On Measurement Error Problems with Predictors Derived from Stationary Stochastic Processes and Application to Cocaine Dependence Treatment Data

Yehua Li
Department of Statistics
University of Georgia

Yongtao Guan, Rajita Sinha
Yale University
Cocaine dependence data

Methodology
  Method of moment estimator
  Subsampling extrapolation (SUBEX)
  Interval censored failure time

Simulation study

Data analysis result

Summary
Cocaine dependence

- Cocaine dependence remains a serious public health problem in the US with more than one million individuals meeting criteria for current dependence (SAMSHA, 2004).
Cocaine dependence remains a serious public health problem in the US with more than one million individuals meeting criteria for current dependence (SAMSHA, 2004).

Cocaine abuse causes serious concerns to the society because of its association with criminal activities as well as the high cost to treat health problems related to it.
Cocaine dependence remains a serious public health problem in the US with more than one million individuals meeting criteria for current dependence (SAMSHA, 2004).

Cocaine abuse causes serious concerns to the society because of its association with criminal activities as well as the high cost to treat health problems related to it.

Despite the availability of efficacious behavioral treatments, relapse rates remain high (Sinha, 2001, 2007).
Cocaine relapse and baseline usage behavior

- Previous studies have revealed that one's **baseline cocaine use behavior** is predictive of cocaine relapse, along with many other risk factors such as age and gender, cocaine withdrawal severity, and stress and negative mood.
Cocaine relapse and baseline usage behavior

- Previous studies have revealed that one’s **baseline cocaine use behavior** is predictive of cocaine relapse, along with many other risk factors such as age and gender, cocaine withdrawal severity, and stress and negative mood.

- To describe the baseline cocaine use behavior, daily cocaine use trajectory data are collected in a short period prior to treatment.
Cocaine relapse and baseline usage behavior

- Previous studies have revealed that one’s baseline cocaine use behavior is predictive of cocaine relapse, along with many other risk factors such as age and gender, cocaine withdrawal severity, and stress and negative mood.
- To describe the baseline cocaine use behavior, daily cocaine use trajectory data are collected in a short period prior to treatment.
- Summary statistics derived from these trajectories are then used as predictors in a subsequent analysis to explain cocaine craving and cocaine relapse.
The data

The research was conducted at the Clinical Neuroscience Research Unit (CNRU) of the Connecticut Mental Health Center.
The research was conducted at the Clinical Neuroscience Research Unit (CNRU) of the Connecticut Mental Health Center.

The study was conducted in two separate periods with 59 enrolling in the first period and 83 enrolling in the second.
The research was conducted at the Clinical Neuroscience Research Unit (CNRU) of the Connecticut Mental Health Center.

The study was conducted in two separate periods with 59 enrolling in the first period and 83 enrolling in the second.

At the baseline, demographic variables (e.g. age, gender, race, years of cocaine use, anxiety level) and daily cocaine use history in the 90 days prior to admission were collected.
The research was conducted at the Clinical Neuroscience Research Unit (CNRU) of the Connecticut Mental Health Center.

The study was conducted in two separate periods with 59 enrolling in the first period and 83 enrolling in the second.

At the baseline, demographic variables (e.g. age, gender, race, years of cocaine use, anxiety level) and daily cocaine use history in the 90 days prior to admission were collected.

The 59 subjects in the first period were interviewed 90 days after treatment.
The research was conducted at the Clinical Neuroscience Research Unit (CNRU) of the Connecticut Mental Health Center.

The study was conducted in two separate periods with 59 enrolling in the first period and 83 enrolling in the second.

At the baseline, demographic variables (e.g. age, gender, race, years of cocaine use, anxiety level) and daily cocaine use history in the 90 days prior to admission were collected.

The 59 subjects in the first period were interviewed 90 days after treatment.

The 83 subjects in the second study were interviewed 14, 30, 90 and 180 days after treatment. Urine screening were conducted on these interviews to ensure the accuracy of the first relapse time.
Response variables of interest

- **Post-treatment cocaine craving score**: desire for using cocaine at this moment, measured by the Tiffany Cocaine Craving Questionnaire-Brief (CCQ-Brief) (Tiffany et al., 1993).
Response variables of interest

- **Post-treatment cocaine craving score**: desire for using cocaine at this moment, measured by the Tiffany Cocaine Craving Questionnaire-Brief (CCQ-Brief) (Tiffany et al., 1993).

- **First relapse time**: the first time that the subject use cocaine after the treatment.
Response variables of interest

- **Post-treatment cocaine craving score**: desire for using cocaine at this moment, measured by the Tiffany Cocaine Craving Questionnaire-Brief (CCQ-Brief) (Tiffany et al., 1993).

- **First relapse time**: the first time that the subject use cocaine after the treatment.
  - At each post treatment interview, the subject will report whether he/she used cocaine, and when.
Response variables of interest

- **Post-treatment cocaine craving score**: desire for using cocaine at this moment, measured by the Tiffany Cocaine Craving Questionnaire-Brief (CCQ-Brief) (Tiffany et al., 1993).

- **First relapse time**: the first time that the subject used cocaine after the treatment.
  - At each post treatment interview, the subject will report whether he/she used cocaine, and when.
  - In the second period, urine test was used to check the validity of the self-reported relapse time.
Response variables of interest

- **Post-treatment cocaine craving score**: desire for using cocaine at this moment, measured by the Tiffany Cocaine Craving Questionnaire-Brief (CCQ-Brief) (Tiffany et al., 1993).

- **First relapse time**: the first time that the subject use cocaine after the treatment.
  - At each post treatment interview, the subject will report whether he/she used cocaine, and when.
  - In the second period, urine test was used to check the validity of the self-reported relapse time.
  - The relapse time are interval censored.
Assessing baseline cocaine use pattern

Upon treatment entry, all subjects were interviewed by well-trained psychologists to collect baseline daily cocaine use history in the 90 days prior to admission, which is documented using a 90-day time-line follow-back (TLFB) Substance Use Calendar.
Assessing baseline cocaine use pattern

- Upon treatment entry, all subjects were interviewed by well-trained psychologists to collect baseline daily cocaine use history in the 90 days prior to admission, which is documented using a 90-day time-line follow-back (TLFB) Substance Use Calendar.

- It has been shown that the TLFB can provide reliable daily cocaine use data that have high 1) retest reliability, 2) correlation with other cocaine use measures and 3) agreement with collateral informants’ reports of patients’ cocaine use as well as results obtained from urine assays.
Assessing baseline cocaine use pattern

Upon treatment entry, all subjects were interviewed by well-trained psychologists to collect baseline daily cocaine use history in the 90 days prior to admission, which is documented using a 90-day time-line follow-back (TLFB) Substance Use Calendar.

It has been shown that the TLFB can provide reliable daily cocaine use data that have high 1) retest reliability, 2) correlation with other cocaine use measures and 3) agreement with collateral informants’ reports of patients’ cocaine use as well as results obtained from urine assays.

All research assistants had been trained by PhD level psychologists and had over three-year experience in administration of similar assessments and they were closely supervised when conducting these interviews.
Assessing baseline cocaine use pattern

- All subjects had been informed upfront that all data are coded and confidential that they can not be summoned in court.
- They were also informed that they would be removed from the study if they were found out not being truthful about their drug use.
Baseline usage pattern
Summary statistics and measurement error

- The true covariates are characteristics of the patients’ long term cocaine use pattern.
- The surrogates are summary statistics from the 90 day baseline usage trajectory, e.g. average daily use amount, use frequency.
- These trajectories are assumed to be stationary, and we consider two kinds of trajectories
  - marked point pattern: use time and amount;
  - point pattern: use time only.
- The point pattern is more reliable because a) people tend to under report the use amount, but there is less incentive to deny a use; b) the subjects use different ways to use cocaine, it is hard to convert use amount into equivalent grams.
The baseline trajectories are subject specific stochastic processes.

The estimation error in the summary statistics are heteroscedastic, non-Gaussian, and dependent on the true covariate. We only have one realization of the random process.

In the joint modeling literature, there is work on using the random slope and intercept of longitudinal processes in a second level regression model. But a strong parametric assumption on the measurement error need to be made.

In functional data analysis, characteristics of random curves are used in second level regression (Crainiceanu et al. 2009).
Notation and setup

- Let $N_i$ be a stationary stochastic process that has generated the $i$th cocaine use trajectory over the baseline period $(0, \tau]$.

\[ W_i = \frac{1}{\tau} \int_0^\tau N_i^*(dt), \]  

(1)

where $N_i^*$ is either $N_i$ or a different stationary process that is derived from $N_i$.

- The distribution of $N_i$ depends on a set of parameters $\Lambda_i$, $X_i = E(W_i|\Lambda_i)$ is the true predictor.

- $W_i = X_i + U_i$, $\sigma_i^2 = \text{var}(U_i)$ might depends on $X_i$ and is different across the subjects.
Let $S(t, p) = (t, t + p\tau]$ be a subinterval, where $0 < p \leq 1$ and $0 \leq t \leq (1 - p)\tau$, $W_i(t, p)$ is the summary statistics defined on $S(t, p)$, and $U_i(t, p) = W_i(t, p) - X_i$. 

$$\hat{\sigma}^2_i = \lim_{\tau \to \infty} (\tau \sigma^2_u - 1 - p\alpha_i + o(1/p\tau^2)).$$ 

Weak Dependency Assumption: We assume that 

$$\text{Var}[U_i(t, p)|\Lambda_i] = \hat{\sigma}^2_i - \frac{1}{p\tau} \alpha_i + o(1),$$ 

where $\alpha_i$ is a constant that is typically nonnegative.
Let $S(t, p) = (t, t + p\tau]$ be a subinterval, where $0 < p \leq 1$ and $0 \leq t \leq (1 - p)\tau$, $W_i(t, p)$ is the summary statistics defined on $S(t, p)$, and $U_i(t, p) = W_i(t, p) - X_i$.

Define $\tilde{\sigma}_i^2 = \lim_{\tau \to \infty} (\tau \sigma_{u_i}^2)$.

**Weak Dependency Assumption:** We assume that

$$(p\tau) \text{Var}[U_i(t, p)|\Lambda_i] = \tilde{\sigma}_i^2 - \frac{1}{p\tau} \alpha_i + o\left(\frac{1}{p\tau}\right), \quad (2)$$

where $\alpha_i$ is a constant that is typically nonnegative.
Notation and setup (Cont.)

- Let $S(t, p) = (t, t + p\tau]$ be a subinterval, where $0 < p \leq 1$ and $0 \leq t \leq (1 - p)\tau$, $W_i(t, p)$ is the summary statistics defined on $S(t, p)$, and $U_i(t, p) = W_i(t, p) - X_i$.
- Define $\tilde{\sigma}_i^2 = \lim_{\tau \to \infty} (\tau\sigma_{ui}^2)$.
- **Weak Dependency Assumption:** We assume that

\[(p\tau) \text{Var}[U_i(t, p)|\Lambda_i] = \tilde{\sigma}_i^2 - \frac{1}{p\tau} \alpha_i + o \left(\frac{1}{p\tau}\right), \tag{2}\]

where $\alpha_i$ is a constant that is typically nonnegative.
- By the weak dependence assumption,

\[\text{Var}[U_i(t, p)|X_i] \approx \frac{\sigma_{ui}^2}{p} - \frac{1 - p}{p^2\tau^2} \alpha_i, \tag{3}\]

if $p\tau$ is sufficiently large.
Method of moment bias correction in linear model

- We consider $Y_i = X_i \beta + Z_i^T \eta + \epsilon_i$, $\theta = (\beta, \eta)$.
- The naive estimator using the surrogate $W$,

$$
\hat{\theta}_{\text{naive}} = \left[ \begin{array}{cc} W^T W & W^T Z \\ Z^T W & Z^T Z \end{array} \right]^{-1} \left[ \begin{array}{c} W^T \\ Z^T \end{array} \right] Y \equiv A^{-1} B.
$$
We consider \( Y_i = X_i\beta + Z_i^T\eta + \epsilon_i, \theta = (\beta, \eta). \)

The naive estimator using the surrogate \( W \),

\[
\hat{\theta}_{\text{naive}} = \left[ \begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array} \right]^{-1} \left[ \begin{array}{c} W^TY \\ Z^TY \end{array} \right] \equiv A^{-1}B.
\]

\[
E(A|X, Z) = \left( \begin{array}{cc} X^TX + \sigma^2 & X^TZ \\ Z^TX & Z^TZ \end{array} \right), \quad E(B|X, Z) = \left( \begin{array}{cc} X^TX & X^TZ \\ Z^TX & Z^TZ \end{array} \right) \theta,
\]

where \( \sigma^2 = E(U^TU) = \sum_{i=1}^n \sigma_{u_i}^2 \).
Method of moment bias correction in linear model

- We consider $Y_i = X_i \beta + Z_i^T \eta + \epsilon_i, \theta = (\beta, \eta)$.
- The naive estimator using the surrogate $W$,

\[ \hat{\theta}_{naive} = \left[ \begin{array}{cc} W^T W & W^T Z \\ Z^T W & Z^T Z \end{array} \right]^{-1} \left[ \begin{array}{c} W^T \\ Z^T \end{array} \right] Y \equiv A^{-1} B. \]

\[ E(A|X, Z) = \left( \begin{array}{cc} X^T X + \sigma^2 & X^T Z \\ Z^T X & Z^T Z \end{array} \right), \quad E(B|X, Z) = \left( \begin{array}{cc} X^T X & X^T Z \\ Z^T X & Z^T Z \end{array} \right) \theta, \]

where $\sigma^2 = E(U^T U) = \sum_{i=1}^{n} \sigma_{u_i}^2$.
- The method of moment estimator $\hat{\theta}_{\text{mom}}$ is obtained by subtracting $\hat{\sigma}^2$ from the first entry of $A$. 
Subsampling estimation of the variance

- Let \( t \) be an arbitrary time point with \( 0 \leq t < \tau - kp\tau \), where \( k = [1/p] \). We then partition the interval \((t, t + kp\tau]\) into \( k \) nonoverlapping subintervals each of length \( p\tau \). We require \( p \leq 1/2 \) so that \( k \geq 2 \).

- Let \( W_i(t + jp\tau, p) \) be the summary statistic defined on the \((j + 1)\)th subinterval for the \(i\)th trajectory, where \( j = 0, 1, \ldots, k - 1 \). Define

\[
\tilde{\sigma}^2_{ui}(t, p) = \frac{1}{k} \sum_{j=0}^{k-1} [W_i(t + jp\tau, p) - W_i(t, kp)]^2.
\]
Let $t$ be an arbitrary time point with $0 \leq t < \tau - kp\tau$, where $k = [1/p]$. We then partition the interval $(t, t + kp\tau]$ into $k$ nonoverlapping subintervals each of length $p\tau$. We require $p \leq 1/2$ so that $k \geq 2$.

Let $W_i(t + jp\tau, p)$ be the summary statistic defined on the $(j + 1)$th subinterval for the $i$th trajectory, where $j = 0, 1, \ldots, k - 1$. Define

$$\tilde{\sigma}_{ui}^2(t, p) = \frac{1}{k} \sum_{j=0}^{k-1} \left[ W_i(t + jp\tau, p) - W_i(t, kp) \right]^2.$$

$$\tilde{\sigma}_{ui}^2(p) = \frac{1}{\tau - kp\tau} \int_0^{\tau - kp\tau} \tilde{\sigma}_{ui}^2(t, p) dt,$$
Subsampling estimation of the variance (cont.)

By condition (1), and define $\alpha = \sum_{i=1}^{n} \alpha_i$, then

$$E\left[\sum_{i=1}^{n} \tilde{\sigma}_{u_i}^2(p)\right] \approx \frac{1}{p} \left(1 - \frac{1}{k}\right) \sigma^2 - \frac{1}{p^2 \tau^2} \left(1 - p - \frac{1 - kp}{k^2}\right) \alpha. \quad (4)$$
Subsampling estimation of the variance (cont.)

- By condition (1), and define $\alpha = \sum_{i=1}^{n} \alpha_i$, then
  \[
  E\left[ \sum_{i=1}^{n} \tilde{\sigma}_{u_i}^2(p) \right] \approx \frac{1}{p} \left( 1 - \frac{1}{k} \right) \sigma^2 - \frac{1}{p^2 \tau^2} \left( 1 - p - \frac{1 - kp}{k^2} \right) \alpha. \tag{4}
  \]

- **Scheme 1**: Ignore correlation, and $\hat{\sigma}^2 = p/(1 - \frac{1}{k}) \sum_{i=1}^{n} \tilde{\sigma}_{u_i}^2(p)$. 
Subsampling estimation of the variance (cont.)

- By condition (1), and define \( \alpha = \sum_{i=1}^{n} \alpha_i \), then
  
  \[
  \mathbb{E} \left[ \sum_{i=1}^{n} \tilde{\sigma}_{u_i}^2(p) \right] \approx \frac{1}{p} \left( 1 - \frac{1}{k} \right) \sigma^2 - \frac{1}{p^2 \tau^2} \left( 1 - p - \frac{1 - kp}{k^2} \right) \alpha. \tag{4}
  \]

- **Scheme 1**: Ignore correlation, and \( \hat{\sigma}^2 = \frac{p}{1 - \frac{1}{k}} \sum_{i=1}^{n} \tilde{\sigma}_{u_i}^2(p) \).

- **Scheme 2**: Take into account of correlation, let
  
  \[
  \tilde{Y}(p) = \frac{p \sum_{i=1}^{n} \tilde{\sigma}_{u_i}^2(p)}{1 - 1/k} = \sum_{i=1}^{n} \tilde{Y}_i(p) \text{ and } \tilde{X}(p) = \frac{1 - p - (1 - kp)/k^2}{p\tau^2(1 - 1/k)}. \tag{5}
  \]

  Regress \( \tilde{Y}(p) \) on \( \tilde{X}(p) \) at some preselected values \( (p_1, \cdots, p_J) \) where \( J \geq 2 \). The resulting intercept and (minus) slope estimators, are \( \hat{\sigma}^2 \) and \( \hat{\alpha} \).
Within trajectory dependence

**Figure:** Scatter plot of $\tilde{X}(p)$ and $\tilde{Y}(p)$ for $p = k/\tau$ where $k = 30, 31, \cdots, 40$ and $\tau = 80$. 

Yehua Li, University of Georgia
Subsampling extrapolation

- Like SIMEX, we inflate the error in the estimator, find out how the estimator change as the level of error increases.
- Not knowing the true distribution of measurement error, we use subsampling to create surrogate with inflated error variance, instead of adding simulated Gaussian error.
Subsampling extrapolation

- Like SIMEX, we inflate the error in the estimator, find out how the estimator change as the level of error increases.

- Not knowing the true distribution of measurement error, we use subsampling to create surrogate with inflated error variance, instead of adding simulated Gaussian error.

- Let $W_i(t, p)$ be the summary statistic defined over the subintervals $S(t, p) = (t, t + p\tau]$,

\[
\text{Var}[W_i(t, p)|\Lambda_i] \approx \frac{\sigma^2_{ui}}{p} - \frac{1 - p}{p^2\tau^2} \alpha_i = \left(\frac{1}{p} - \frac{1 - p}{p^2\tau^2} \frac{\alpha_i}{\sigma^2_{ui}}\right) \sigma^2_{ui} \\
\equiv \chi_i(p)\sigma^2_{ui}.
\]
Let $\hat{\theta}(p; t_1, \cdots, t_n)$ be the estimated regression coefficient based on 
$\{Y_i, W_i(t_i, p), Z_i\}$, where $t_i \in [0, (1 - p)\tau]$, $i = 1, \cdots, n$, without accounting for measurement error.

$$
\hat{\theta}(p) = \frac{1}{[(1 - p)\tau]^n} \int_0^{(1-p)\tau} \cdots \int_0^{(1-p)\tau} \hat{\theta}(p; t_1, \cdots, t_n) \, dt_1 \cdots dt_n.
$$

(7)

The integral is evaluated by Monte-Carlo.

The procedure is repeated for a set of preselected $(p_1, \ldots, p_J)$. We fit a parametric model for $(\hat{x}(p), \hat{\theta}(p))$, e.g.

$$
G_Q(p, \Gamma) = \gamma_1 + \hat{x}(p)\gamma_2 + \hat{x}(p)^2\gamma_3,
$$

where $\Gamma = (\gamma_1, \gamma_2, \gamma_3)$. Normally, a quadratic or cubic extrapolant function is used.

$$
\hat{\theta}_{\text{subex}} = G_Q(p = \infty, \hat{\Gamma}).
$$
The variance inflation factor is

\[ x(p) = \frac{1}{p} \left( 1 - \frac{1 - p}{p \tau^2} \frac{\alpha}{\sigma^2} \right). \]

**Scheme 1:** Ignore correlation, \( x_i(p) \approx 1/p \). This is valid if \( \alpha_i/(\tau^2 \sigma^2_{u_i}) \) is small and/or \( p \tau \) is large.

**Scheme 2:** Plug in \( \hat{\sigma}^2 \) and \( \hat{\alpha} \) from the regression between \( \hat{Y}(p) \) and \( \hat{X}(p) \).
Two naive alternatives that ignore within-subject correlation

Let $W_i(0, 1/2)$ and $W_i(\tau/2, 1/2)$ be the counterparts of $W_i$ based on the data in $(0, \tau/2]$ and $(\tau/2, \tau]$.

- **Naive MOM:** estimate $\sigma^2$ by

$$\hat{\sigma}_{naive}^2 = \sum_{i=1}^{n} \tilde{\sigma}_{u_i}^2 (1/2) = \frac{1}{4} \sum_{i=1}^{n} [W_i(0, 1/2) - W_i(\tau/2, 1/2)]^2.$$
Two naive alternatives that ignore within-subject correlation

Let $W_i(0, 1/2)$ and $W_i(\tau/2, 1/2)$ be the counterparts of $W_i$ based on the data in $(0, \tau/2]$ and $(\tau/2, \tau]$.

- **Naive MOM:** estimate $\sigma^2$ by

  $$\hat{\sigma}_{naive}^2 = \sum_{i=1}^{n} \tilde{\sigma}_{u_i}^2(1/2) = \frac{1}{4} \sum_{i=1}^{n} [W_i(0, 1/2) - W_i(\tau/2, 1/2)]^2.$$ 

- **Empirical SIMEX** (Devanarayan and Stefanski, 2002): use the pseudo “remeasurements”

  $$W_i(\zeta) = W_i + \frac{\sqrt{\zeta}}{2} [W_i(0, 1/2) - W_i(\tau/2, 1/2)], \text{ for each } \zeta > 0.$$
We model the first relapse time $T_i$ through a Cox proportional hazard model (Cox, 1972),

$$
\lambda(t|X_i, Z_i) = \lambda_0(t) \exp(X_i \beta + Z_i \eta).
$$

(8)
We model the first relapse time $T_i$ through a Cox proportional hazard model (Cox, 1972),

$$\lambda(t|X_i, Z_i) = \lambda_0(t) \exp(X_i\beta + Z_i\eta).$$  \hspace{1cm} (8)

In our data, about 50.6% of the subjects have observed relapse time, 31.6% are interval censored and 17.8% are right censored.
We model the first relapse time $T_i$ through a Cox proportional hazard model (Cox, 1972),

$$\lambda(t|X_i, Z_i) = \lambda_0(t) \exp(X_i \beta + Z_i \eta).$$  \hspace{1cm} (8)

In our data, about 50.6% of the subjects have observed relapse time, 31.6% are interval censored and 17.8% are right censored.

Following Ruppert et al. (2003), Cai and Betensky (2003),

$$\psi(t) = \log \lambda_0(t) = a_0 + a_1 t + \sum_{k=1}^{K} b_k (t - \kappa_k)_+,$$  \hspace{1cm} (9)

where $x_+ \equiv \max(x, 0)$, and $\kappa_k$’s are the knots.
Penalized spline approach of Cai and Betensky (2003)

- We observe $n$ independent tuples, $(T_{li}, T_{ri}, \delta_i)$, where $[T_{li}, T_{ri}]$ gives the censoring interval, $\delta_i$ is the indicator for right censoring. When $\delta_i = 0$ and $T_{li} = T_{ri}$, the event time $T_i$ is right censored at $T_{ri}$; when $\delta_i = 1$ and $T_{li} < T_{ri}$, $T_i$ is interval censored within $[T_{li}, T_{ri}]$; when $\delta_i = 1$ and $T_{li} = T_{ri}$, $T_i$ is observed at $T_{ri}$. In addition, let $\delta_{0i}$ be the indicator for observed data, i.e. $\delta_{0i} = I(\delta_i = 1, T_{li} = T_{ri})$. 

\[ \ell(\Theta) = n \sum_{i=1}^{n} \delta_{0i} \left\{ \log \lambda_0(T_{ri}) + X_i^T \theta \right\} - \exp(X_i^T \theta) \Lambda_0(T_{ri}) + \delta_i (1 - \delta_{0i}) \log \left( \exp \left\{ \Lambda_0(T_{ri}) - \Lambda_0(T_{li}) \right\} \exp(X_i^T \theta) - 1 \right) \]

(10)

Where $\Theta$ is the collection of all parameters.
Penalized spline approach of Cai and Betensky (2003)

- We observe $n$ independent tuples, $(T^l_i, T^r_i, \delta_i)$, where $[T^l_i, T^r_i]$ gives the censoring interval, $\delta_i$ is the indicator for right censoring. When $\delta_i = 0$ and $T^l_i = T^r_i$, the event time $T_i$ is right censored at $T^r_i$; when $\delta_i = 1$ and $T^l_i < T^r_i$, $T_i$ is interval censored within $[T^l_i, T^r_i]$; when $\delta_i = 1$ and $T^l_i = T^r_i$, $T_i$ is observed at $T^r_i$. In addition, let $\delta_{0i}$ be the indicator for observed data, i.e. $\delta_{0i} = I(\delta_i = 1, T^l_i = T^r_i)$.

- Denote $\mathbf{X} = (X^T, Z^T)^T$ and $\Lambda_0(t) = \int_0^t \lambda_0(u)du$.

\[
\ell(\Theta) = \sum_{i=1}^{n} \delta_{0i} \{\log \lambda_0(T^r_i) + \mathbf{X}_i^T \theta\} - \exp(\mathbf{X}_i^T \theta)\Lambda_0(T^r_i) + \delta_i (1 - \delta_{0i}) \log \left( \exp[\{\Lambda_0(T^r_i) - \Lambda_0(T^l_i)\} \exp(\mathbf{X}_i^T \theta)] - 1 \right),
\]

where $\Theta$ is the collection of all parameters.
Penalized spline approach of Cai and Betensky (Cont.)

- The spline coefficients $b$ are modeled as random effects with the joint distribution $\text{Normal}(0, \sigma_b^2 I)$. The joint likelihood is proportional to

$$\ell_p(\Theta) = \ell(\Theta) - \frac{K}{2} \log(\sigma_b^2) - \frac{1}{2\sigma_b^2} b^T b. \quad (11)$$

$\Theta = (\theta^T, a^T, b^T)^T$ is estimated by maximizing the penalized likelihood (11).

- The tuning parameter $\sigma_b^2$ is chosen by maximizing the marginal likelihood using Laplace approximation (Breslow and Clayton, 1993).

- When $X$ is measured with error, SUBEX method is applicable.
Simulation 1: Linear model

Let $N_i(\cdot)$ be a stationary point process with intensity $X_i$, where $X_i \sim 1 + \text{Log Normal}(0, 0.5)$; $Z_i$ is an error-free Bernoulli random variable independent of $X_i$, with $P(Z_i = 1) = 0.5$.

$Y_i = \theta_0 + \theta_1 X_i + \theta_2 Z_i + \epsilon_i, \quad i = 1, 2, \cdots, n$, where $\theta = (\theta_0, \theta_1, \theta_2)^T = (1, 1, 1)^T$ and $\epsilon_i \sim \text{Normal}(0, 0.5^2)$. 
Simulation 1: Linear model

- Let $N_i(\cdot)$ be a stationary point process with intensity $X_i$, where $X_i \sim 1 + \text{Log Normal}(0, 0.5)$; $Z_i$ is an error-free Bernoulli random variable independent of $X_i$, with $P(Z_i = 1) = 0.5$.
- $Y_i = \theta_0 + \theta_1 X_i + \theta_2 Z_i + \epsilon_i$, $i = 1, 2, \ldots, n$, where $\theta = (\theta_0, \theta_1, \theta_2)^T = (1, 1, 1)^T$ and $\epsilon_i \sim \text{Normal}(0, 0.5^2)$.
- We assume that $X_i$ is unknown but can be estimated by $W_i = N_i((0, \tau]) / \tau$, where $N_i((0, \tau])$ denotes the number of events of $N_i$ contained in the time interval $(0, \tau]$.
Simulation 1: Linear model (Cont.)

We consider two scenarios for $N_i$.

- **Scenario 1:** $N_i$ is a homogeneous Poisson process with intensity $X_i$.

- **Scenario 2:** $N_i$ is a homogeneous Poisson cluster process: we generate a homogeneous Poisson process with intensity $\rho_i = X_i/3$ as the parent process; each parent generates $m = 3$ children on average according to a Poisson distribution, and disperse the children locations independently following a normal distribution centered at the parent location and with standard deviation $\omega = 2$. The final process contains only the children event times.
Simulation 1 (Cont.)

- We set $n = 200$ and repeat the simulation 200 times in each case.
- We apply the naive estimator, the empirical SIMEX method, two versions of Method of Moment estimator and two versions of SUBEX ($\text{MOM}_1$ and $\text{SUBEX}_1$ ignore within-trajectory correlation, $\text{MOM}_2$ and $\text{SUBEX}_2$ take into account of the correlation).
- For SUBEX, we set the $p$ values to be from 0.6 to 0.95 with an increment of 0.05, and use a quadratic function for the extrapolation.
## Simulation results for linear model

### Scenario 1: Poisson process, $\tau = 10$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\theta_1$ Bias</th>
<th>$\theta_1$ SE</th>
<th>Rel. Eff.</th>
<th>$\theta_2$ Bias</th>
<th>$\theta_2$ SE</th>
<th>Rel. Eff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>-.3710</td>
<td>.0686</td>
<td></td>
<td>.0075</td>
<td>.0893</td>
<td></td>
</tr>
<tr>
<td>MOM$_1$</td>
<td>.0134</td>
<td>.1590</td>
<td>5.6190</td>
<td>.0086</td>
<td>.0973</td>
<td>.8406</td>
</tr>
<tr>
<td>MOM$_2$</td>
<td>.0163</td>
<td>.1765</td>
<td>4.5517</td>
<td>.0086</td>
<td>.0975</td>
<td>.8376</td>
</tr>
<tr>
<td>SUBEX$_1$</td>
<td>-.0999</td>
<td>.1317 (.1334)</td>
<td>5.2228</td>
<td>.0052</td>
<td>.1255 (.1308)</td>
<td>.5084</td>
</tr>
<tr>
<td>SUBEX$_2$</td>
<td>-.1073</td>
<td>.1346 (.1317)</td>
<td>4.8206</td>
<td>.0063</td>
<td>.1246 (.1297)</td>
<td>.5158</td>
</tr>
<tr>
<td>SIMEX</td>
<td>-.1348</td>
<td>.0989 (.0896)</td>
<td>5.0983</td>
<td>.0075</td>
<td>.0929 (.0913)</td>
<td>.9237</td>
</tr>
</tbody>
</table>

### Scenario 2: Poisson cluster process, $\tau = 10$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\theta_1$, $\tau = 10$ Bias</th>
<th>$\theta_1$, $\tau = 10$ SE</th>
<th>Rel. Eff.</th>
<th>$\theta_2$, $\tau = 20$ Bias</th>
<th>$\theta_2$, $\tau = 20$ SE</th>
<th>Rel. Eff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>-.6678</td>
<td>.0616</td>
<td></td>
<td>-.5163</td>
<td>.0670</td>
<td></td>
</tr>
<tr>
<td>MOM$_1$</td>
<td>-.4205</td>
<td>.1106</td>
<td>2.3798</td>
<td>-.1388</td>
<td>.1519</td>
<td>6.4177</td>
</tr>
<tr>
<td>MOM$_2$</td>
<td>-.1869</td>
<td>.2974</td>
<td>3.6586</td>
<td>-.1014</td>
<td>.1551</td>
<td>7.9226</td>
</tr>
<tr>
<td>SUBEX$_1$</td>
<td>-.4417</td>
<td>.1245 (.1227)</td>
<td>2.1361</td>
<td>-.2263</td>
<td>.1508 (.1463)</td>
<td>3.6700</td>
</tr>
<tr>
<td>SUBEX$_2$</td>
<td>-.4009</td>
<td>.1544 (.1556)</td>
<td>2.4387</td>
<td>-.1997</td>
<td>.1750 (.1692)</td>
<td>3.8524</td>
</tr>
<tr>
<td>SIMEX</td>
<td>-.5189</td>
<td>.0831 (.0802)</td>
<td>1.6292</td>
<td>-.2959</td>
<td>.0958 (.0913)</td>
<td>2.8029</td>
</tr>
</tbody>
</table>
Simulation 2: interval-censored failure time data

We simulate $X_i$, $N_i$ and $Z_i$ as in Simulation 1, and simulate the failure time $T_i$ through a Cox model

$$
\lambda(t|X_i, Z_i) = \lambda_0(t) \exp(X_i \beta + Z_i \eta),
$$

where $\lambda_0(t) = t$ and $(\beta, \eta)^T = (1, 1)^T$. 

We assume censoring at random and set the censoring times to be $(0.2, 0.5, 1)$. Let the censoring indicator $\delta_i$ be a binary variable independent of $X_i$ and $Z_i$, with $P(\delta_i = 1) = 0.5$. When $\delta_i = 1$, the event time $T_i$ is censored in the interval between the two censoring times closest to $T_i$; if $T_i$ is less than 0.2, it is censored in $[T_l_i = 0, T_r_i = 0.2]$. If an event time is over 1, it is automatically right censored at 1.

Overall, about 12% of the observations are right censored, 43% are interval censored, and the rest 45% are observed.
Simulation 2: interval-censored failure time data

- We simulate $X_i$, $N_i$ and $Z_i$ as in Simulation 1, and simulate the failure time $T_i$ through a Cox model
  \[
  \lambda(t|X_i, Z_i) = \lambda_0(t) \exp(X_i \beta + Z_i \eta),
  \]
  where $\lambda_0(t) = t$ and $(\beta, \eta)^T = (1, 1)^T$.

- We assume censoring at random and set the censoring times to be $(0.2, 0.5, 1)$. Let the censoring indicator $\delta_i$ be a binary variable independent of $X_i$ and $Z_i$, with $P(\delta_i = 1) = 0.5$. When $\delta_i = 1$, the event time $T_i$ is censored in the interval between the two censoring times closest to $T_i$; if $T_i$ is less than 0.2, it is censored in $[T_i^l = 0, T_i^r = 0.2]$; if an event time is over 1, it is automatically right censored at 1.
Simulation 2: interval-censored failure time data

- We simulate $X_i$, $N_i$ and $Z_i$ as in Simulation 1, and simulate the failure time $T_i$ through a Cox model
  \[
  \lambda(t|X_i, Z_i) = \lambda_0(t) \exp(X_i\beta + Z_i\eta),
  \]
  where $\lambda_0(t) = t$ and $(\beta, \eta)^T = (1, 1)^T$.

- We assume censoring at random and set the censoring times to be $(0.2, 0.5, 1)$. Let the censoring indicator $\delta_i$ be a binary variable independent of $X_i$ and $Z_i$, with $P(\delta_i = 1) = 0.5$. When $\delta_i = 1$, the event time $T_i$ is censored in the interval between the two censoring times closest to $T_i$; if $T_i$ is less than 0.2, it is censored in $[T_i^l = 0, T_i^r = 0.2]$; if an event time is over 1, it is automatically right censored at 1.

- Overall, about 12% of the observations are right censored, 43% are interval censored, and the rest 45% are observed.
Simulation results in the interval-censored failure time data

<table>
<thead>
<tr>
<th>Poisson, $\tau = 10$, $n = 200$</th>
<th>Poisson cluster, $\tau = 10$, $n = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bias</strong></td>
<td><strong>SE</strong></td>
</tr>
<tr>
<td>No Error</td>
<td>.0103</td>
</tr>
<tr>
<td>Naive</td>
<td>-4.273</td>
</tr>
<tr>
<td>SUBEX$_1$</td>
<td>-1.424</td>
</tr>
<tr>
<td>SUBEX$_2$</td>
<td>-1.486</td>
</tr>
<tr>
<td>SIMEX</td>
<td>-1.674</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Poisson cluster, $\tau = 20$, $n = 200$</th>
<th>Poisson cluster, $\tau = 20$, $n = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bias</strong></td>
<td><strong>SE</strong></td>
</tr>
<tr>
<td>No Error</td>
<td>.0186</td>
</tr>
<tr>
<td>Naive</td>
<td>-5.825</td>
</tr>
<tr>
<td>SUBEX$_1$</td>
<td>-2.967</td>
</tr>
<tr>
<td>SUBEX$_2$</td>
<td>-2.698</td>
</tr>
<tr>
<td>SIMEX</td>
<td>-3.750</td>
</tr>
</tbody>
</table>
## Results for the analysis of the CCQ-Brief score data

<table>
<thead>
<tr>
<th></th>
<th>NAIVE</th>
<th>MOM&lt;sub&gt;1&lt;/sub&gt;</th>
<th>SIMEX</th>
<th>MOM&lt;sub&gt;2&lt;/sub&gt;</th>
<th>SUBEX&lt;sub&gt;1&lt;/sub&gt;</th>
<th>SUBEX&lt;sub&gt;2&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>.476 (.176)</td>
<td>.529 (.195)</td>
<td>.537 (.223)</td>
<td>.565 (.211)</td>
<td>.577 (.204)</td>
<td>.662 (.257)</td>
</tr>
<tr>
<td>indicator</td>
<td>-.046 (.062)</td>
<td>-.044 (.062)</td>
<td>-.049 (.064)</td>
<td>-.043 (.062)</td>
<td>-.040 (.064)</td>
<td>-.033 (.065)</td>
</tr>
<tr>
<td>gender</td>
<td>-.108 (.093)</td>
<td>-.107 (.093)</td>
<td>-.112 (.091)</td>
<td>-.106 (.093)</td>
<td>-.096 (.092)</td>
<td>-.083 (.095)</td>
</tr>
<tr>
<td>race</td>
<td>.305 (.099)</td>
<td>.306 (.100)</td>
<td>.311 (.101)</td>
<td>.307 (.100)</td>
<td>.289 (.100)</td>
<td>.271 (.103)</td>
</tr>
<tr>
<td>age</td>
<td>-.088 (.081)</td>
<td>-.096 (.082)</td>
<td>-.087 (.081)</td>
<td>-.101 (.083)</td>
<td>-.095 (.081)</td>
<td>-.101 (.083)</td>
</tr>
<tr>
<td>cocyrs</td>
<td>-.013 (.087)</td>
<td>-.011 (.087)</td>
<td>-.019 (.085)</td>
<td>-.010 (.088)</td>
<td>-.015 (.083)</td>
<td>-.013 (.086)</td>
</tr>
<tr>
<td>curanxs</td>
<td>.139 (.085)</td>
<td>.142 (.085)</td>
<td>.134 (.079)</td>
<td>.145 (.086)</td>
<td>.163 (.084)</td>
<td>.178 (.086)</td>
</tr>
</tbody>
</table>
Results for the analysis of the time to first relapse data

<table>
<thead>
<tr>
<th></th>
<th>NAIVE</th>
<th>SIMEX</th>
<th>SUBEX₁</th>
<th>SUBEX₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>cocuse</td>
<td>-.203 (.081)</td>
<td>-.210 (.096)</td>
<td>-.097 (.095)</td>
<td>.017 (.206)</td>
</tr>
<tr>
<td>gender</td>
<td>-.377 (.299)</td>
<td>-.387 (.301)</td>
<td>-.351 (.319)</td>
<td>-.301 (.347)</td>
</tr>
<tr>
<td>race</td>
<td>-.196 (.274)</td>
<td>-.182 (.277)</td>
<td>-.187 (.272)</td>
<td>-.162 (.276)</td>
</tr>
<tr>
<td>age</td>
<td>-.053 (.024)</td>
<td>-.053 (.024)</td>
<td>-.054 (.024)</td>
<td>-.056 (.024)</td>
</tr>
<tr>
<td>cocyrs</td>
<td>.110 (.028)</td>
<td>.111 (.028)</td>
<td>.111 (.028)</td>
<td>.109 (.029)</td>
</tr>
<tr>
<td>curanxs</td>
<td>.279 (.221)</td>
<td>.275 (.224)</td>
<td>.308 (.218)</td>
<td>.352 (.232)</td>
</tr>
</tbody>
</table>
In many scientific problems in particular substance use research, it is common to have summary statistics derived from some stochastic processes as covariates. The estimation error in these summary statistics causes estimation bias in the regression coefficients like in classical measurement error problems.

We propose a new method-of-moment approach for linear models and a subsampling extrapolation method that is generally applicable to both linear and nonlinear models.

The proposed methods are based on novel subsampling techniques that take into account of the correlation within individual processes, and have shown good performance in both simulation and real data analysis.