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

- 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

- 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

- 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 data

- 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 data

- 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 data

- 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 data

- 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 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.

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

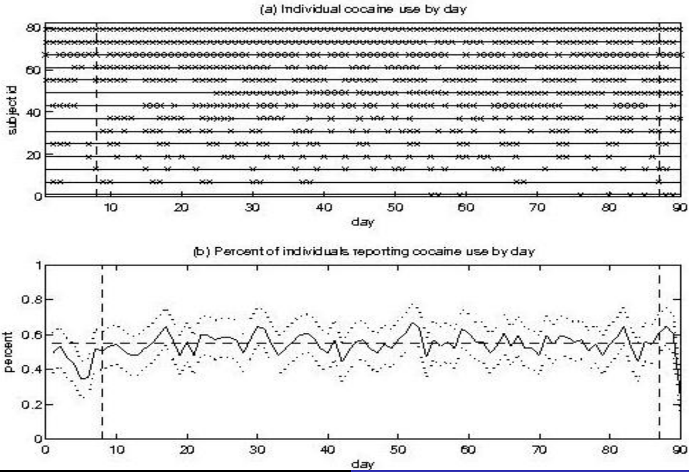
Assessing baseline cocaine use pattern

- 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.

Difference with the classical measurement error problems

- 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 slop 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 ith cocaine use trajectory over the baseline period $(0, \tau]$.
$$W_i = \frac{1}{\tau} \int_0^\tau N_i^*(dt), \tag{1}$$
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 Λ_i, $X_i = \mathrm{E}(W_i | \Lambda_i)$ is the true predictor.
- $W_i = X_i + U_i$, $\sigma_i^2 = \mathrm{var}(U_i)$ might depends on X_i and is different across the subjects.

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 $U_i(t,p) = W_i(t,p) - X_i$.

Notation and setup (Cont.)

- Let $S(t,p) = (t, t+p\tau]$ be a subinterval, where $0 < p \le 1$ and $0 \le t \le (1-p)\tau$, $W_i(t,p)$ 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

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

where α_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 $\tilde{\sigma}_i^2 = \lim_{\tau\to\infty}(\tau\sigma_{u_i}^2)$.
- **Weak Dependency Assumption:** We assume that

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

  where α_i is a constant that is typically nonnegative.
- By the weak dependence assumption,

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

  if $p\tau$ is sufficiently large.

Method of moment bias correction in linear model

- We consider $Y_i = X_i\beta + \mathbf{Z}_i^T\boldsymbol{\eta} + \epsilon_i$, $\boldsymbol{\theta} = (\beta, \boldsymbol{\eta})$.
- The naive estimator using the surrogate W,

$$\hat{\boldsymbol{\theta}}_{\text{naive}} = \begin{bmatrix} \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{W} & \mathbf{Z}^T\mathbf{Z} \end{bmatrix}^{-1} \left[\begin{pmatrix} \mathbf{W}^T \\ \mathbf{Z}^T \end{pmatrix} \mathbf{Y}\right] \equiv \mathbf{A}^{-1}\mathbf{B}.$$

Method of moment bias correction in linear model

- We consider $Y_i = X_i\beta + \mathbf{Z}_i^T\boldsymbol{\eta} + \epsilon_i$, $\boldsymbol{\theta} = (\beta, \boldsymbol{\eta})$.
- The naive estimator using the surrogate W,

$$\hat{\boldsymbol{\theta}}_{\text{naive}} = \begin{bmatrix} \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{W} & \mathbf{Z}^T\mathbf{Z} \end{bmatrix}^{-1} \left[\begin{pmatrix} \mathbf{W}^T \\ \mathbf{Z}^T \end{pmatrix} \mathbf{Y} \right] \equiv \mathbf{A}^{-1}\mathbf{B}.$$

$$\mathrm{E}(\mathbf{A}|\mathbf{X},\mathbf{Z}) = \begin{pmatrix} \mathbf{X}^T\mathbf{X} + \sigma^2 & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} \end{pmatrix}, \quad \mathrm{E}(\mathbf{B}|\mathbf{X},\mathbf{Z}) = \begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} \end{pmatrix} \boldsymbol{\theta},$$

where $\sigma^2 = \mathrm{E}(\mathbf{U}^T\mathbf{U}) = \sum_{i=1}^n \sigma_{u_i}^2$.

Method of moment bias correction in linear model

- We consider $Y_i = X_i\beta + \mathbf{Z}_i^T\boldsymbol{\eta} + \epsilon_i$, $\boldsymbol{\theta} = (\beta, \boldsymbol{\eta})$.
- The naive estimator using the surrogate W,

$$\hat{\boldsymbol{\theta}}_{\text{naive}} = \begin{bmatrix} \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{W} & \mathbf{Z}^T\mathbf{Z} \end{bmatrix}^{-1} \left[\begin{pmatrix} \mathbf{W}^T \\ \mathbf{Z}^T \end{pmatrix} \mathbf{Y} \right] \equiv \mathbf{A}^{-1}\mathbf{B}.$$

$$\mathrm{E}(\mathbf{A}|\mathbf{X},\mathbf{Z}) = \begin{pmatrix} \mathbf{X}^T\mathbf{X} + \sigma^2 & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} \end{pmatrix}, \quad \mathrm{E}(\mathbf{B}|\mathbf{X},\mathbf{Z}) = \begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} \end{pmatrix}\boldsymbol{\theta},$$

where $\sigma^2 = \mathrm{E}(\mathbf{U}^T\mathbf{U}) = \sum_{i=1}^n \sigma_{u_i}^2$.

- The method of moment estimator $\boldsymbol{\theta}_{\text{mom}}$ is obtained by subtracting $\widehat{\sigma}^2$ from the first entry of $\boldsymbol{A}$

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 ith trajectory, where $j = 0, 1, \cdots, k-1$. Define

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

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 ith trajectory, where $j = 0, 1, \cdots, k-1$. Define

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

$$\tilde{\sigma}^2_{u_i}(p) = \frac{1}{\tau - kp\tau} \int_0^{\tau - kp\tau} \tilde{\sigma}^2_{u_i}(t, p) dt,$$

Subsampling estimation of the variance (cont.)

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

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

Subsampling estimation of the variance (cont.)

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

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

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

Subsampling estimation of the variance (cont.)

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

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

- **Scheme 1:** Ignore correlation, and $\widehat{\sigma}^2 = p/(1 - \frac{1}{k})\sum_{i=1}^{n} \tilde{\sigma}^2_{u_i}(p)$.
- **Scheme 2:** Take into account of correlation, let

$$\tilde{Y}(p) = \frac{p\sum_{i=1}^{n} \tilde{\sigma}^2_{u_i}(p)}{1 - 1/k} \equiv \sum_{i=1}^{n} \tilde{Y}_i(p) \text{ and } \tilde{X}(p) = \frac{1 - p - (1 - kp)/k^2}{p\tau^2(1 - 1/k)}. \quad (5)$$

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

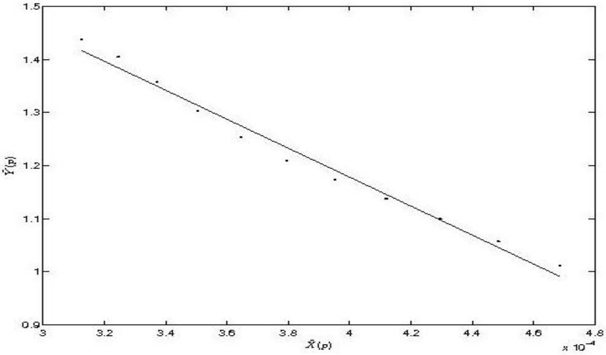


Figure: Scatter plot of $\tilde{X}(p)$ and $\tilde{Y}(p)$ for $p = k/\tau$ where $k = 30, 31, \cdots, 40$ and $\tau = 80$.

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 $S(t,p) = (t, t+p\tau]$,

$$\begin{aligned}\mathrm{Var}[W_i(t,p)|\Lambda_i] &\approx \frac{\sigma^2_{u_i}}{p} - \frac{1-p}{p^2\tau^2}\alpha_i = \left(\frac{1}{p} - \frac{1-p}{p^2\tau^2}\frac{\alpha_i}{\sigma^2_{u_i}}\right)\sigma^2_{u_i} \\ &\equiv x_i(p)\sigma^2_{u_i}. \end{aligned} \quad (6)$$

SUBEX (cont.)

- Let $\hat{\boldsymbol{\theta}}(p; t_1, \cdots, t_n)$ be the estimated regression coefficient based on $\{Y_i, W_i(t_i, p), \mathbf{Z}_i\}$, where $t_i \in [0, (1-p)\tau]$, $i = 1, \cdots, n$, without accounting for measurement error.

$$\hat{\boldsymbol{\theta}}(p) = \frac{1}{[(1-p)\tau]^n} \int_0^{(1-p)\tau} \cdots \int_0^{(1-p)\tau} \hat{\boldsymbol{\theta}}(p; t_1, \cdots, t_n) dt_1 \cdots dt_n. \tag{7}$$

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 $(\hat{x}(p), \hat{\boldsymbol{\theta}}(p))$, e.g.

$$\mathcal{G}_Q(p, \boldsymbol{\Gamma}) = \gamma_1 + \hat{x}(p)\gamma_2 + \hat{x}(p)^2\gamma_3,$$

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

- $\hat{\boldsymbol{\theta}}_{\text{subex}} = \mathcal{G}_Q(p = \infty, \hat{\boldsymbol{\Gamma}})$.

SUBEX (cont.)

- 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 $\tilde{Y}(p)$ and $\tilde{X}(p)$.

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 σ^2 by

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

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 σ^2 by

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

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

$$W_i(\zeta) = W_i + \frac{\sqrt{\zeta}}{2}\left[W_i(0,1/2) - W_i(\tau/2,1/2)\right], \text{ for each } \zeta > 0.$$

Interval censored relapse time

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

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

Interval censored relapse time

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

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

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

Interval censored relapse time

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

$$\lambda(t|X_i, Z_i) = \lambda_0(t) \exp(X_i\beta + Z_i\eta). \quad (8)$$

- 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),

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

where $x_+ \equiv \max(x, 0)$, and κ_k's are the knots.

Penalized spline approach of Cai and Betensky (2003)

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

Penalized spline approach of Cai and Betensky (2003)

- We observe n independent tuples, (T_i^l, T_i^r, δ_i), where $[T_i^l, T_i^r]$ gives the censoring interval, δ_i is the indicator for right censoring. When $\delta_i = 0$ and $T_i^l = T_i^r$, the event time T_i is right censored at T_i^r; when $\delta_i = 1$ and $T_i^l < T_i^r$, T_i is interval censored within $[T_i^l, T_i^r]$; when $\delta_i = 1$ and $T_i^l = T_i^r$, T_i is observed at T_i^r. In addition, let δ_{0i} be the indicator for observed data, i.e. $\delta_{0i} = I(\delta_i = 1, T_i^l = T_i^r)$.
- Denote $\mathbb{X} = (X^T, Z^T)^T$ and $\Lambda_0(t) = \int_0^t \lambda_0(u)du$.

$$\begin{aligned}\ell(\Theta) &= \sum_{i=1}^{n} \delta_{0i}\{\log \lambda_0(T_i^r) + \mathbb{X}_i^T\theta\} - \exp(\mathbb{X}_i^T\theta)\Lambda_0(T_i^r) \\ &\quad + \delta_i(1-\delta_{0i}) \log\left(\exp[\{\Lambda_0(T_i^r) - \Lambda_0(T_i^l)\}\exp(\mathbb{X}_i^T\theta)] - 1\right), \quad (10)\end{aligned}$$

where Θ is the collection of all parameters.

Penalized spline approach of Cai and Betensky (Cont.)

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

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

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

- The tuning parameter σ_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 $X_i \sim 1+$ Log Normal $(0, 0.5)$; Z_i is an error-free Bernoulli random variable independent of X_i, with $P(Z_i = 1) = 0.5$.
- $Y_i = \theta_0 + \theta_1 X_i + \theta_2 Z_i + \epsilon_i$, $i = 1, 2, \cdots, n$, where $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)^T = (1, 1, 1)^T$ 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 $X_i \sim 1+$ Log Normal $(0, 0.5)$; Z_i is an error-free Bernoulli random variable independent of X_i, with $P(Z_i = 1) = 0.5$.
- $Y_i = \theta_0 + \theta_1 X_i + \theta_2 Z_i + \epsilon_i$, $i = 1, 2, \cdots, n$, where $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)^T = (1, 1, 1)^T$ 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 (MOM_1 and SUBEX_1 ignore within-trajectory correlation, MOM_2 and 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, $\tau = 10$						
		θ_1			θ_2	
	Bias	SE	Rel. Eff.	Bias	SE	Rel. Eff.
Naive	-.3710	.0686		.0075	.0893	
MOM_1	.0134	.1590	5.6190	.0086	.0973	.8406
MOM_2	.0163	.1765	4.5517	.0086	.0975	.8376
SUBEX_1	-.0999	.1317 (.1334)	5.2228	.0052	.1255 (.1308)	.5084
SUBEX_2	-.1073	.1346 (.1317)	4.8206	.0063	.1246 (.1297)	.5158
SIMEX	-.1348	.0989 (.0896)	5.0983	.0075	.0929 (.0913)	.9237
Scenario 2: Poisson cluster process						
		θ_1, $\tau = 10$			θ_1, $\tau = 20$	
	Bias	SE	Rel. Eff.	Bias	SE	Rel. Eff.
Naive	-.6678	.0616		-.5163	.0670	
MOM_1	-.4205	.1106	2.3798	-.1388	.1519	6.4177
MOM_2	-.1869	.2974	3.6586	-.1014	.1551	7.9226
SUBEX_1	-.4417	.1245 (.1227)	2.1361	-.2263	.1508 (.1463)	3.6700
SUBEX_2	-.4009	.1544 (.1556)	2.4387	-.1997	.1750 (.1692)	3.8524
SIMEX	-.5189	.0831 (.0802)	1.6292	-.2959	.0958 (.0913)	2.8029

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

$$\lambda(t|X_i, Z_i) = \lambda_0(t)\exp(X_i\beta + Z_i\eta), \tag{12}$$

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

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

$$\lambda(t|X_i, Z_i) = \lambda_0(t)\exp(X_i\beta + Z_i\eta), \tag{12}$$

  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 δ_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
$$\lambda(t|X_i, Z_i) = \lambda_0(t) \exp(X_i\beta + Z_i\eta), \tag{12}$$
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 δ_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

	Poisson, $\tau = 10$, $n = 200$			Poisson cluster, $\tau = 10$, $n = 200$		
	Bias	SE	Rel. Eff.	Bias	SE	Rel. Eff.
No Error	.0103	.1471 (.1469)		.0053	.1480 (.1462)	
Naive	-.4273	.1168 (.1129)		-.7225	.0890 (.0800)	
$SUBEX_1$	-.1424	.2787 (.2587)	2.0113	-.5073	.2176 (.2016)	1.7402
$SUBEX_2$	-.1486	.2789 (.2570)	1.9709	-.4687	.2768 (.2590)	1.7908
SIMEX	-.1674	.1830 (.1732)	3.1990	-.5931	.1360 (.1237)	1.4318

	Poisson cluster, $\tau = 20$, $n = 200$			Poisson cluster, $\tau = 20$, $n = 400$		
	Bias	SE	Rel. Eff.	Bias	SE	Rel. Eff.
No Error	.0186	.1545 (.1465)		.0081	.1041 (.1015)	
Naive	-.5825	.1067 (.0971)		-.5923	.0730 (.0695)	
$SUBEX_1$	-.2967	.2717 (.2602)	2.1715	-.3007	.2001 (.1843)	2.7329
$SUBEX_2$	-.2698	.3223 (.3052)	1.9915	-.2713	.2351 (.2178)	2.7691
SIMEX	-.3750	.1712 (.1558)	2.0654	-.3871	.1193 (.1107)	2.1713

Results for the analysis of the CCQ-Brief score data

	NAIVE	MOM_1	SIMEX	MOM_2	$SUBEX_1$	$SUBEX_2$
frequency	.476 (.176)	.529 (.195)	.537 (.223)	.565 (.211)	.577 (.204)	.662 (.257)
indicator	-.046 (.062)	-.044 (.062)	-.049 (.064)	-.043 (.062)	-.040 (.064)	-.033 (.065)
gender	-.108 (.093)	-.107 (.093)	-.112 (.091)	-.106 (.093)	-.096 (.092)	-.083 (.095)
race	.305 (.099)	.306 (.100)	.311 (.101)	.307 (.100)	.289 (.100)	.271 (.103)
age	-.088 (.081)	-.096 (.082)	-.087 (.081)	-.101 (.083)	-.095 (.081)	-.101 (.083)
cocyrs	-.013 (.087)	-.011 (.087)	-.019 (.085)	-.010 (.088)	-.015 (.083)	-.013 (.086)
curanxs	.139 (.085)	.142 (.085)	.134 (.079)	.145 (.086)	.163 (.084)	.178 (.086)

Results for the analysis of the time to first relapse data

	NAIVE	SIMEX	SUBEX_1	SUBEX_2
cocuse	-.203 (.081)	-.210 (.096)	-.097 (.095)	.017 (.206)
gender	-.377 (.299)	-.387 (.301)	-.351 (.319)	-.301 (.347)
race	-.196 (.274)	-.182 (.277)	-.187 (.272)	-.162 (.276)
age	-.053 (.024)	-.053 (.024)	-.054 (.024)	-.056 (.024)
cocyrs	.110 (.028)	.111 (.028)	.111 (.028)	.109 (.029)
curanxs	.279 (.221)	.275 (.224)	.308 (.218)	.352 (.232)

Summary

- 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.