Bayesian Analysis of Episodic Drug Abuse Data

New Models and New Inferences

Adam J. King

Ph.D. Candidate, Department of Biostatistics
Graduate Student Researcher, Integrated Substance Abuse Programs
University of California, Los Angeles
aking@ucla.edu

March 21, 2014

Purpose

Many fields of research use and develop statistical methods:

- social scientists often call their statistical tools methodology
- medical researchers call their statistical tools biostatistics

Substance abuse research straddles the social sciences and medicine, and so ideally will draw the best available statistical methods from each field.

- the purpose of my research is to develop and adapt recent advances in biostatistics for substance abuse research
- the purpose of my talk here today is to share some of these with you

Outline

1. Episodic Drug Abuse Data
   - Background
   - Treatment Utilization and Effectiveness Project
2. Statistical Approaches
   - Event History Analysis
   - Semiparametric and Additive Models
   - Bayesian Inference and Computing
3. Cocaine Episode Termination in TUE
   - Software and Model
   - Inferences

Episodic Data

An episode is a time span during which a certain trait, behavior, or circumstance was continually present for a study subject.

Observing repeated episodes of various types is common in many fields:

- Ex: in a study of chronic disease, subjects may experience episodes of disease activity, remission, and treatment
- Ex: in a study of labor markets, subjects may experience episodes of schooling, employment, and unemployment
- Ex: in a study of illicit drug use, subjects may experience episodes of drug use, abstinence, drug treatment, and incarceration

Examining episodes directly, instead of aggregate summaries of episodes, allows us to better observe temporal relationships.

- Ex: we observe a positive association between incarceration and drug use; which is causing which?

The Treatment Utilization and Effectiveness Project (TUE)

TUE was a study of illicit drug users in L.A. County from 1992–1997.

- NIDA funded and conducted by ISAP's predecessor organization (Hser, Anglin, and Longshore were PI's)
- 5,168 subjects were contacted and completed baseline interviews across three sources (jails, STD clinics, and hospital ER's)
- baseline interviews and urinalysis confirmed high drug use prevalence (53% of jail, 9% of STD, 18% of ER subj. had positive cocaine test)

A subsample completed the Natural History Interview (NHI) component.

- NHI subjects had to report use of an illegal drug during previous year
- selection priority given to subjects reporting dependent drug use or with elevated risk of dependent drug use (due to criminal activity) or with elevated risk of HIV infection (due to intravenous drug use)
- 787 selected, 566 ultimately interviewed, 508 available for analysis

The Natural History Interview (NHI)

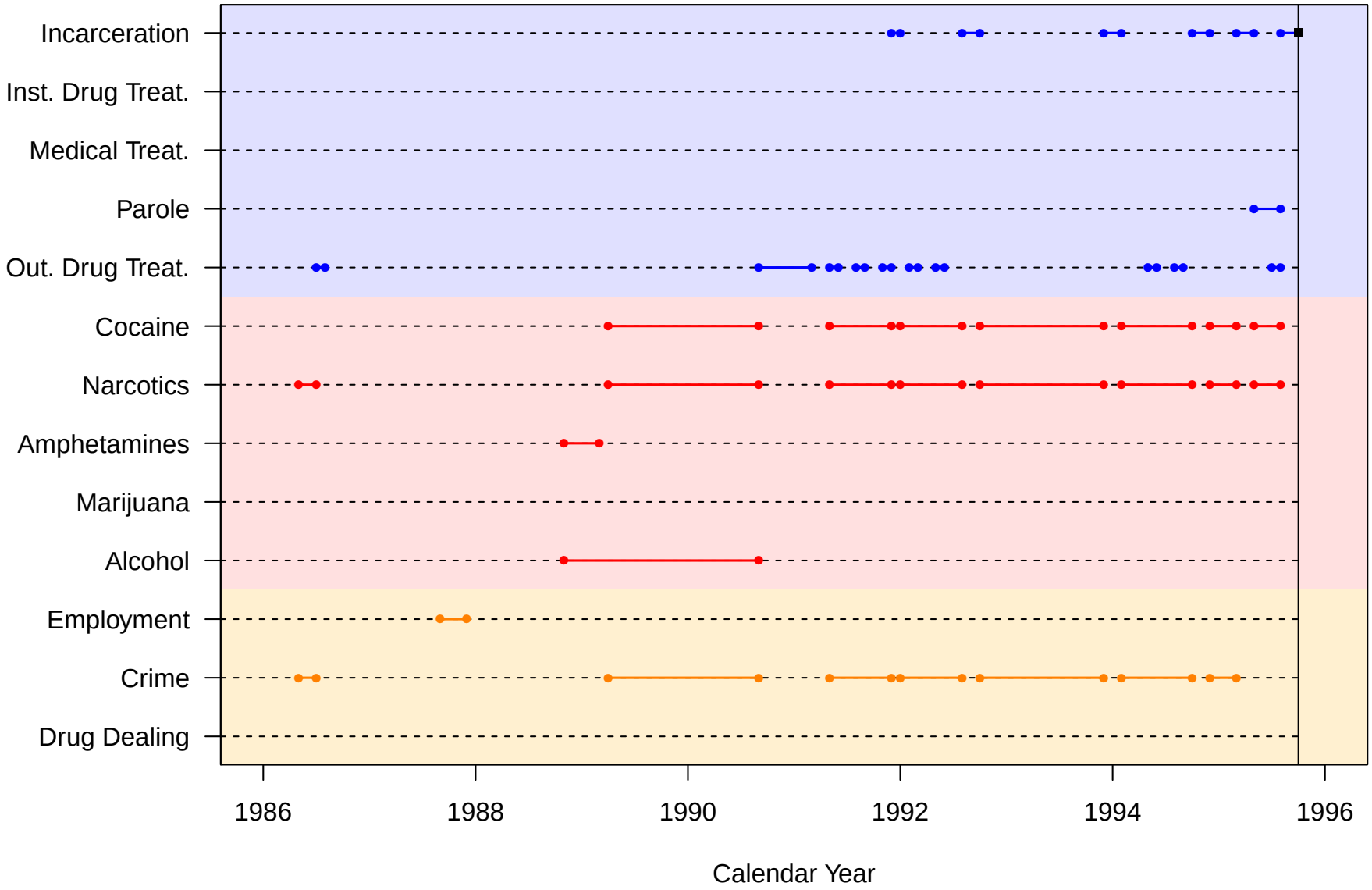
The goal of the NHI is to provide a comprehensive picture of the lifetime drug use, drug treatment, and criminal history of interviewees.

- recorded all lifetime episodes of 13 types of behaviors/circumstances
- episodes recorded as contiguous segments of whole calendar months
- Timeline Follow-back (TLFB) interview process used a timeline of major events (e.g., arrests, marriages, deaths, elections) to aid recall
- TLFB shown to have high test-retest reliability over shorter spans (e.g., daily records of alcohol use over a period of months)
- comparison with records from other data collection processes suggests the NHI is reliable

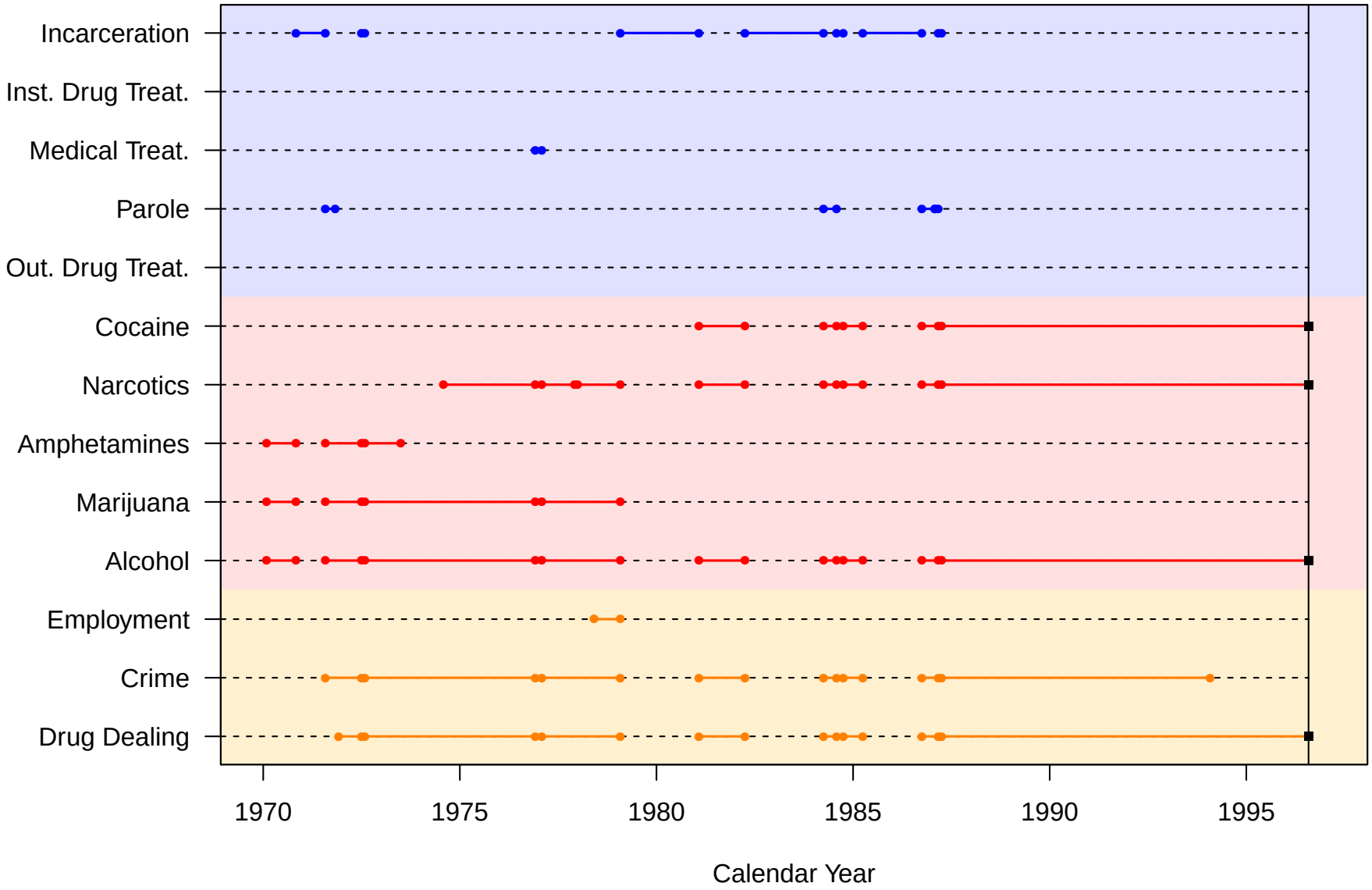
Episode Types and Counts

Episode Type	No. Epis.	No. Subj.	No. of Epis. per Subject			
			Mean	Q2	Q3	Max.
Incarceration	1480	312	4.74	4	6	28
Inst. Drug Treat.	295	168	1.76	1	2	8
Med./Psych. Treat.	58	45	1.29	1	1	5
Legal Status	899	256	3.51	3	5	21
Non-Inst. Drug Treat.	312	153	2.04	1	2	12
Cocaine/Crack	2056	428	4.80	4	7	19
Narcotics	718	155	4.63	2	7	29
Amph./Speed/Crystal	567	223	2.54	1	3	13
Marijuana	1319	421	3.13	2	4	25
Alcohol	1994	477	4.18	3	6	24
Legal Employment	1312	438	3.00	2.5	4	15
Crime	877	226	3.88	2.5	5	28
Dealing	722	258	2.80	2	4	25

Episodic Data from One Subject (age 26 at interview)



Episodic Data from One Subject (age 39 at interview)



Statistical Approaches and Tools

Today, I'm going to focus on three methodologies not in wide use in substance abuse research:

1. Event History Analysis: the study of the timings of events (relevant for episodic data where subjects experience transition events)
2. Semiparametric and Additive Models: which relaxes the assumption in various forms of regression that covariates have linear effects
3. Bayesian Inference and Computing: alternative paradigm of statistics, but for our purposes allows us to build and fit richer statistical models

The Study of the Timings of Events

An event is an occurrence or qualitative change which happens at some specific point in time.

- we are often interested in both whether the event of interest occurs, and if so, when the event occurs
- we want a statistical model that allows us to simultaneously address both questions
  - Ex: how likely is it a subject resumes cocaine use following treatment?
  - Ex: how likely is it a subject resumes use in the next 12 months?
  - Ex: how long do former users remain abstinent on average?

Event time data is common in many fields, and so has many names:

- survival analysis in biostatistics
- event history analysis in the social sciences
- mortality analysis in demography
- failure time analysis or reliability theory in engineering

Features of Event Histories from Episodic Data

In studies of episodic data, the events of interest are the occurrences which begin or end episodes.

- there may be competing risks, which are distinct ways or reasons that an episode begins or ends
  - Ex: an episode of cocaine use may end because the subject voluntarily stops use or because he was arrested
- there may be recurrent events, which are events of interest that subjects may experience multiple times
  - Ex: if subjects are followed for several years, they may repeatedly cease and resume cocaine use
- there may be multiple types of events, since the events of beginning and ending episodes are different (called a multistate model)
  - Ex: may simultaneously model risks of starting and stopping cocaine use

Cox Proportional Hazards Regression Model

The hazard rate $\lambda(t)$ at time t is the probability that the event of interest occurs at t given that the event has not already occurred prior to t.

- when time t is continuous, we need to define $\lambda(t)$ using a limit, since the probability the event occurs at exactly $t = 5.5628271\ldots$ is zero
- often called the instantaneous risk, since it's the chances the event will occur in the next instant

The Cox proportional hazards model relates the the hazard rate $\lambda(t)$ to time t and a collection of predictor variables $X_1, X_2, \ldots X_M$ as follows:

$$\lambda(t) \equiv \exp(\eta(t)), \qquad \eta(t) \equiv \beta_0(t) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_M X_M,$$

where $\beta_0(t)$ is an arbitrary function of time t and each term $\beta_m X_m$ is a linear function of the predictor variable X_m.

Time Scales

What exactly is time t?

- usually, t is time-at-risk (also called gap time or simply duration), and is the time elapsed between the moment the subject first becomes at risk of event occurrence and the current time point in question
  - Ex: if the event of interest is ceasing cocaine use, then t denotes how long the current episode of cocaine use has lasted up to that point
- other notions of time may also have strong relationships with hazard
  - Ex: current age, current calendar time period, birth cohort, time since last treatment, time since first use of the drug, age at first use of the drug

Why does time t deserve special status in our model?

- maybe it doesn't, but the standard method of fitting the Cox model only allows us to include one single arbitrary function $\beta_0(t)$ at a time

Semiparametric and Additive Models

Nonlinear relationships between predictor variables and response variables are common in many types of regression (linear, logistic, Poisson, Cox).

An additive model replaces the linear predictor with an additive predictor:

$$\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_M X_M \quad \longrightarrow \quad \beta_1(X_1) + \beta_2(X_2) + \cdots + \beta_M(X_M)$$

where each $\beta_m(\cdot)$ is an arbitrary function of X_m, not a single coefficient.

We need an estimation procedure which can fit additive models for survival time response variables with various extra features:

- time-varying covariates
- competing risks
- recurrent events

Bayesian Inference

There are three components of Bayesian inference:

1. a prior probability distribution $p(\beta)$ for our parameters β capturing our knowledge about these parameters prior to looking at the data
   - Ex: if we believe before seeing the data that the parameter β_5 is between 1 and 3, then we might use the prior $\beta_5 \sim N(2, 0.5^2)$
   - Ex: if we believe before seeing the data that $\beta_5(\cdot)$ is a smooth function of X_5, then we might place a prior on $\beta_5(\cdot)$ that ensures its smoothness
2. a likelihood or model $p(y|\beta)$ for how data y depend on parameters β
   - Ex: if $y =$ height and $x =$ weight, we may have $p(y|\beta) = N(\beta_0 + \beta_1 x, \sigma^2)$
3. a posterior distribution $p(\beta|y)$ capturing our knowledge about β after examining the data, which we may calculate using Bayes' Theorem:

$$p(\beta|y) = \frac{p(y|\beta)p(\beta)}{\int p(y|\beta)p(\beta)d\beta}$$

Inferences from Monte Carlo Simulation

Bayes' Theorem gives us a formula for calculating $p(\beta|y)$, but several problems stand in the way of obtaining practical inferences from $p(\beta|y)$.

- the integral $\int p(y|\beta)p(\beta)d\beta$ in the denominator may be hard to calculate, especially if the dimension of β is high
- even with a closed form expression for $p(\beta|y)$, we may need to do further calculations to make inferences about quantities of interest

  Ex: the quantity $\exp(\beta_{\text{age}}(30) - \beta_{\text{age}}(20))$ is the ratio of the risk of event occurrence in 30-year-olds to the risk of occurrence in 20-year-olds

A solution is to use Monte Carlo Simulation to draw a sample from $p(\beta|y)$.

- once we have a large enough sample $\{\beta^{(1)}, \ldots, \beta^{(S)}\}$ from $p(\beta|y)$, it's easy to make inferences about almost any aspect of the posterior

  Ex: we can estimate the quantity $\exp(\beta_{\text{age}}(30) - \beta_{\text{age}}(20))$ by calculating

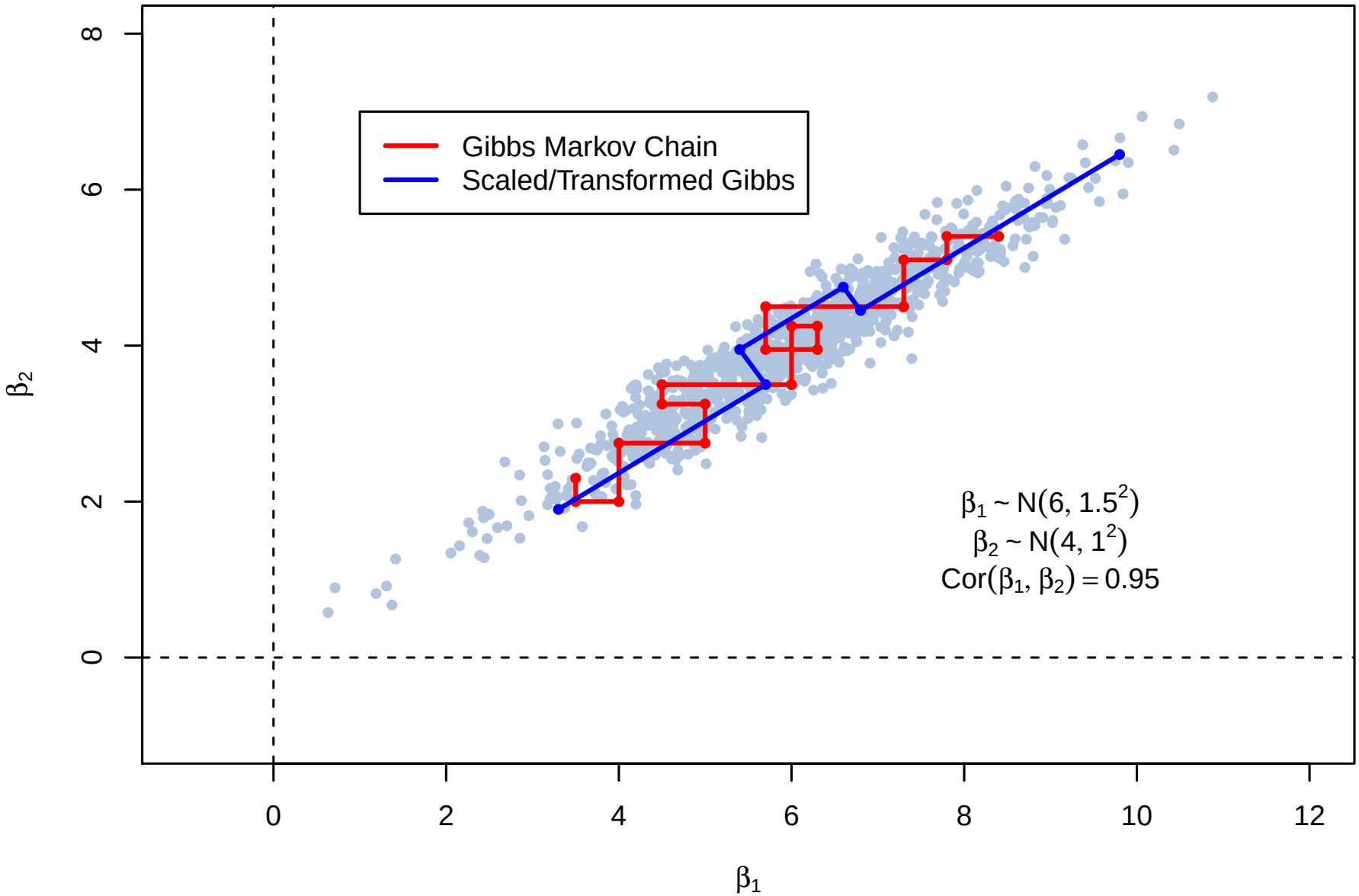
$$\frac{1}{S}\sum_{s=1}^{S}\exp(\beta_{\text{age}}^{(s)}(30) - \beta_{\text{age}}^{(s)}(20))$$

Markov Chain Monte Carlo (MCMC) Simulation

The most common way to simulate from $p(\beta|y)$ is with a Markov Chain.

- a sequence of random values $\{\beta^{(1)}, \beta^{(2)}, \ldots\}$ where each link $\beta^{(s)}$ in the chain is only determined by the previous link $\beta^{(s-1)}$ is Markov
- the elements $\beta^{(s)}$ of the chain come from the posterior distribution, but they are not a simple random sample, because they are correlated
- more highly correlated samples provide less information about $p(\beta|y)$, since they expore the distribution more slowly (they have slow mixing)
- we can reduce the correlation in the chain and speed up mixing by tuning our MCMC algorithm to the shape and scale of the $p(\beta|y)$

MCMC Example with $p(\beta_1, \beta_2|y)$ Multivariate Normal



R Software Package

I implemented my models and algorithms in a collection of R functions, which will be published on the CRAN R software repository this spring.

The user passes the following dataframes and variables to the functions:

1. dataframes of covariates at the subject, episode, or person-time level (person-time level specification allows time-varying covariates)
2. prior-type specifiers for how each covariate is to be incorporated (choices: independent/categorical, autoregressive, random effects)
3. episode-level study time and censoring/competing risks indicators

Example Dataset Characteristics

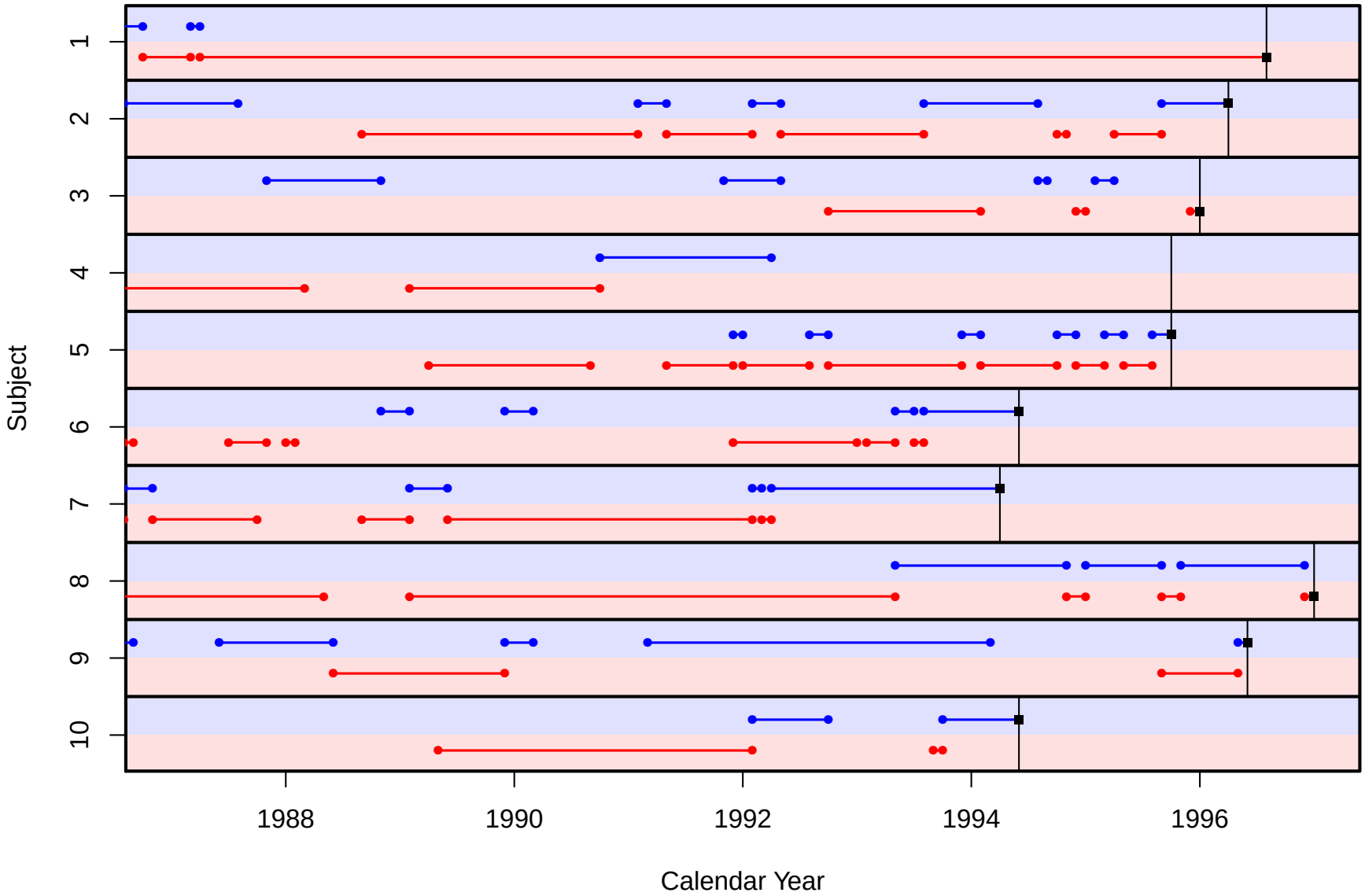
We analyze termination event risk in recurrent episodes of cocaine use with two competing risks (incarceration and voluntary use cessation).

- 408 subjects with ≥ 1 cocaine use episode (plus a few other criteria)
- 1527 cocaine use episodes:
  - 100 right-censored by interview time
  - 689 ended because the subject was incarcerated
  - 738 ended because the subject voluntarily stopped use
- average episode lasts 19.4 months

We include six predictor variables plus correlated subject random effects.

- time-at-risk, age, and calendar time included using smooth functions
- number of previous cocaine episodes, sex, and race are categorical

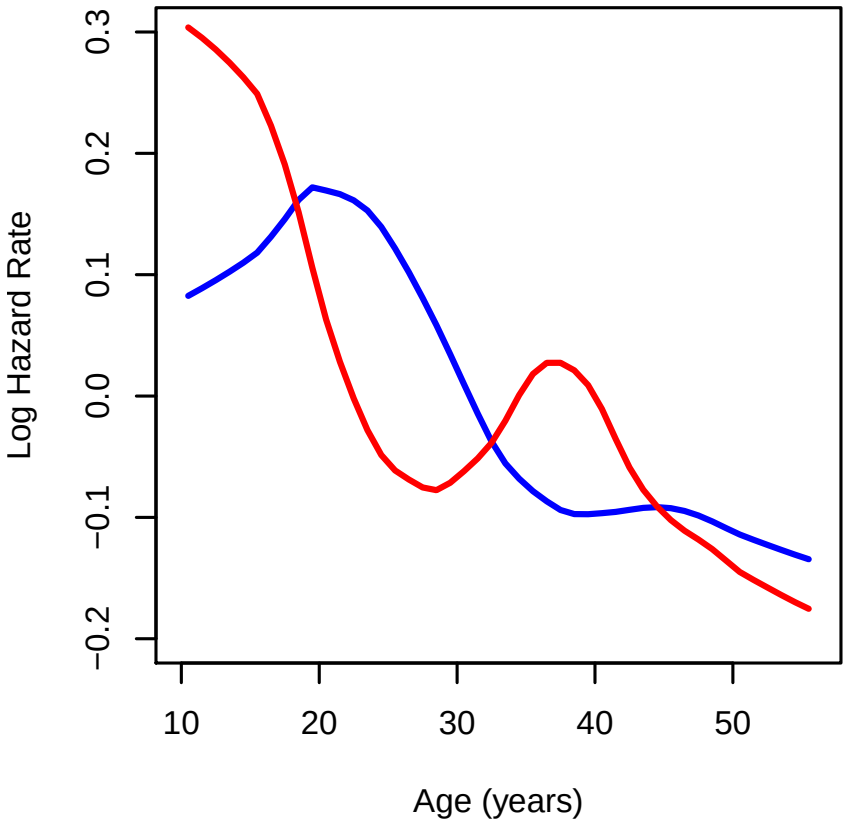
10 Years of Data From 10 Subjects

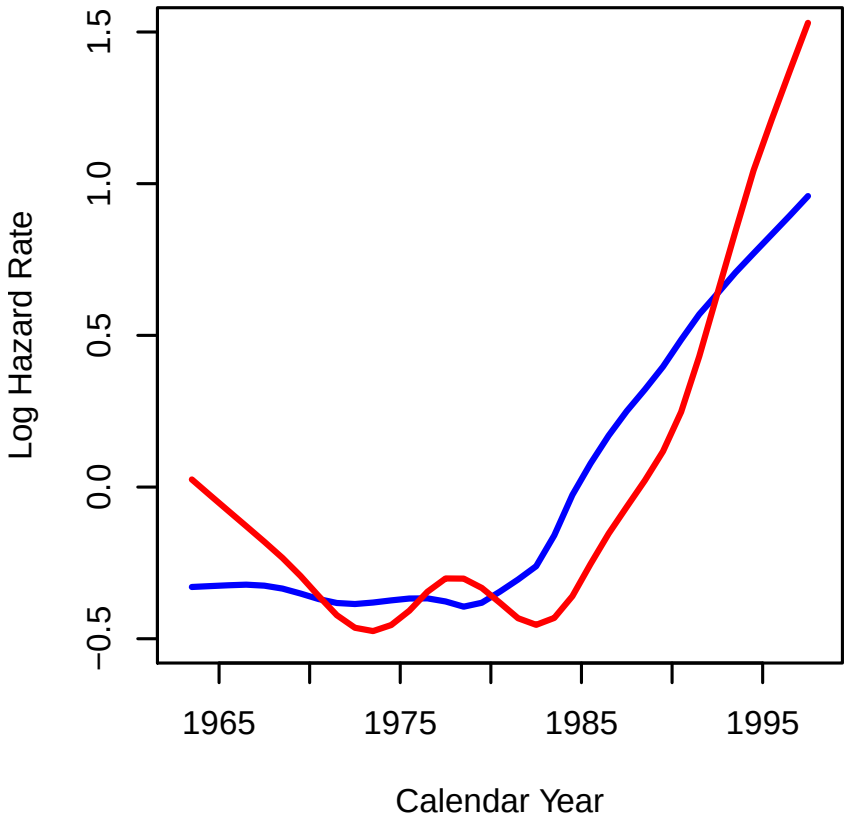


Age and Calendar Time Period Inferences

Subjects in their early teens have the highest likelihood of quitting voluntarily, while subjects in their early 20's are worst off.

The hazards of both voluntary quitting and becoming incarcerated increase dramatically beginning in the early 1980's.

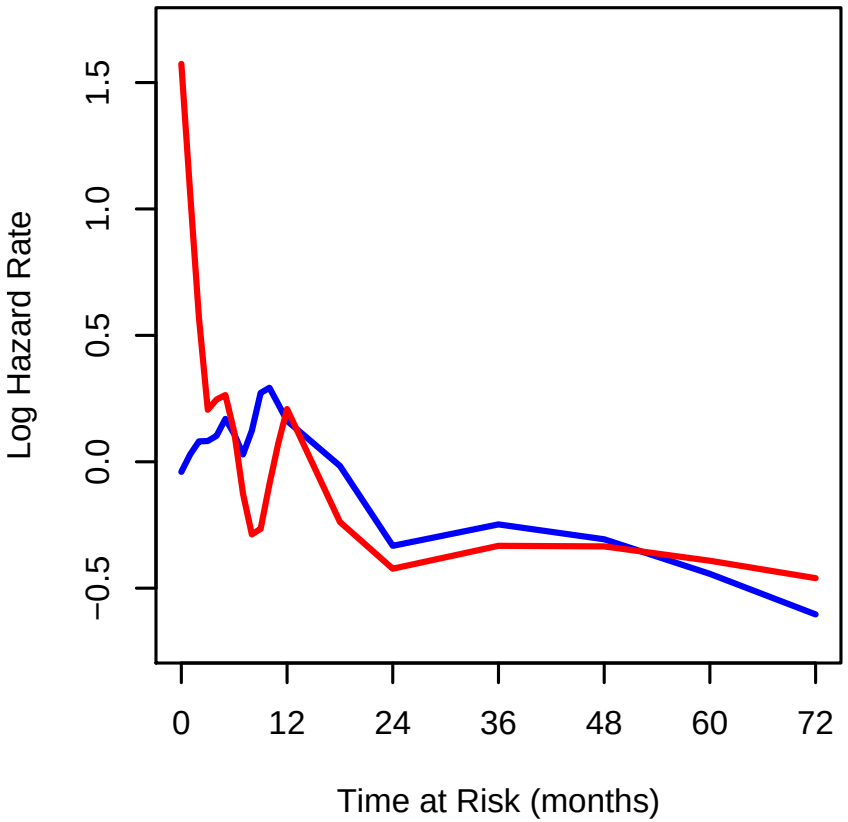


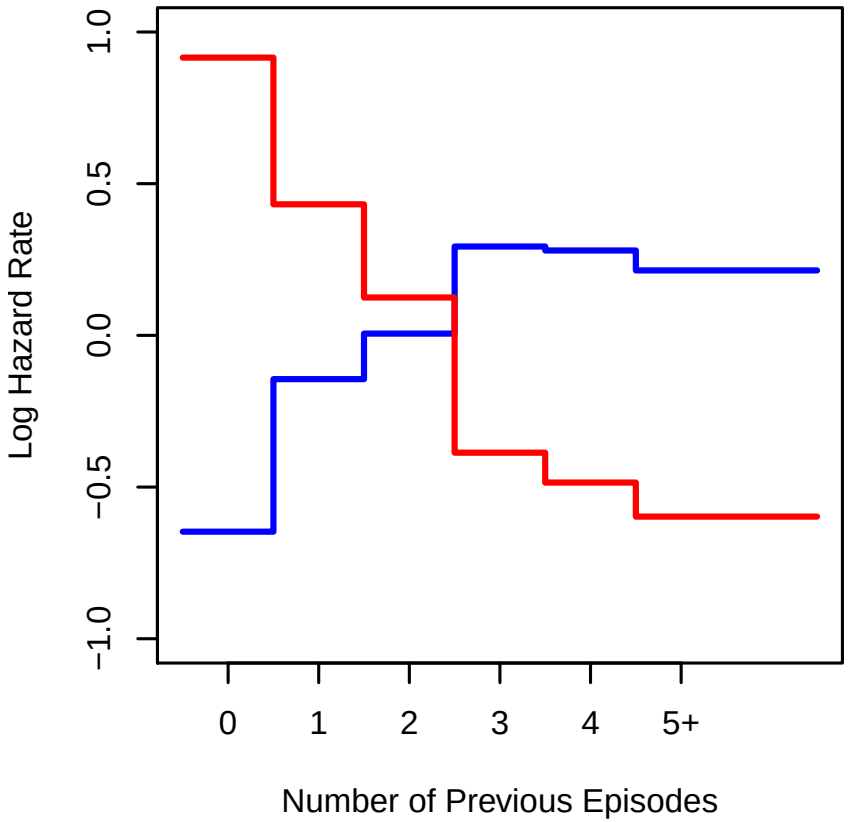


Time-at-Risk and Number of Previous Episodes Inferences

The chances of voluntarily stopping use drop dramatically after the first month of use; both hazards decrease moderately during the second year.

With increasing numbers of previous cocaine use episodes, voluntary cessation becomes less likely while incarceration risk increases.





Thank You!

Acknowledgements:

Dr. Robert Weiss (UCLA Biostatistics)

Dr. Yih-Ing Hser (UCLA ISAP)