I. Introduction

I.1. Examples

I.1.1. Event History Data

Example 1 — Tumorgenicity Experiments

Tumorgenicity experiments usually refer to studies on animals for investigating toxic effect of certain chemicals to cause cancer. In these situations, the variable of interest is usually the time to tumor onset, which is often not directly observable. Instead, only the death time of animals under study and the status of tumor onset at the death time are observed.
Example 2 — Bone Marrow Transplantation Studies
Example 3 — National Cooperative Gallstone Study

The data in this example comprise first year follow up of the patients with floating gallstones in a follow-up study in two study groups, placebo (48) and high-dose chenodiol (65), from the National Cooperative Gallstone Study. The data include the successive visit-times in study weeks and the associated counts of episodes of nausea for 113 patients. One of the objectives of the study is to test the difference of the two treatments with respect to the incidence rate of nausea. During the study, patients were scheduled to return for clinical visits at 1, 2, 3, 6, 9, and 12 months, and asked to report the total number of each type of symptom relating to the disease (e.g., nausea) that had occurred since the last clinical visit.

Variable of interest — counts of nausea

Other variables — observation times & covariates including treatment indicators

Goal — Comparison of different treatments
Example 4 — Prevention program for traffic violations

An injury is the leading cause of death and disability among young people today and traumatic brain injury (TBI) is among the most devastating of these injuries. Those typically affected by TBI are youths between the ages of 15 and 24 and the majority of those injuries are sustained in motor vehicle crashes with the main cause being driving behaviors such as driving at excessive speed. To reduce the rate of motor vehicle crashes, one way is to conduct educational prevention programs. The University of Missouri-Columbia recently designed and conducted a one-day educational traffic injury prevention program on young traffic offenders with speeding convictions. To evaluate the program, a longitudinal retrospective study was conducted using the state traffic violation database.

Study target: Young people with speeding offenses
Objective: To assess the usefulness of a one-day educational injury prevention program
Study groups: program participants (92) & non-participants (87)
Data: Each traffic violation conviction was recorded since obtaining driving license
Covariates: Sex and other possibly related variables
Variable of interest — # of traffic violation convictions
I.1.2. Longitudinal Data

Example 1 — Growth of Sitka spruce

Study target: Sitka spruce tree
Objective: To assess the effect of ozone pollution on tree growth
Treatments: Control (54 trees) & ozone exposure groups (25 trees)
Outcome: Sizes of trees
Data: The sizes of each tree were measured at 13 different time points (13 × 79 observations)

Example 2 — Epileptic seizures

Study target: Epileptics
Objective: To assess efficacy of the anti-epileptic drug progabide
Study performed: A randomized clinical trial of 59 epileptics
Treatment: progabide or placebo in addition to chemotherapy
Outcome: The numbers of epileptic seizures during 8 (baseline) or 2 weeks
Data: The numbers of epileptic seizures were recorded at baseline and four subsequent periods
Covariates: Ages of patients in addition to treatment & baseline seizure counts
Example 3 — AIDS Clinical Trial

The data in this example arose from protocol 116A of the AIDS Clinical Trials Group and the study is a three arms randomized double-blind trial comparing three treatments, ZDV (AZT), 500mg/day DDI, and 750mg/day DDI. The HIV positive patients were randomized into one of the three treatments and the median time of follow-up was 85 weeks. During the study, the protocol called each patient to visit a clinical center to be inspected at week 2, 8, 12, 16, 24, 32, 40, 48, 56, or 64 and at each visit, their CD4 cell counts were measured in addition to other variables.

Except treatments, there are three covariates of interest, DXAIDS, AGE and RNA, where DXAIDS = 1 if a patient was diagnosed with AIDS at entry and 0 otherwise, AGE = 1 if the entry age of a patient is larger than or equal to 36 and 0 otherwise, and RNA = 1 if the baseline RNA of a patient is not below 200,000 copy numbers and 0 otherwise.

Variable of interest — CD4 counts

Goal — Treatment comparison & regression analysis
I.1.3. Issues needed to be concerned in the analysis

1. fixed numbers of observations — non-fixed numbers of observations
2. same observation times — different observation times
3. balanced observations — unbalanced observations
4. non-missing observations — missing observations
5. continuous outcomes — count outcomes
6. non-monotonic outcomes over time — increasing outcomes over time
7. real observations — summary observations
8. same stopping (censoring) times – different stopping (censoring) times
9. outcome-independent (noninformative) observation times — outcome-dependent (informative) observation times
I.2. Review of Stochastic Processes

I.2.1. Concepts

Stochastic process — a family of random variables \( X = \{ X(t) ; t \in \mathbb{R} \} \) indexed by time \( t \).

Multi-state process — a stochastic process \( X = \{ X(t) ; t \in [0, \tau] \} \) with finite state space \( S = \{1, \ldots, p\} \) and with right-continuous sample paths: \( X(t+) = X(t) \), where \( \tau \leq \infty \). For a multi-state process \( X \), the initial distribution is defined as

\[
\pi_h(0) = P\{ X(0) = h \} , \; h \in S
\]

and the history \( \mathcal{X}_t \) consists of the observation of the process in the interval \([0, t]\). Given the history \( \mathcal{X}_t \), the transition probabilities are defined as

\[
P_{hj}(s, t) = P\{ X(t) = j \mid X(s) = h , \mathcal{X}_s \}
\]

for \( h, j \in S, \; s, t \in [0, \tau], \; s < t \). The corresponding transition intensities are defined as

\[
\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P_{hj}(t, t + \Delta t)}{\Delta t}
\]

which are assumed to exist.
A state $h \in S$ is absorbing if for all $t \in [0, \tau]$, $j \neq h$, $\alpha_{hj}(t) = 0$, otherwise $h$ is transient. The state probabilities $\pi_h(t) = P\{X(t) = h\}$ are given by

$$\pi_h(t) = \sum_{j \in S} \pi_j(0) P_{jh}(0, t).$$

(Continuous time) Markov process — a stochastic process $X$ for which the transition intensities $\alpha_{hj}(t)$ depend only on the current state of $X(t)$ (and possibly via time-fixed covariates). Two popular classes of Markov processes are these with constant intensities

$$\alpha_{hj}(t) = \alpha_{hj}$$

and piecewise constant intensities

$$\alpha_{hj}(t) = \alpha^k_{hj}, \ t_{k-1} < t \leq t_k, \ t_0 = 0.$$  

Counting process — a stochastic process $\{N(t) ; t \geq 0\}$ such that $N(0) = 0$, $N(t) < \infty$, and the paths are right-continuous with probability one, piecewise constant, jumps of size of +1.

A multi-state process $X$ can be equivalently described by the record $X(0)$ and the counting processes

$$N^i_{hj}(t) = \# \text{ (direct transitions } h \to j \text{ in } [0, t] \text{ for } i).$$
**Poisson process** — a counting process \( \{ N(t); t \geq 0 \} \) such that

\[
P\{ N(t) - N(t-h) = 1 \mid \mathcal{N}(t-h) \} = \lambda(t) h + o(h)
\]

and

\[
P\{ N(t) - N(t-h) > 1 \mid \mathcal{N}(t-h) \} = o(h)
\]

for small \( h > 0 \) and all \( t > 0 \), where \( \lambda(t) \geq 0 \) is a left continuous function satisfying \( \int_0^t \lambda(u) \, du = \Lambda(t) < \infty \), called the intensity of the Poisson process, and \( \mathcal{N}(t-h) = \{ N(u) : 0 \leq u \leq t \} \). The above condition can be equivalently written as

\[
P\{ dN(t) = 1 \mid \mathcal{N}(t^-) \} = P\{ dN(t) = 1 \} = \lambda(t) \, dt
\]

and

\[
P\{ dN(t) = 0 \mid \mathcal{N}(t^-) \} = 1 - \lambda(t) \, dt .
\]

For a Poisson process and any fixed \( t > 0 \), we have

\[
N(t) \sim Poisson \left( \Lambda(t) \right) .
\]

If \( N_1(t) \sim PP(\lambda_1(t)) \) and \( N_2(t) \sim PP(\lambda_2(t)) \) are independent, then

\[
N(t) = N_1(t) + N_2(t) \sim PP(\lambda_1(t) + \lambda_2(t)) .
\]

In this case,

\[
P\{ dN_1(t) = 1 \mid dN(t) = 1 \} = \lambda_1(t) / [\lambda_1(t) + \lambda_2(t)] .
\]
For a Poisson process $N(t)$, let $T_i$ denote the time to the $i$th event. If $n$ events are observed from a Poisson process of intensity $\lambda(t)$ over the interval $[0, \tau]$, then the likelihood is

$$\prod_{i}^{n} \lambda(t_i) \exp\{-\Lambda(\tau)\}.$$ 

If $\lambda(t) = \lambda$, the process is called *time-homogeneous*. In this case, $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are independent exponential variables with rate $\lambda$. If $N(t) \sim PP(\lambda(t))$, then $N^*(s) = N(t)$ is a time-homogeneous Poisson process with intensity $\lambda^*(s) = 1, s > 0$, where $s = \int_{0}^{t} \lambda(u) \, du = \Lambda(t)$.

**Martingale**— a right-continuous stochastic process $M = \{ M(t); t \geq 0 \}$ with left-hand limits such that

$$E[M(t) \mid \mathcal{F}_s] = M(s) \text{ a.s.}, \ 0 \leq s \leq t.$$ 

**Predictable stochastic process**— a stochastic process $\{Y(t); t \geq 0\}$ such that its value at $t$ is known just before $t$ (left-continuous sample paths).
I.3. Review of Missing Data Mechanisms

Definition of Littel and Rubin (1987) — Consider a study involving two variables $X$ and $Y$ and suppose that $Y$ is subject to missing, but not $X$. We say that the missing data are

*missing completely at random* (MCAR) if the probability of obtaining an observation (e.g., response) is independent of $X$ and $Y$, or

*missing at random* (MAR) if the probability depends on $X$, but not $Y$.

If the probability depends on both $X$ and $Y$, we say that the missing data mechanism is nonignorable.

More formal definition — Suppose that in a study the variables of interest are $\{Y_i, X_i, i = 1, ..., n\}$ and only $Y_i$’s are subject to missing. Define $R_i = 1$ if $Y_i$ is missing and 0 otherwise. Then MCAR and MAR correspond to

$$f(R_i \mid Y_i, X_i, \phi) = f(R_i \mid X_i, \phi)$$

and

$$f(R_i \mid Y_i, X_i, \phi) = f(R_i \mid Y_i^{obs}, X_i, \phi),$$

respectively. The nonignorable missing data mechanism means that $f(R_i \mid Y_i, X_i, \phi)$ depends on both $Y_i^{mis}$ and $Y_i^{obs}$.
I.4. Organization of Subsequent Lectures

Recurrent Event Data

Complete observations on multi-state processes

*Counting Process Models for Life History Data: A Review*

Panel (interval-censored) Count Data

Incomplete observations on multi-state processes

Sun and Kalbfleisch (JASA, 1993; 1449-1454; Statistica Sinica, 1995, 279-290)

Random Effect Models Approach

\[ Y_{i,j} = \beta_0 + b_{0i} + \beta_1 t_{i,j} + b_{1i} t_{i,j} + \epsilon_{ij} \]

\[ Y = X \beta + Z b + \epsilon \]

Laird and Ware (Biometrics, 1982, 963-974),

Davidian and Giltinan (Chapman & Hall, 1995)

DLZ, Chapter 9
Marginal Models Approach — Generalized Estimating Equation

\[ Y = (Y_1, Y_2, ..., Y_n) \quad Y_i = (y_{i,1}, y_{i,2}, ..., y_{i,n_i}) \]

\[ E[Y] = \mu(\beta) \quad Var[Y] = \sigma^2(\beta, \alpha) \]

\[ \phi(\beta, \alpha | Y) = 0 \]


DLZ, Chapter 8

Transition Models Approach

Extensions of generalized linear models

\[ g(\mu) = X \beta \quad \mu = E(Y) \]

\[ E[y_{i,j} | (y_{i,1}, ..., y_{i,j-1})] = ??? \]

DLZ, Chapter 10