LATENT VARIABLE PROBLEMS IN LONGITUDINAL AND LIFE HISTORY DATA

Richard Cook
Department of Statistics and Actuarial Science
University of Waterloo

OUTLINE

A INCOMPLETE HIERARCHICAL LONGITUDINAL DATA

B MISSING COVARIATES IN LONGITUDINAL STUDIES

C LATENT VARIABLES AND MULTISTATE MODELS

D TRUNCATED LIFE HISTORY DATA

Responses	Y_{j1}	Y_{j2}	$\cdots$	Y_{jK}
Missing Indicator	R_{j1}	R_{j2}	$\cdots$	R_{jK}
Covariates	$\boldsymbol{x}_{j1}$	$\boldsymbol{x}_{j2}$	$\cdots$	$\boldsymbol{x}_{jK}$

$$\boldsymbol{Y}_j = (Y_{j1}, Y_{j2}, \ldots, Y_{jK})' \qquad \boldsymbol{x}_j = (\boldsymbol{x}'_{j1}, \boldsymbol{x}'_{j2}, \ldots, \boldsymbol{x}'_{jK})' \qquad R_{jk} = 1 \text{ if } Y_{jk} \text{ observed}$$

RANDOM EFFECT MODELS

- Formulated based on $g(E[Y_{jk}|\boldsymbol{x}_j, u_i]) = u_j + \boldsymbol{x}'_{jk}\beta$ where $u_j \sim G(u_j; \theta)$
- Likelihood/Bayesian methods (EM, MCEM, MCMC, etc.) for random effects and incomplete data

MARGINAL METHODS

- Focus is in typically on $E(Y_{jk}|\boldsymbol{x}_j) = \mu_{jk}$ and sometimes $COV(Y_{jk}, Y_{jk'}|\boldsymbol{x}_j)$
- Inverse probability weighted generalized estimating equations [Selection Models]

TRANSITIONAL METHODS

- Formulated based on $g(E[Y_{jk}|Y_{j1}, Y_{j2}, \ldots, Y_{j,k-1}]) = \boldsymbol{x}'_{jk}\beta$ [missing responses $\rightarrow$ missing covariates]
- Likelihood the most common basis for inference, though increasing interest in robust methods for marginal "transition probabilities" and estimating function framework

A.1 INCOMPLETE CLUSTERED LONGITUDINAL DATA

EXAMPLE: SCHOOL-BASED LONGITUDINAL STUDIES

- schools form the clusters (J_i students for school i) and each student is followed longitudinally for smoking status

ISSUES TO CONSIDER

Random effect models generalize nicely for incomplete hierarchical data

For marginal methods, "selection model" for inverse probability weighted estimating functions should deal with variation between clusters in propensity for "missing data"

- via latent cluster-specific random effects?
- via multivariate marginal methods?

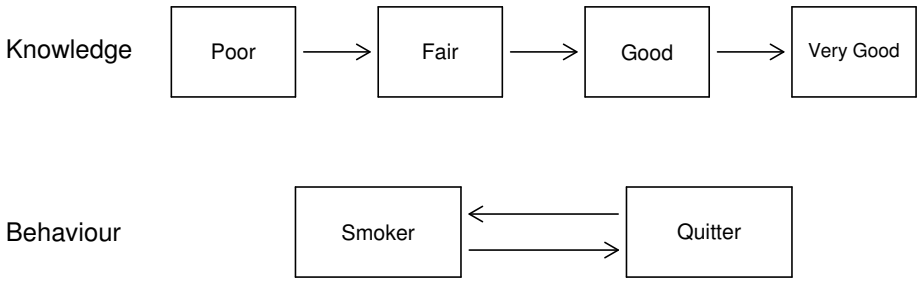
For transitional models, clustering of transition occurrences must be dealt with

- multivariate latent random effects at cluster level for each transition probability?
- dealing with incomplete data via latent variables as well? [Recall: $g(E[Y_{jk}|Y_{j1}, Y_{j2}, \ldots, Y_{j,k-1}]) = \boldsymbol{x}'_{jk}\beta$]

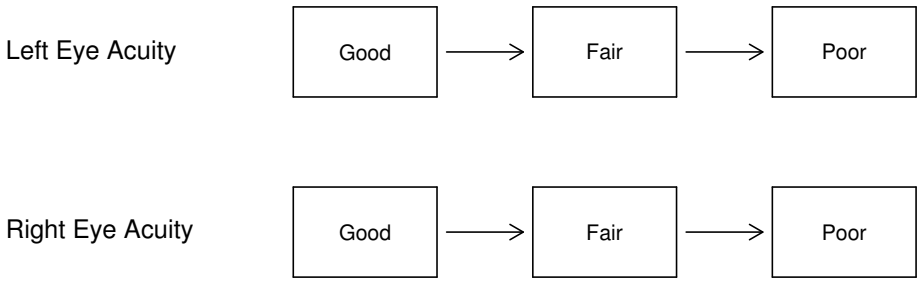
A.2 MULTIVARIATE MULTI-STATE PROCESSES

Attention may be directed towards two or more multi-state processes when

- studying effects of health promotion interventions on knowledge, attitudes, and behaviour

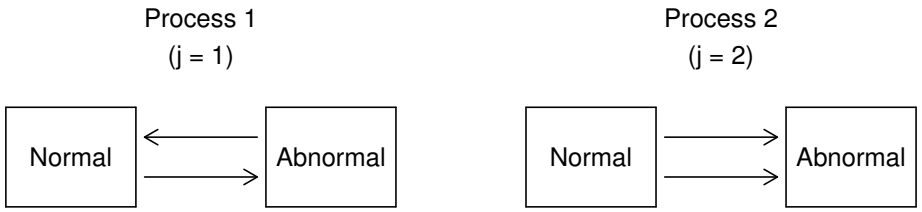


- examining deterioration of paired organ systems (opthalmology, nephrology, etc.)



A.2 MULTIVARIATE MULTI-STATE PROCESSES - CTD

- when assessing observer agreement on dynamic disease processes



Interest may lie in

- characterizing how these processes change together or track one another
- estimating or testing global effects
- improving efficiency through joint modeling

ISSUES TO CONSIDER

Insight can be gained in the nature of the associations via latent variables

- correlated random effects on transition probabilities
- associations induced by fully specified latent multi-state processes
- tractability of likelihood functions? Framework for model fitting and inference?

NOTATION FOR LONGITUDINAL STUDIES WITH TWO PHASE SAMPLING

Y, X, V Nonvalidation Sample $\Delta = 0$	Y, V Validation Sample $\Delta = 1$

- Y_j response vector
- V_j is a fully observed (auxiliary) covariate vector
- X_j is covariate vector where

$$\Delta_j = R_j = \begin{cases} 1 \text{ if } X_j \text{ is observed (validation sample)} \\ 0 \text{ otherwise} \end{cases}$$

- $f(y_j|x_j, v_j; \beta)$ where β is the parameter of interest
- $E(R_j|Y_j, X_j, V_j) = P(R_j = 1|Y_j, V_j) = \pi(Y_j, V_j)$

OBJECTIVE: To conduct inferences about β robust to misspecification of covariate distribution(s).

COMPLETE DATA LOG-LIKELIHOOD

$$\log L_C \;\propto\; \Delta\left[\log f(y|x,v;\beta)+\log f(x|v;\theta)\right]+(1-\Delta)\left[\log f(y|x,v;\beta)+\log f(x|v;\theta)\right]$$

E-STEP : Take expectation of $\log L_C$ w.r.t. $f(X|Y,V;\tilde{\beta}^{(k-1)},\tilde{\theta}^{(k-1)})$

$$\begin{aligned} E_{X|Y,V}(\log L_C) \;=\; & \Delta\left[\log f(y|x,v;\beta)+\log f(x|v;\theta)\right]+ \\ & (1-\Delta)\left\{E_{X|Y,V}[\log f(y|X,v;\beta);\tilde{\beta}^{(k-1)},\tilde{\theta}^{(k-1)}]+E_{X|Y,V}[\log f(X|v;\theta);\tilde{\beta}^{(k-1)},\tilde{\theta}^{(k-1)}]\right\} \end{aligned}$$

M-STEP : Solve

$$\begin{aligned} \frac{\partial E_{X|Y,V}(\log L_C)}{\partial\beta} &= \Delta\cdot S(y|x,v;\beta)+(1-\Delta)\cdot E_{X|Y,V}[S(y|X,v;\beta);\tilde{\beta}^{(k-1)},\tilde{\theta}^{(k-1)}] \qquad (1)\\ \frac{\partial E_{X|Y,V}(\log L_C)}{\partial\theta} &= \Delta\cdot S(x|v;\theta)+(1-\Delta)\cdot E_{X|Y,V}[S(X|v;\theta);\tilde{\beta}^{(k-1)},\tilde{\theta}^{(k-1)}] \end{aligned}$$

B.1 TWO-PHASE STUDIES WITH REPEATED MEASURES DESIGNS

Responses	Y_{j1}	Y_{j2}	$\cdots$	Y_{jK}
Covariates	$\boldsymbol{x}_{j1}$	$\boldsymbol{x}_{j2}$	$\cdots$	$\boldsymbol{x}_{jK}$
Auxiliary Covariates	$\boldsymbol{v}_{j1}$	$\boldsymbol{v}_{j2}$	$\cdots$	$\boldsymbol{v}_{jK}$
Indicator	R_{j1}	R_{j2}	$\cdots$	R_{jK}

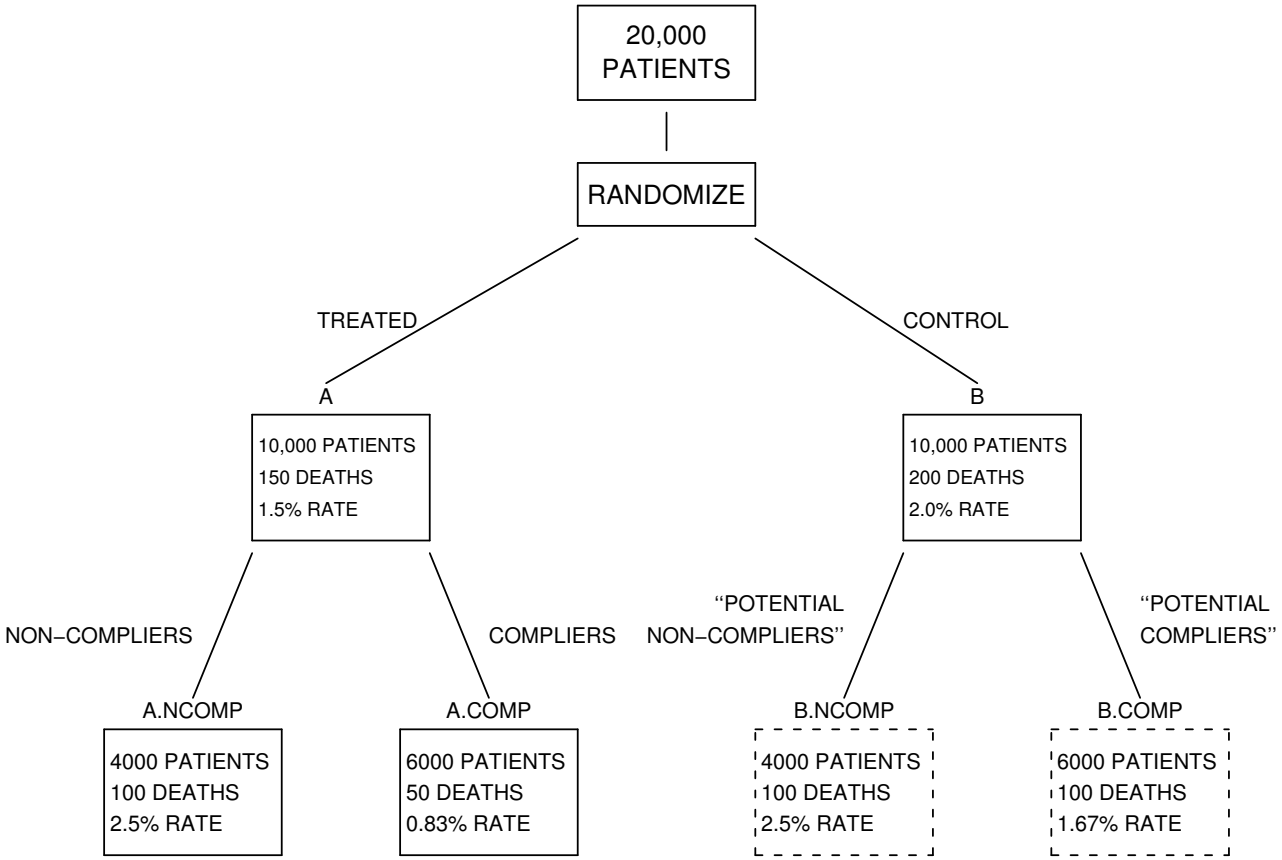
ISSUES TO CONSIDER

Factors that should affect the sampling probabilities for x_{jk} given $(\boldsymbol{y}_{jk}, \boldsymbol{v}_{jk})$ and process history for

- random effect models
- marginal models
- transitional models

Robustness and efficiency trade-offs between likelihood and estimating function approaches

B.2 NONCOMPLIANCE IN LONGITUDINAL STUDIES



B.2 NONCOMPLIANCE IN LONGITUDINAL STUDIES

There are a variety of ways in which patients can become non-compliant regarding their assigned medication.

To simplify things, let us suppose that patients can be classified as compliant or non-compliant (Cuzick et al. 1997).

Responses	Y_{j1}	Y_{j2}	$\cdots$	Y_{jK}
Treatment	x_{j1}	x_{j2}	$\cdots$	x_{jK}
Compliance Indicator	c_{j1}	c_{j2}	$\cdots$	c_{jK}

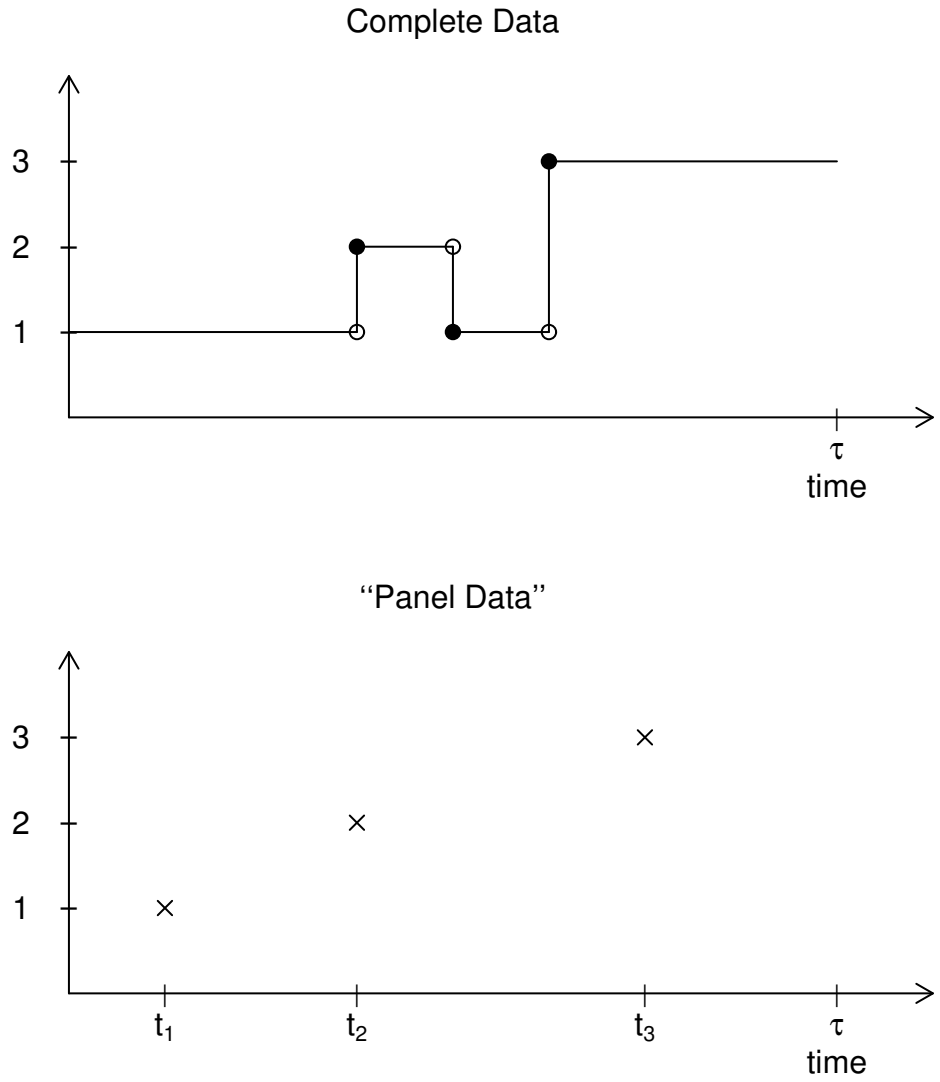
ISSUES TO CONSIDER

Need to specify models $\boldsymbol{C}_j$ given covariates

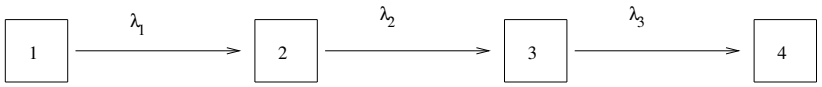
Counterfactual arguments may be helpful to gain estimates of treatment effect under good compliance

Fully specified models for $\boldsymbol{Y}_j$?

C.1 MARKOV MODELS UNDER PANEL OBSERVATION



C.1 MIXED MARKOV MODELS UNDER PANEL OBSERVATION: JOINT DAMAGE IN RHEUMATOLOGY



State 1: Normal or soft tissue swelling State 2: surface erosions

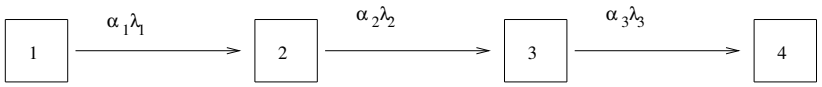
State 3: joint space narrowing State 4: disorganization/damage/surgery

Data for a single subject may be represented as follows

- $t_0 < t_1 < \cdots < t_m$ are observation times (in, say months since clinic entry)
- $Y(t_r)$ is the state occupied at time t_r (e.g. $Y(t_0) = 1$ if joint starts out normal or with soft tissue swelling).
  - predictor variables (e.g. demographic factors, family history, genetic data)
  - time-dependent predictor variables only observed at $t_0, t_1, ..., t_m$

C.1 MIXED MARKOV MODELS UNDER PANEL OBSERVATION: JOINT DAMAGE IN RHEUMATOLOGY

Consider



where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)'$ has a "genuine" trivariate distribution.

- var(α_k) reflects degree of variation in $k \longrightarrow k+1$ transition intensity
- corr$(\alpha_k, \alpha_{k'})$ reflects extent to which higher $k \longrightarrow k+1$ intensities are associated with higher $k' \longrightarrow k'+1$ intensities

ISSUES TO CONSIDER

- Likelihood functions are intractable
- MCMC algorithms for dealing with latent random effects ?
- MCMC algorithms for dealing with latent random effects and transition times?

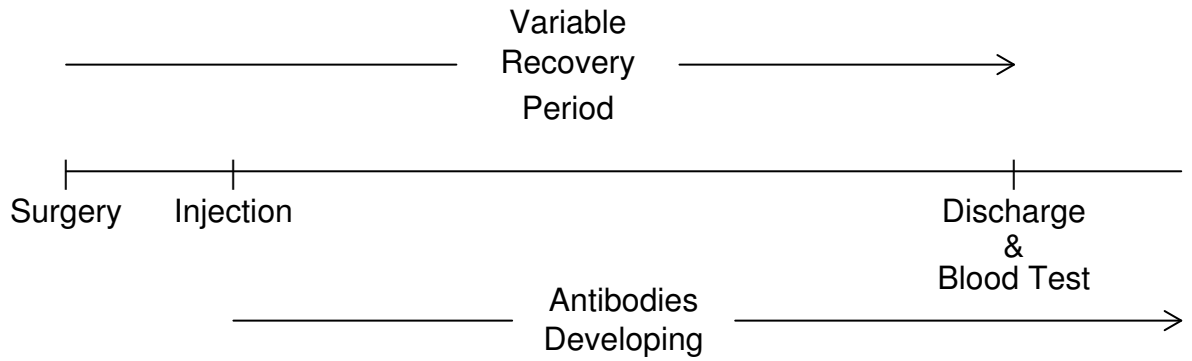
C.2 CURE RATE MODELS UNDER CURRENT STATUS OBSERVATION

Features of some recent orthopedic surgery trials

- Drugs were injected after surgery:
  - Enoxaparin - 12-24 hours
  - Fondaparinux - 4-8 hours
- Patients recovered from surgery in the hospital.
- Upon discharge, blood tests were conducted to assess seroconversion.

Study	Drug	Patients	Med. Time to Discharge
Pentamaks	Enoxaparin	365	4.82
(Knee)	Fondaparinux	388	5.34
Pentathlon	Enoxaparin	984	4.88
(Hip)	Fondaparinux	989	5.42

C.2 CURE RATE MODELS UNDER CURRENT STATUS OBSERVATION



Interest is not in when patients seroconvert, but *whether* they seroconvert.

Suggests the use of a "cure-rate model"

CURRENT STATUS DATA

- Extreme case of interval censoring.
- All subjects tested for the presence of a condition (seroconversion) at only one point in time (discharge time).

Notation :

- $t_i =$ test time (or discharge time) for subject i.
- $\delta_i = \begin{cases} 1, & \text{if subject } i \text{ tested positive (seroconverted)} \\ 0, & \text{otherwise} \end{cases}$
- $\bar{F}(\cdot) =$ s.f. for the time to condition onset (seroconversion) for full population.

Likelihood :

$$\mathcal{L}_{CS} = \prod_{i=1}^{n} \left[1 - \bar{F}(t_i)\right]^{\delta_i} \left[\bar{F}(t_i)\right]^{1-\delta_i}$$

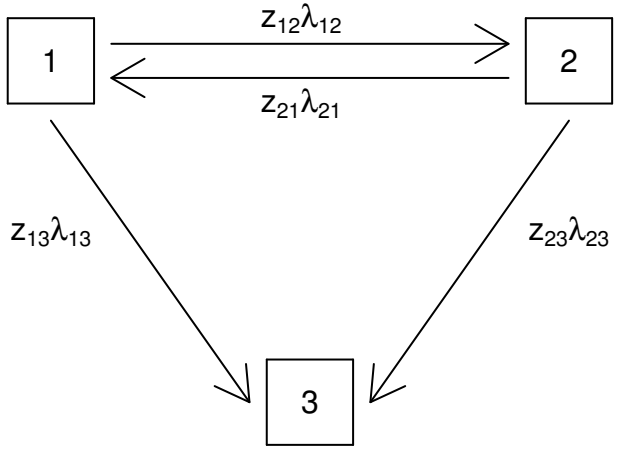
$$l_{CS} = \sum_{i=1}^{n} \left\{\delta_i \log\left[1 - \bar{F}(t_i)\right] + (1 - \delta_i) \log\left[\bar{F}(t_i)\right]\right\} \tag{2}$$

MOVER-STAYER MODELS FOR CURRENT STATUS DATA

Let

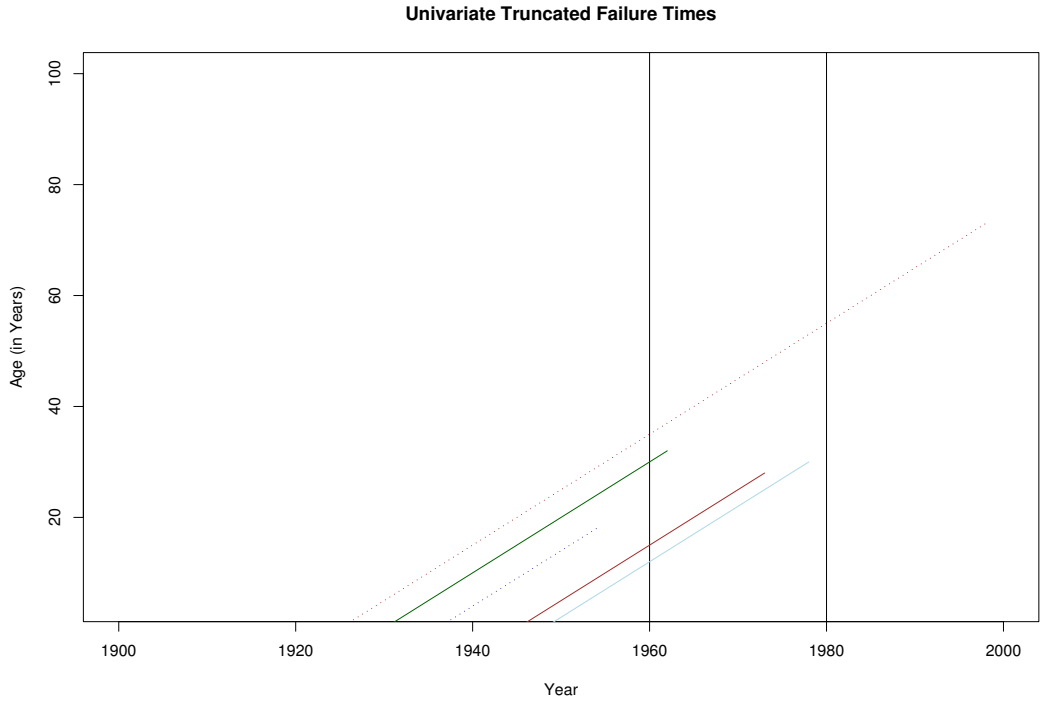
- π = probability of seroconversion
- $z_i = \begin{cases} 1, & \text{if patient } i \text{ will eventually seroconvert} \\ 0, & \text{otherwise} \end{cases}$ (Latent Variable)
- $P(z_i = 1) = \pi$ and $P(z_i = 0) = 1 - \pi$
- S = time to seroconversion ($S \to \infty$ if $z_i = 0$)
- $\bar{F}_1(\cdot)$ = s.f. for population of seroconverters ($z_i = 1$)

LATENT VARIABLE MODELS FOR MORE GENERAL MULTI-STATE PROCESSES



Such models are useful to describe unusually long sojourns in particular states.

UNIVARIATE TRUNCATED DATA



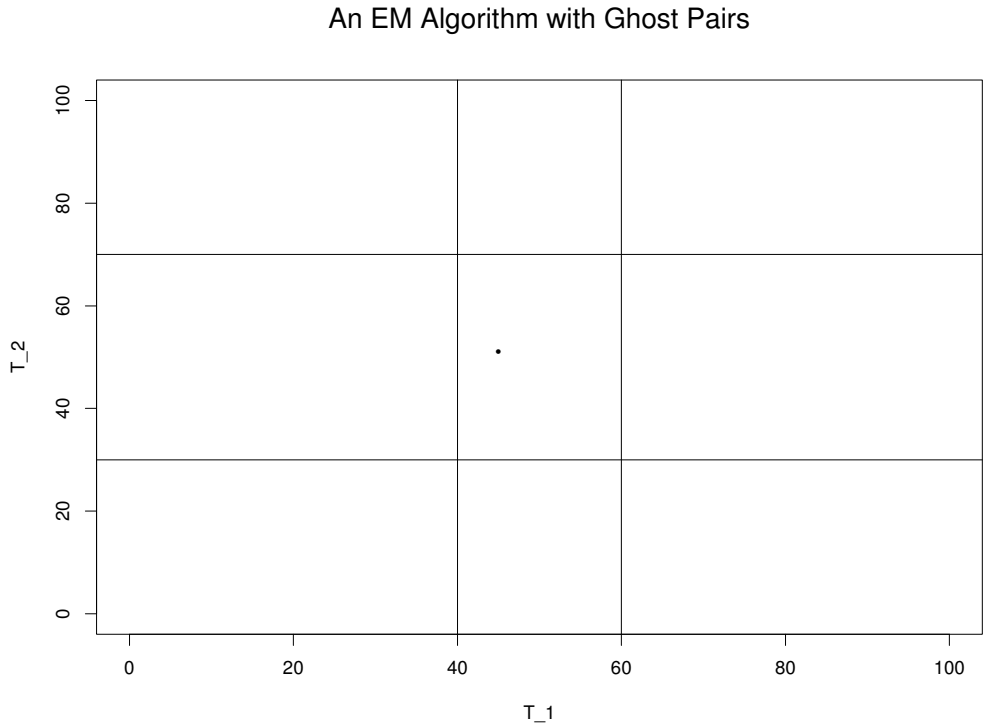
Data on failure times are available only between 1960 and 1980.

- Y_i is the year of birth for the ith subject
- X_i is the year of event for the ith subject
- $T_i = X_i - Y_i$ is the age when event occurs
- $B_i = [L_i, R_i] = [1960 - Y_i, 1980 - Y_i]$

Likelihood contribution from subject i:

$$F(t_i|T_i \in B_i) = P(T_i \leq t_i|T_i \in B_i)$$

COMPLETE DATA



Let $\mathbf{C}_i = \mathbf{B}_i^c$ be the complement of the bivariate truncation region

Let $\mathbf{C}_i = \cup_{j=1}^{J} \mathbf{C}_{ij}$

K_{ij} is the number of "ghost pairs" in $\mathbf{C}_{ij}$ for pair i

$\boldsymbol{u}_{ijk} = (u_{ijk1}, u_{ijk2})'$ are times for kth ghost pair in C_{ij}, and $\boldsymbol{u}_{ij} = (\boldsymbol{u}'_{ij1}, \ldots, \boldsymbol{u}'_{ijK_{ij}})'$

$\mathbf{Y}_i = (\mathbf{X}_i, (\boldsymbol{u}_{ij}, K_{i1}), \ldots, (\boldsymbol{u}_{iJ}, K_{iJ}))$ is the complete data for the ith pair

$\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ is the complete data

COMPLETE DATA LIKELIHOOD

The complete data likelihood can be written as

$$L_c(\theta; \mathbf{Y}) = \prod_{i=1}^{n} \Big[f(t_{i1}, t_{i2}) \times \prod_{j=1}^{J} \prod_{k=1}^{K_{ij}} f(\boldsymbol{u}_{ijk}) \Big] ,$$

- $K_{i1}, \ldots, K_{iJ}$ are unobserved
- $\mathbf{u}_{ijk}$ times for kth "ghost pair" in C_{ij} are also unobserved

The corresponding complete data log-likelihood is

$$l_c(\theta; \mathbf{Y}) = \sum_{i=1}^{n} \Big[\log f(t_{i1}, t_{i2}) + \sum_{j=1}^{J} \sum_{k=1}^{K_{ij}} \log(f(\boldsymbol{u}_{ijk})) \Big] .$$

SOME ROLES OF LATENT VARIABLE MODELS

- Sometimes latent variables are introducted simply to facilitate estimation
- Role is sometimes simply to increase flexibility of a model - interpretation of latent process may not be critical
- Sometimes interpretation of parameters hinges critically on notion of a latent process

QUESTIONS

- How complex should models for latent processes get (parameter driven versus data driven models)?
- Diagnostics for latent class models?
- Sensitivity analyses may play an important role