Health and Behavior Study

The AHB study is an 11-year follow-up study consisting of 489 subjects recruited from over 3000 incoming, first-time freshmen at the University of Missouri-Columbia in 1987. They were selected based on responses to a test that measures alocoholism in their biological parents and some other criteria. The detailed discussion on the test and criteria is given in Sher *et al.* (1991). Among these 489 subjects, approximately half of them are male and also approximately half of them are considered family history positive for paternal alcoholism on the basis of endorsement of self-report questionnaire and interview items concerning paternal alcoholism. During the follow-up, participants completed structured interview measures of alcohol and drug diagnoses and questionnaire measures of problems associated with these substances at each of 6 assessments at years 1, 2, 3, 4, 7, and 11.

During the AHB study, the investigators collected information on a number of variables related to alcohol and drug uses and about characteristics of study subjects (Sher *et al.*, 1991). In particular, the response variables include ACON and DCON, representing the numbers of negative events related to alcohol and drug uses or often called negative alcohol and drug consequences, respectively. They are defined as the numbers of positive answers to 14 items in the Short Michigan Alcoholism Screening Test (Selzer, Vinokur and Van Rooijen, 1975). These items include hangovers, blackouts and driving while intoxicated and others specifically designed for college students such as missing classes or receiving a low grade as the result of alcohol and drug uses. The top panel of Figure 1 presents

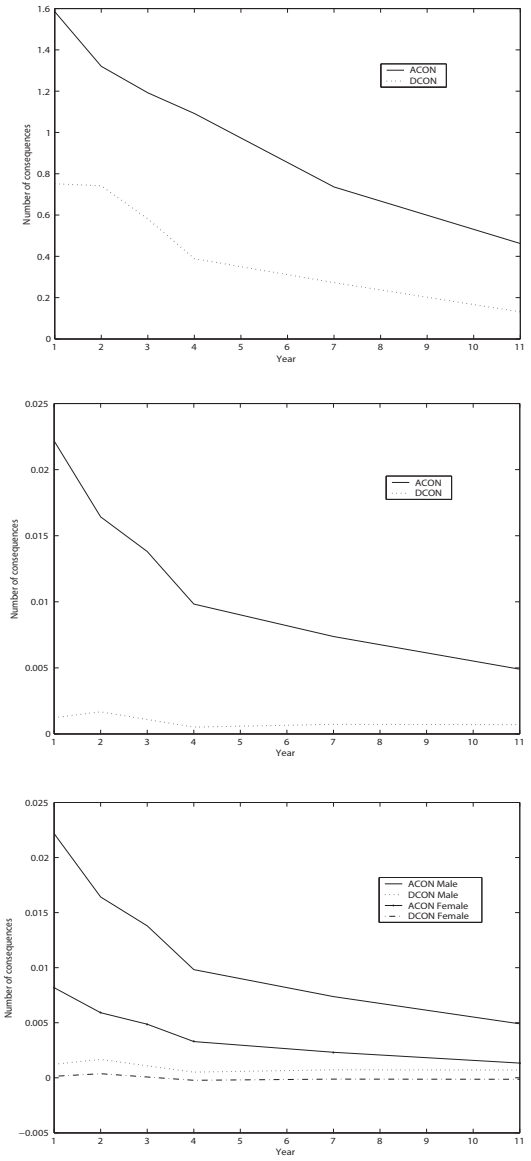

Figure 1: Top panel: Sample means of the consequences of alcohol and drug uses. Middle panel: Estimated averages of the consequences of alcohol and drug uses. Bottom panel: Estimated averages of the consequences of alcohol and drug uses

the sample means of these two variables over the study period and indicates that both alcohol and drug consequences decrease along with ages. In the analysis below, we will focus on these two variables. Some other response variables that were measured and are directly related to alcohol and drug uses include the

frequencies of drinking occasions and heavy drinking occasions as well as the quantity consumed in an average day.

Among the covariate variables that were measured in the AHB study, in addition to the gender variable SEX, one is the variable LESECY, which is commonly of interest in psychology and represents the total number of life events, either positive or negative, happened in the previous year (Sarason, Johnson and Siegel, 1978). Also the information is available for three variables that concern personal characteristics of the study subject and they are EEPQ, NEPQ and NEOC. The variables EEPQ and NEPQ are defined by using the Eysenck Personality Questionnaire (Eysenck, 1988), which consists of 90 items designed to assess the personality traits, and based on the subtests of extroversion and neuroticism, respectively. The variable NEOC measures conscientiousness and is defined by using the NEO personality inventory questionnaire (Costa and McCrae, 1992). Table 1 provides a summary of these five variables.

Table 1: Summary of covariate variables

Variable	Type	Range	Mean	Median
SEX	binary	(0,1)	0.5256	1
LESECY	count	(0,47)	11.2100	11
EEPQ	count	(1,21)	14.9161	16
NEPQ	count	(0,23)	8.2037	7
NEOC	count	(11,48)	32.6725	33

As in most longitudinal studies, there exist many missing values in the AHB study. For example, at the last assessment by year 11, 79 of 489 subjects had left the study and no measurements were available from these subjects. Also EEPQ and NEPQ were measured only at the 1st, 5th and 6th assessment years, while NEOC was measured only at the 4th, 5th and 6th assessment years. For this study, the questions of interest include how alcohol and drug consequences change over time, if and how much the consequences are related, and what characteristic variables are predictive of alcohol and drug consequences.

Many authors have analyzed the AHB study from different points of view with the focus on alcohol and drug use frequencies. For example, Sher *et al.* (1991) first described the study in detail and discussed several basic problems. One of these is about sampling bias and they found that the study subjects and those who chose not participating in the study are generally similar. Jackson, Sher and Wood (2000) studied the comorbidity between alcohol use disorders and tobacco use disorders and Jackson and Sher (2003) examined the association between alcohol use disorders and psychological distress. Trull, Waudby and Sher (2004) discussed the relationship between the personality and alcohol and drug uses. Most of the analysis methods discussed in these papers apply only to complete

data or require Poisson process-related or parametric model assumptions. In the following, the focus will be on joint analysis and the methods presented not only allow missing data but also do not need the Poisson process assumption.

3. Notation and Models

Consider a longitudinal study that involves n independent subjects and in which there exist K response variables or processes of interest and each subject is measured at M different time points. For subject i, let Y_{ikm} denote the measurement on the kth response variable at the mth time point and suppose that there exists a p-dimensional vector of covariates X_{im} that may be time-dependent, $i = 1, ..., n$, $k = 1, ..., K$, $m = 1, ..., M$. Also for subject i, suppose that there exists a latent variable U_{im} representing all characteristics or factors of the subject at the mth time point that affect the response variables. We assume that given U_{im}, Y_{ikm} can be described by the following linear mixed model

$$Y_{ikm} = \beta_{0k} + \beta_{1k} U_{im} + b_{ik} + e_{ikm}, \tag{3.1}$$

where the β_{0k}'s and β_{1k}'s are regression parameters, the b_{ik}'s are random intercepts and the e_{ikm}'s are measurement errors. Note that in model (3.1), the Y_{ikm}'s for different k share the same U_{im}. For the analysis of the AHB study, Y_{ikm} will be taken to be the logarithm of ACON and DCON in model (3.1) as well as model (3.3) described below. Some comments on this will be given later. It will be assumed that the b_{ik}'s and e_{ikm}'s are independent of each other and distributed as the normal distributions with mean zero and variances σ_{bk}^2 and σ_{ek}^2, respectively.

It the U_{im}'s were known, one could simply determine the relationship between the U_{im}'s and the X_{im}'s and then easily regress the Y_{ikm}'s on the U_{im}'s to estimate covariate effects. Since the U_{im}'s are unobserved, we need to specify the relationship between the U_{im}'s and the X_{im}'s and for this, we assume that

$$U_{im} = X_{im}'\alpha + a_i + \epsilon_{im}, \tag{3.2}$$

where α is a p-dimensional vector of regression parameters, the a_i's are random effects and the ϵ_{im}'s denote measurement errors. Again we will assume that the a_i's and the ϵ_{im}'s are independent of each other and distributed as the normal distributions with mean zero and variances σ_a^2 and σ_ϵ^2, respectively. They will be assumed to be independent of the b_{ik}'s and e_{ikm}'s.

The model defined by (3.1 and (3.2) means that all K response variable Y_{ikm}'s are linearly determined by the latent variables U_{im}'s with some errors and the U_{im}'s follow a linear mixed model determined by covariates X_{im}. However, different response variables may have different relationships with the latent variables.

Both b_{ik} and a_i characterize the correlation among repeated measurements on the same response variable. As usual, it will be assumed that different response variables are independent given the latent variables.

Under model (3.1), each response variable is determined by two parameters, β_{0k} and β_{1k} given the latent variables U_{im}'s. Sometimes this may be too restrictive. For this, we can consider the model

$$Y_{ikm} = \beta_{0km} + \beta_{1k} U_{im} + b_{ik} + e_{ikm}, \tag{3.3}$$

where β_{1k}, b_{ik} and e_{ikm} are defined as in model (3.1) and $\beta_{0k1}, ..., \beta_{0kM}$ are intercept parameters for the kth response variable over measurement times. It is apparent that if $\beta_{0k1} = ... = \beta_{0kM}$, model (3.3) reduces to model (3.1). That is, model (3.1) is a simplified version of model (3.3).

In terms of the alcohol and drug use study, the advantages of models (3.1) and (3.3) include that they clearly define the mechanisms behind the alcohol and drug uses and allow one to study all aspects together. As with most latent variable models, their main disadvantage is that it may be difficult to verify the assumptions about the linear relationship and the normality. Corresponding to these, we note that the response variables about the alcohol and drug uses are count variables and thus instead of treating them as general longitudinal variables, they can be represented using the counting process notation as defined below.

Consider a longitudinal study on some recurrent events and let N_{ikm} denote the cumulative number of the kth type event that have occurred up to the mth measurement time point for subject i, $i = 1, ..., n$, $k = 1, ..., K$, $m = 1, ..., M$. For the AHB study, we can define N_{ikm} as the total number of the consequences of alcohol or drug use during the first m measurements periods. For N_{ikm}, a natural marginal mean model is given by

$$E(N_{ikm} \,|\, X_{ikm}) = \exp(X'_{ikm}\tau + \gamma_{km}) \tag{3.4}$$

(Cai and Schaubel, 2004; Lawless and Nadeau, 1995; Lin and Ying, 2001), where as α, τ is a p-dimensional vector of regression parameters and the γ_{km}'s are unknown baseline parameters.

A major advantage of model (3.4) is that it is a marginal model and does not require the normality assumption. Also it makes use of the count nature of the response variables considered here and leaves the correlation among response variables arbitrary. Of course, this model cannot be used for prediction, but it seems to be more natural and appropriate if one is only interested in covariate effects. In the next section, we discuss estimation procedures for the models described above.

4. Estimation Procedures

Let $Y_i = (Y_{i11}, ..., Y_{i1M}, ..., Y_{iK1}, ..., Y_{iKM})'$, $U_i = (U_{i1}, ..., U_{iM})'$, and $b_i = (b_{i1}, ..., b_{iK})'$. Also let $\beta = (\beta_{01}, \beta_{11}, ..., \beta_{0K}, \beta_{1K})'$ for model (3.1) or $\beta = (\beta_{011}, ...,$
$\beta_{01M}, \beta_{111}, ..., \beta_{11M}, ..., \beta_{0K1}, \ ..., \beta_{0KM}, \beta_{1K1}, ..., \beta_{1KM})'$ for model (3.3), $\sigma_b^2 = (\sigma_{b1}^2,$
$..., \sigma_{bK}^2)'$, and $\sigma_e^2 = (\sigma_{e1}^2, ..., \sigma_{eK}^2)'$. Then assuming no missing data and under models (3.1) and (3.3), we have the full likelihood function that is proportional to

$$L(\theta) = \prod_{i=1}^{n} \int f(Y_i|U_i, b_i; \beta, \sigma_e^2)\, f(U_i|a_i; \sigma_\epsilon^2)\, f(b_i; \sigma_b^2)\, f(a_i; \sigma_a^2)\, dY_i\, dU_i\, db_i\, da_i\,.$$

In the above, θ denotes all parameters β, α, σ_b^2, σ_e^2, σ_a^2 and σ_ϵ^2 together,

$$\begin{aligned} f(Y_i|U_i, b_i; \beta, \sigma_e^2) &= \prod_{k=1}^{K} \prod_{m=1}^{M} f_0(Y_{ikm}; \beta_{0k} + \beta_{1k} U_{im} + b_{ik}, \sigma_e^2) \\ f(U_i|a_i; \sigma_a, \sigma_\epsilon) &= \prod_{m=1}^{M} f_0(U_{im}; X_{im}'\alpha + a_i, \sigma_\epsilon^2) \\ f(b_i; \sigma_b^2) &= \prod_{k=1}^{K} f_0(b_{ik}; 0, \sigma_b^2) \\ f(a_i; \sigma_a^2) &= f_0(a_i; 0, \sigma_a^2), \end{aligned}$$

where $f_0(\cdot; \mu, \sigma^2)$ denotes the normal density function with mean μ and variance σ^2.

For estimation of θ, it is natural to apply the EM algorithm. For this, one can treat (Y_i, U_i, b_i, a_i) as the complete data, which gives the complete data log-likelihood function

$$l(\theta) = \sum_{i=1}^{n} \left[\log f(Y_i|U_i, b_i; \beta, \sigma_e^2) + \log f(U_i|a_i; \sigma_\epsilon^2) + \log f(b_i; \sigma_b^2) + \log f(a_i; \sigma_a^2) \right]. \tag{4.1}$$

The EM algorithm starts with choosing initial estimates of the parameters and then iterates between the M- and E-steps. The M-step maximizes the complete data log-likelihood function (4.1) to obtain the estimates of parameters, which involve the expectations of the functions of the latent variables U_i's, b_i's and a_i's. In the E-step, these expectations are calculated. Once the finial estimates are

obtained, their covariance matrix can be estimated by the inverse of the observed information matrix (Louis, 1982)

For the AHB data, as pointed before, there exist missing data for both response variables and covariates. For the response variables, one can simply base the analysis on the available values assuming that the missing values missed at random. For the missing covariates, one way is to treat the missing values the same way as that for the latent variables U_i's, b_i's and a_i's in the EM algorithm discussed above. It can be easily seen that this would make the estimation process much more complicated. For simplicity, in the following data analysis, we impute the missing covariate values using the first previous values that are available. We also tried other simple approaches and got similar results.

Now we consider estimation of the parameters in model (3.4). First assume that one is only interested in regression parameter τ. Under model (3.4), it is apparent that we have $E[\, N_{ikm} \exp(-X'_{ikm}\tau) \,|\, X_{ikm}] = \exp(\gamma_{km})$. This suggests a natural unbiased estimating equation given by

$$U_1(\tau) = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{m=1}^{M} X_{ikm}\, N_{ikm} \exp(-X'_{ikm}\tau) = 0 \tag{4.2}$$

assuming that $E(X_{ikm}) = 0$. If the expectation of X_{ikm} is not zero, we can simply replace it by $X_{ikm} - \bar{X}_{km}$, where $\bar{X}_{km} = n^{-1} \sum_{i=1}^{n} X_{ikm}$. The similar equations have been used by, among others, Sun and Wei (2000) for the analysis of univariate panel count data. Let $\hat{\tau}_1$ denote the solution to equation (4.2) and τ_0 the true value of τ. Then it can be shown that $\hat{\tau}_1$ is consistent and $\sqrt{n}\,(\hat{\tau}_1 - \tau_0)$ converges in distribution to a normal random vector with mean zero and the covariance matrix that can be estimated by $\hat{\Sigma}_\tau^{(1)} = \hat{F}_1^{-1}\, \hat{\Sigma}_U^{(1)}\, \hat{F}_1^{-1}$. In this formula,

$$\hat{F}_1 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{m=1}^{M} X_{ikm}\, X'_{ikm}\, N_{ikm} \exp(-X'_{ikm}\hat{\tau}_1)$$

and

$$\hat{\Sigma}_U^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \sum_{m=1}^{M} X_{ikm}\, N_{ikm} \exp(-X'_{ikm}\hat{\tau}_1) \right]^{\otimes 2}$$

with $a^{\otimes 2} = a\, a'$ for a vector a.

Sometimes we may be also interested in estimation of the γ_{km}'s in addition to τ. For this, note that under model (3.4), $E[\, N_{ikm} - \exp(X'_{ikm}\tau + \gamma_{km}) \,|\, X_{ikm}\,] = 0$. This suggests that we can consider the unbiased estimating equations

$$\sum_{m=1}^{M} \left(N_{ikm} - e^{X'_{ikm}\tau + \gamma_{km}} \right) X_{ikm} = 0\,. \tag{4.3}$$

for τ instead of (4.2), and

$$U_\gamma(\tau, \gamma'_{km}s) = \sum_{i=1}^{n} \left(N_{ikm} - e^{X'_{ikm}\tau+\gamma_{km}} \right) = 0 \tag{4.4}$$

for the γ_{km}'s. Let $\hat{\tau}_2$ and the $\hat{\gamma}_{km}$'s denote the solutions to equations (4.3) and (4.4). Then one can show that as $\hat{\tau}_1$, $\hat{\tau}_2$ is consistent and $\sqrt{n}\,(\hat{\tau}_2 - \tau_0)$ converges in distribution to a normal random vector with mean zero and the covariance matrix that can be estimated by $\hat{\Sigma}_\tau^{(2)} = \hat{F}_2^{-1}\,\hat{\Sigma}_U^{(2)}\,\hat{F}_2^{-1}$, where

$$\hat{F}_2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{m=1}^{M} X_{ikm}\, X'_{ikm}\, \exp(X'_{ikm}\hat{\tau}_2 + \hat{\gamma}_{km})$$

and

$$\hat{\Sigma}_U^{(2)} = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \sum_{m=1}^{M} X_{ikm}\, (N_{ikm} - e^{X'_{ikm}\hat{\tau}_2+\hat{\gamma}_{km}}) \right]^{\otimes 2}.$$

The proofs about the asymptotic results described above are beyond the scope of this paper and will be given somewhere else.

5. Analysis of the AHB Study

In this section, we apply the methods discussed in the previous sections to the AHB study described in Section 2. For the analysis, as mentioned before, we will focus on two response variables, ACON and DCON, representing the negative consequences of the alcohol and drug uses, respectively, and their relationship with covariates SEX, EEPQ, LESECY, NEOC and NEPQ. During the study, six measurements ($M = 6$) were supposed to be collected about all variables ($K = 2$, $p = 5$) from each study subject. However, all variables except SEX have some missing values as discussed before.

To obtain some basic ideas about the data, we first conducted a preliminary analysis by fitting the data to the commonly used Poisson model that assumes that ACON ($k = 1$) and DCON ($k = 2$) follow the Poisson process with the conditional mean

$$\exp(X'_{im}\,\tilde{\beta}_k + b) \tag{5.1}$$

given random effect b. Here X_{im} denotes the values at the mth time point of the vector of the intercept, YEAR and the five covariates, and $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are vectors of regression parameters, where YEAR represents the year at which ACON or DCON was measured. Model (5.1) can be easily fitted by using, for example, SAS PROC GLIMMIX. Table 2 presents the obtained results and includes the estimated effects of covariates, their standard error estimates, and the p-values

for testing covariate effects being zero. In the results given here and below, the covariate representing SEX is defined to be equal to 1 for female and 0 otherwise. It can be seen from Table 2 that all covariates are strongly associated with both ACON and DCON except that it seems that LESECY has no effect on ACON. It is interesting to note that the females in the study had significantly less consequences of the alcohol and drug uses than the males.

Table 2: Estimated effects of covariates based on Poisson model

	ACON			DCON		
	Estimate	SD	p-value	Estimate	SD	p-value
Intercept	-0.5466	0.4568	0.2317	-1.6151	0.6179	0.0091
YEAR	-0.0557	0.0215	0.0097	-0.0647	0.0370	0.0808
SEX	-0.8583	0.1473	< 0.0001	-0.9434	0.1876	< 0.0001
EEPQ	0.0544	0.0139	< 0.0001	0.04855	0.0182	0.0079
LESECY	0.0199	0.0127	0.1164	0.0714	0.0169	< 0.0001
NEOC	-0.0285	0.0085	0.0008	-0.0536	0.0112	< 0.0001
NEPQ	0.0505	0.0123	< 0.0001	0.0854	0.0157	< 0.0001

Table 3 gives the results obtained by fitting the model defined by (3.1) and (3.2) to the AHB study data. Here and also for the results given below under the model defined by (3.2) and (3.3), Y_{ikm} was taken to be the logarithm of ACON or DCON plus 0.001. The results in Table 3 suggest that both ACON and DCON are positively correlated with the latent scores U_{im}'s, which represent the overall effect of covariates on ACON and DCON measured by the estimates of the β_{0k}'s and β_{1k}'s. In terms of the latent scores and covariates, SEX and NEOC are negatively associated with the latent scores, while EEPQ, LESECY and NEPQ are positively associated with the latent scores. That is, SEX and NEOC are negatively correlated with the consequences of alcohol and drug uses and EEPQ, LESECY and NEPQ are positively correlated with the consequences of alcohol and drug uses. Also the analysis indicates that the consequences of alcohol and drug uses are strongly related ($\hat{\sigma}_a^2 = 0.600$ with the estimated standard deviation being 0.247). Although this is expected, there do not exist many analyses that provide such evidence since most of them are individual analysis. Note that the covariate YEAR represents the time effect on the latent scores and thus on ACON and DCON. The negative sign of the estimated coefficient for this covariate suggests that the consequences of alcohol and drug uses decrease when people become old.

Table 3: Results of the joint analysis based on the model defined by (3.1) and (3.2)

	ACON			DCON		
	Estimate	SD	p-value	Estimate	SD	p-value
$\hat{\beta}_{0k}$	-2.006	0.090	< 0.0001	-4.651	0.073	< 0.0001
$\hat{\beta}_{1k}$	0.026	0.001	< 0.0001	0.016	0.001	< 0.0001
	$\hat{\alpha}_1$(YEAR)	$\hat{\alpha}_2$(SEX)	$\hat{\alpha}_3$(EEPQ)	$\hat{\alpha}_4$(LESECY)	$\hat{\alpha}_5$(NEOC)	$\hat{\alpha}_6$(NEPQ)
Estimate	-6.393	-8.081	0.852	0.745	-0.607	1.053
SD	0.013	0.075	0.003	0.004	0.002	0.005
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

We now consider fitting the model defined by (3.2) and (3.3) to the data and the estimated regression parameters are given in Table 4 along with their standard error estimates and their corresponding p-values as in Tables 2 and 3. Note that under the model defined by (3.2) and (3.3), the time effects on ACON and DCON are nonparametrically incorporated into model (3.3) and thus unlike for the model defined by (3.1) and (3.2), we did not consider the covariate YEAR in this case. Although Table 4 shows that the model defined by (3.2) and (3.3) gave conclusions similar to those obtained under the model defined by (3.1) and (3.2), the estimated effects of all covariates on both latent scores and the consequences of alcohol and drug uses are much more significant. One possible reason for this is that model (3.1) assumes that the consequences are linear functions of time, which may be too restrictive, while model (3.3) imposes no restriction on the shapes of the consequences with respect to time, thus providing more insights. Again the analysis suggests that the consequences of alcohol and drug uses are significantly related with $\hat{\sigma}_a^2 = 0.2926$ and the estimated standard error of 0.029.

Table 4: Results of the joint analysis based on the model defined by (3.2) and (3.3)

	ACON			DCON		
	Estimate	SD	p-value	Estimate	SD	p-value
$\hat{\beta}_{0k1}$	-3.765	0.014	< 0.0001	-6.107	0.016	< 0.0001
$\hat{\beta}_{0k2}$	-4.050	0.014	< 0.0001	-5.924	0.012	< 0.0001
$\hat{\beta}_{0k3}$	-4.213	0.014	< 0.0001	-6.170	0.012	< 0.0001
$\hat{\beta}_{0k4}$	-4.526	0.014	< 0.0001	-6.491	0.012	< 0.0001
$\hat{\beta}_{0k5}$	-4.783	0.015	< 0.0001	-6.361	0.012	< 0.0001
$\hat{\beta}_{0k6}$	-5.133	0.016	< 0.0001	-6.376	0.013	< 0.0001
$\hat{\beta}_{1k}$	0.058	0.0002	< 0.0001	0.042	0.0002	< 0.0001
	$\hat{\alpha}_1$(SEX)	$\hat{\alpha}_2$(EEPQ)	$\hat{\alpha}_3$(LESECY)	$\hat{\alpha}_4$(NEOC)	$\hat{\alpha}_5$(NEPQ)	
Estimate	-15.951	1.880	1.669	-1.273	2.106	
SD	0.106	0.005	0.006	0.002	0.008	
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	

It can be seen from Table 4 that the consequence of alcohol use seems to decrease along with time, but the consequence of drug use seems to stay at the same level respect to time. To give a graphical idea, the middle panel of Figure

1 displays the estimated averages of the consequences of alcohol and drug uses for males with all EEPQ, LESECY, NEOC and NEPQ set to equal to zero. As commented above, although the consequences of both alcohol and drug uses showed the decreasing trends, the change of the consequence of drug use with respect to time is relatively much less significant. To give a comparison between females and males, the bottom panel of Figure 1 presents the estimated averages of the consequences of alcohol and drug uses for both females and males with all EEPQ, LESECY, NEOC and NEPQ set to be zero. As shown in Table 4, the males in the study clearly had significantly higher consequences of alcohol and drug uses than the females. It is worth noting by comparing the middle and bottom panels of Figure 1 to top panel that after controlling the covariates, the decreasing rates of the consequences of both alcohol and drug uses are much mild.

Table 5: Estimated effects of covariates based on the marginal model (4.1)

	$\hat{\tau}_1$			$\hat{\tau}_2$		
Factor	Estimate	SD	p-value	Estimate	SD	p-value
SEX	-0.533	0.136	0.0001	-0.459	0.083	< 0.0001
EEPQ	0.036	0.013	0.0079	0.041	0.008	< 0.0001
LESECY	0.026	0.009	0.0050	0.034	0.018	0.0633
NEOC	-0.016	0.008	0.0589	-0.027	0.011	0.0116
NEPQ	0.059	0.012	< 0.0001	0.054	0.008	< 0.0001

Table 6: Estimates of the baseline parameters for the marginal model (4.1)

Parameter	$\hat{\gamma}_{k1}$	$\hat{\gamma}_{k2}$	$\hat{\gamma}_{k3}$	$\hat{\gamma}_{k4}$	$\hat{\gamma}_{k5}$	$\hat{\gamma}_{k6}$
			ACON			
Estimate	-0.323	0.372	0.771	1.018	1.391	1.634
SD	0.0566	0.0494	0.0549	0.0310	0.0260	0.0867
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
			DCON			
Estimate	-1.069	-0.298	0.097	0.283	0.620	0.815
SD	0.0722	0.0869	0.1186	0.0506	0.0558	0.0816
p-value	< 0.0001	0.0006	0.4127	< 0.0001	< 0.0001	< 0.0001

To fit model (3.4) to the AHB study data, as discussed above, define N_{ikm} to be the total number of the consequences of alcohol or drug use recorded at the first m time points for subject i. The application of the estimation procedures described in the previous section yielded the results presented in Table 5, which includes both $\hat{\tau}_1$ and $\hat{\tau}_2$ as well as the estimates of their standard errors and the p-values for testing each of covariate effects equal to zero. The results about the estimates of the γ_{km}'s corresponding to $\hat{\tau}_2$ and obtained from equations

(4.3) and (4.4) are given in Table 6. It can be seen from Table 5 that the two estimation procedures for τ gave similar conclusions about covariate effects except that the estimated effects for LESECY and NEOC are little different. Also the results obtained here support the conclusions given by the model defined by (3.1) and (3.2) or (3.2) and (3.3), and this suggests that these models provide reasonable approximations to the underlying processes of the alcohol and drug use consequences.

In summary, the analyses indicate that all covariates considered had significant effects on the consequences of alcohol and drug uses. The female subjects seem to have much less consequences of alcohol and drug uses than the male subjects. One possible reason for this is that females may have lower rates of alcohol and drug uses, or they have better self-control than males. The study suggests that the consequences of alcohol and drug uses are in general positively associated with EEPQ, LESECY and NEPQ but negatively associated with NEOC. Also it indicates that as expected, the consequences due to alcohol and drug uses are significantly and positively related. Furthermore, the consequences seem to decrease when people become old with the consequence of alcohol use decreasing more significantly than that of drug use. This could be because people have more control on themselves or use alcohol and drug less when they are old.

6. Concluding Remarks

Several statistical approaches have been proposed for the analysis of alcohol and drug uses with the focus on the temporary trends of the consequences of alcohol and drug uses and their relationship with characteristics of alcohol and drug users. One key feature of these methods is that they allow one to perform joint analysis of alcohol and drug uses and to assess the association between them. Also unlike many existing approaches for this type of studies, the method based on model (3.4) does not require a distribution assumption and thus is more robust than the methods relying on, for example, the normality or Poisson assumption.

With respect to the comparison of the two types of models presented in the preceding sections, the latent variable models specifically model the underlying processes of the alcohol and drug uses and are appropriate if one needs the detailed description of the processes. This would be the case if one is interested in the specific process for a particular group of individuals or prediction. The idea behind these models and the estimation procedure are straightforward although the implementation of the estimation procedure may not be simple. As mentioned before, a major shortcoming is the normality assumption as the Poisson assumption. In contrast, the marginal mean model should be used if the focus is on the effects of covariates such as gender and personality. It does not need the normality assumption, but apparently cannot be used for prediction.

As mentioned before, the main goal of this paper is to provide statistical methods for alcohol and drug studies that do not rely on Poisson assumptions. In terms of the variables ACON and DCON, of course, it seems natural to apply Poisson-based methods as many authors did before. On the other hand, as it is well-known, the Poisson assumption is quite restrictive, while the models presented in the previous sections are relatively less restrictive and easier to check. Also they seem to be more appropriate for longitudinal alcohol and drug data.

Although the main purposes of this paper are to analyze the AHB study and to provide statistical methodology for assessing alcohol and drug uses, the methods discussed also apply to the analysis of similar longitudinal studies, which are common in, for example, medical follow-up studies and clinical trials. One example is studies on HIV or AIDS patients where many markers such as CD4 and RNA are usually measured together. In these cases, the proposed models and inference procedures, especially the marginal model (3.4), provide flexible methodology for their joint analysis. The same is true for many cancer studies.

Acknowledgements

The authors wish to thank a referee for his/her many helpful and insightful comments and suggestions which greatly improved the paper.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Spring-Verlag.

Cai, J. and Schaubel, D. E (2004). Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Analysis* **10**, 121-138.

Costa P. T. Jr. and McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEOFFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *The Analysis of Longitudinal Data*. Oxford.

Eysenck and H. J. (1988). *Eysenck Personality Questionnaire-Revised*. Educational and Industrial Testing Services.

Jackson, K. M. and Sher, K. J. (2003). Alcohol use disorders and psychological distress: A prospective state-trait analysis. *Journal of Abnormal Psychology* **112**, 599-613.

Jackson, K. M., Sher, K. J., Cooper, M. L. and Wood, P. K. (2002). Adolescent alcohol and tobacco use: Onset, persistence and trajectories of use across two samples. *Addiction* **97**, 517-531.

Jackson, K. M., Sher, K. J. and Schulenberg, J. E. (2005). Conjoint developmental trajectories of young adult alcohol and tobacco use. *Journal of Abnormal Psychology* **114**, 612-626.

Jackson, K. M., Sher, K. J. and Wood, P. K. (2000). Prospective analysis of comorbidity: Tobacco and alcohol use disorders. *Journal of Abnormal Psychology* **109**, 679-694.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-74.

Lawless, J. F. and Nadeau, J. C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**, 158-168.

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

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine* **13**, 2233-2247.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **96**, 103-126.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of og the Royal Statistical Society B* **44**, 226-233.

Nadeau, J. C. and Lawless, J. F. (1998). References for means and covariances of point processes through estimating functions. *Biometrika* **85**, 983-906.

Sarason, I. G., Johnson, J. H. and Siegel, J. M. (1978). Assessing the impact of life changes: Development of the life experiences survey. *Journal of Consulting and Clinical Psychology* **46**, 932-946.

Selzer, M., Vinokur, A. and van Rooijen, L. (1975). A self-administrated short Michigan alcoholism screening test (SMAST). *Journal of Studies on Alcohol* **36**, 117-126.

Shah, A., Laird, N. M. and Schoenfeld, D. (1997). Random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association* **92**, 775-779.

Sher, K. J., Walitzer, K. S., Wood, P. K. and Brent, E. E. (1991). Characteristics of children of alcoholics: Putative risk factors, substance use and abuse, and psychopathology. *Journal of Abnormal Psychology* **100**, 427-448.

Sun, J. and Wei, L. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society: Series B* **62**, 293-2000.

Trull, T. J., Waudby, C. J. and Sher, K. J. (2004). Alcohol, tobacco, and drug use disorders and personality disorder symptoms. *Experimental and Clinical Psychopharmacology* **12**, 65-75.

Zeger, S. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

Received October 2, 2007; accepted January 2, 2008.

Liang Zhu
Department of Statistics
University of Missouri
146 Middlebush Hall
Columbia, MO, 65211, USA
sunj@missouri.edu

Jianguo Sun
Department of Statistics
University of Missouri
146 Middlebush Hall
Columbia, MO, 65211, USA
lzkn7@mizzou.edu

Phillip Wood
Department of Psychological Sciences
University of Missouri
210 McAlester Hall
Columbia, MO 65211, USA
phillipkwood@yahoo.com