3-Lecture Minicourse on Statistics of Survival Data

Eric Slud

I. (11/6) Death Hazards & Competing Risks
Concepts:
(i) Statistical Estimation as mathematical problem,
(ii) Identifiability, nonparametric vs. nonparametric.

II. (11/13) Population Cohorts & Martingales
Concepts:
(iii) Counting process models,
(iv) “Innovations” and Statistics

III. (11/20) Models and Likelihoods with $\infty$-Dimensional Parameters
Concepts:
(v) Nuisance parameters,
(vi) Statistical Efficiency.

Lecture Slides (incl. annotated references) at:
www.math.umd.edu/~evs/SurvSlid.pdf
Data Format for a Survival Study

Subjects enter at random times $E_i$, ‘followed up’ until $E_i + T_i = \min(E_i + X_i, E_i + C_i)$ (not both observed) ‘death-time’ ($X_i = \text{lifetime}$), or ‘censoring time’ (e.g., $C_i = E_{\text{max}} - E_i + \tau$ administrative)

DATA: $\{(E_i, T_i, \Delta_i, Z_i), i = 1, \ldots, n\}$ or

$$D = \{(T_i, \Delta_i), i = 1, \ldots, n\}$$

where $T_i =$ time-on-test or event time

$\Delta_i = I_{[X_i \leq C_i]}$ death indicator

$Z_i$ auxiliary covariates, e.g. treatment gp

Assumptions: random vectors $(E_i, X_i, C_i, Z_i)$ independent and identically distributed (iid), $i = 1, \ldots, n$;

also $(X_i, C_i)$ have continuous joint density, i.e.

$$\lim_{\delta \to 0} \frac{1}{\delta^2} P(X_1 \in (x, x+\delta), C_1 \in (c, c+\delta)) = f_{X,C}(x,c)$$

OBJECTIVE: to estimate the marginal survival function $S_X(t) = P(X_1 > t) = 1 - F_X(t)$ consistently from the data $D$, which means ...
Definition. Say that a sequence of (measurable) mappings

$$\hat{F}^{(n)} : (\mathbb{R}^+ \times \{0, 1\})^n \longrightarrow \{F : \text{dist.fcns on } \mathbb{R}^+ \}$$

are consistent estimators of \( F_X = 1 - S_X \) based on the iid pairs \((T_i, \Delta_i)\) if

$$\|\hat{F}^{(n)}(\{(T_i, \Delta_i)\}_{i=1}^n) - F_X\|_\infty \longrightarrow 0$$
in expectation or (more restrictively) with probability 1.

Say that a function(al) \( \vartheta = \vartheta(f_{X,C}) \) of the joint probability density of (each subject’s underlying) data \((X_i, C_i)\) is identifiable if it depends only on \( f_{T,\Delta}(t, j) = \lim_{\delta \downarrow 0} \frac{1}{\delta} P(\min(X_1, C_1) \in (t, t + \delta), \Delta_1 = j) \)

\[= \int_t f_{X,C}(t, c) dc I_{[j=1]} + \int_t f_{X,C}(s, t) ds I_{[j=0]} \]

Joint prob. density function \( f_{T,\Delta} \) (mixed-type continuous & discrete) is, due to the Law of Large Numbers (pointwise for each \( t \)) and Glivenko-Cantelli Theorem (uniformly in \( t \in [0, \infty) \)) consistently estimated by the (derivative with respect to \( t \) of the) empirical subdistribution functions

$$\hat{F}_{T,\Delta=j}(t) = \frac{1}{n} \sum_{i=1}^n I_{[T_i \leq t, \Delta_i = j]}$$
Summary So Far: from def’ns, consistent estimation generally requires identifiability. Identifiability plus suitable continuity in the parameter-functional implies a consistent estimator-sequence can be found.

In the LATENT FAILURE MODEL, want to identify in $S_X(t)$ what the survival probabilities would be with no removals due to $C_i$.

This makes clearer sense if $C_i$ is administrative censoring and $E_i$ is unrelated to health than if $C_i$ is due to death from another cause. (In that case, called Competing Risks, death-variable $X_i$ following $C_i$ is counterfactual.) We have no data on deaths following removals!

Assume $X_i, C_i$ independent (densities $f_X, f_C$). Following D. Bernoulli (1760), Kaplan & Meier (1958), S. Berman (1960) we show $S_X$ identifiable, consistently estimated.

Idea: Actuarial concept of Death Hazards. Actuaries call this topic “multiple decrement tables” to find e.g. probabilities of wife’s survival probabilities following husband’s death, for joint-insurance premiums.
Death Hazards

In general, define hazard intensity

\[ h_X(t) \equiv \lim_{\delta \to 0} \frac{1}{\delta} P(X \in (t, t + \delta) \mid X > t) = \frac{f_X(t)}{S_X(t)} \]

Then

\[ h_X(t) = -\frac{d}{dt} \ln S_X(t) \implies S_X(t) = \exp \left( -\int_0^t h_X(s) \, ds \right) \]

In terms of observed data \((T_i, \Delta_i)\),

\[ S_T(t) = P(T_1 > t) = P([X_1 > t] \cap [C_1 > t]) \]
\[ = P(X_1 > t)P(C_1 > t) = S_X(t)S_C(t) \]

and also

\[ P(t < T_1 < t+\delta, \Delta_1 = 1) \approx P(t < X_1 < t+\delta, C_1 > t) \]
\[ \approx \delta f_X(t) S_C(t) \]

So determine

\[ \delta h_X(t) \approx \frac{P(t < T_1 < t < \delta, \Delta_1 = 1)}{P(T_1 > t)} \]

\[ S_X(x) = \exp \left( \int_0^x \frac{1}{S_T(t)} \frac{d}{dt} P(T_1 > t, \Delta_1 = 1) \, dt \right) \]
Kaplan-Meier Curve

The ‘empirical’ estimators are:

for \( S_T(t) : \frac{1}{n} \sum_{i=1}^{n} I_{[T_i > t]} \)

for \((P(T_1 > t, \Delta_1 = 1) : \frac{1}{n} \sum_{i=1}^{n} I_{[t < T_i < C_i]}\)\)

for \( S_X(x) : \prod_{0 \leq t \leq x} \left( 1 - \sum_{i=1}^{n} I_{[T_i = t, \Delta_i = 1]} / \sum_{i=1}^{n} I_{[T_i \geq t]} \right) \)

This is the Kaplan-Meier (1958) estimator, known to actuaries in discretized form 80 years earlier.

**What if \( X_i, C_i \) are dependent?**

Depends on the unknowable counterfactual hazards:

\[ \lim_{s \to 0} \frac{1}{\delta} P(T \in (t, t + \delta) \mid C = s) \quad \text{for} \quad s < t \]

**Anyway:**

\[ \hat{S}_X^{KM}(x) \longrightarrow \exp \left( \int_0^x \frac{1}{S_T(t)} \frac{d}{dt} P(T_1 > t, \Delta_1 = 1) \, dt \right) \]
Parametric vs Nonparametric Models

So what do biostatisticians do to identify $S_X$?

**Main approaches:**

1. *(cf. David & Moeschberger 1978 book)* Parametrize joint density $f_{X,C} = f_{X,C}(\cdot, \cdot | \vartheta)$ and therefore $f_{T,\Delta}(t, j) = f_{T,\Delta}(t, j | \vartheta)$ (untestable assumption!).

**Idea:** $(P(T > t, \Delta = 1), P(T > t)) \leftrightarrow \vartheta \mapsto S_X$

**Example:** $(\log X, \log C) \sim \mathcal{N}$ bivariate-normal!

Assumptions about $f_{T,\Delta}$ can be tested from large datasets using $n^{-1} \sum_{i=1}^{n} I_{[T_i > t, \Delta = j]}$, but assumptions about $f_{X,C}$ cannot!

2. Find qualitative assumption just enough to render $S_X$ identifiable.

3. Models for additional observed variables $V$ and qualitative assumptions they satisfy wrt $(X, C)$.
Since $S_X$ is not identifiable nonparametrically, what additional piece of information would be just enough for identifiability?

Slud & Rubinstein (1983) show based on $f_{T,\Delta}$ that $S_X$ is monotone \( \leq \) functional (expressed as ODE sol’n) of

$$\rho(t) = \lim_{\delta \to 0} \frac{P(X < t + \delta | T < t, X > t)}{P(X < t + \delta | T > t)}$$

One-to-one correspondence: $S_X \leftrightarrow \rho$

**Cases.**

$\rho \equiv 1$ : includes independence of $X, C$
   Kaplan-Meier estimator consistent.

$\rho \approx 0$ : minimal, $S_X(t) \approx P(\Delta = 0 \cup T > t)$
   censored never die.

$\rho \nearrow \infty$ : maximal $S_X(t) \approx S_T(t)$
   death just after censoring.

**Outcome:** bounds $r_1 < \rho(\cdot) < r_2$ give (consistently estimated) bounds on $S_X$.

**General Problem:** Assumptions on $\rho$ needed for identifiability, but not testable from data!

7
Survival Curves in Study on Cardiovascular Deaths
167 Patients, VA study, Heart Attack vs other deaths

Surv. time (months)
Survival Prob

\( \rho = 0 \)
\( \rho = 0.5 \)
\( \rho = 1 \)
\( \rho = 2 \)
\( \rho = \infty \)
Formulations Using ‘Covariates’

(A) \((X, C)\) conditionally independent given \(V\)

Nonparametric regression Cheng 1989 & others

\[
f_{X,C,V}(x, c, v) = f_V(v) f_{X|V}(x|v) f_{C|V}(c|v)
\]

\(S_X\) can be estimated from \((T_i, \Delta_i, V_i)\) using density estimates or regression models. (Think: Kaplan-Meier on subpopulations with \(V_i = v\) if \(V\) is discrete.)

\[
S_X(t) = \int \exp \left( - \int_0^t \frac{dP(X \leq \min(s, C)|V = v)}{P(\min(X, C) \geq s|V = v)} \right) dF_V(v)
\]

Under this assumption, can use testable regression models for \((X, V)\) and/or \((C, V)\) dependence! This is the starting point for Lectures 2,3.

(B) \(C, V\) conditionally indep. given \(X\). Under regularity conditions

(a) \(\int f_{C|X}^2(c|x) f_X(x) dx < \infty \quad \forall c\)

(b) the functions \(\{f_{V|X}(v|\cdot) : v \in \mathbb{R}\}\) are linearly dense in \(L^2(f_X(x)dx)\)

Slud (1992) proves \(S_X\) is uniquely determined by \(F_1, S_T\).

(Estimation involves nonparametric mixture density, very inaccurate.)
Nonparametric assumptions specifying dependence between $X_i, C_i$

**EXAMPLE:** Zheng & Klein (1995, *Biometrika*) assume known form of ‘Copula’ with $f_X$, $f_C$ unknown

$$K(u,v) \equiv P(F_X(X) \leq u, F_C(C) \leq v)$$

and prove: if $K$ is bivariate distribution function on $[0,1]^2$ with Uniform[0,1] marginals assigning positive probability to all open sets of $[0,1]^2$, then $f_X$, $f_C$, $f_{X,C}$ are uniquely determined by $F_1$, $S_T$.

**EXAMPLE:** Emoto & Matthews (1990) assume for (unknown) measure $\pi$ on $[0,1]$ that

$$P(X > x, C > y) = \exp \left( - \int \max(px^\alpha, (1-p)y^\beta) \, \pi(dp) \right)$$

with $\alpha \neq \beta > 0$, and show that $\pi$, $f_{X,C}$ are uniquely determined by $F_1, S_T$.

*Statistically, these models beg the question how dependence model could be known!*?
Annotated References, Lecture 1

Probability Theory:

books by M. Loeve; P. Billingsley, ...

*Law of Large Numbers, Glivenko-Cantelli Thm.*

Statistics:


R. Miller, *Survival Analysis* (1980) *good general book, also back in print*

Competing Risks:

D. Bernoulli (1760): removing Smallpox mortality

Gail, M. (1975), *Biometrics* review article

(also 1980 Encycl. Statist. Sci. article)

Tsiatis, A. (1975) *PNAS nonidentifiability*

Prentice, R. et al. (1976) *Biometrics, counterfactuals*

David, H. and Moeschberger, M. (1978) **Theory of Competing Risks**

Slud, E. and Rubinstein, L. (1983) *Biometrika*

*summarizes Dependent Competing Risks problem, defines \( \rho(t) \) function, obtains estimation results.*
shows that estimating survival under assumption of independence between survival and censoring within two risk-groups could exactly reverse the actual ordering between the groupwise survival functions.

Slud, E. (1992) proceedings paper: showed that $S_X$ is identifiable under regularity conditions from $S_T, F_1$ if $C, V$ are conditionally independent given $X$.


Cox, D.R. (1972) *Jour Roy Stat Soc B*
seminal paper on survival-analysis regression models based on $V$, when $C, X$ are conditionally independent given $V$.


Zheng & Klein (1995) *Biometrika*, identifiability of $f_{X,C}$ under a nonparametric assumption on dependence of $X_i, C_i$. 

12