SYLLABUS, PH 240B SURVIVAL ANALYSIS and CAUSALITY

Instructor: Mark van der Laan
Office: 108 Haviland Hall
Tel: 510-643-9866
email: laan@stat.berkeley.edu
Topics, not necessarily in the following order:

WEEK 1 Formal definition of right censoring of a time-dependent multivariate stochastic process which we will refer to as monotone censoring. Coarsening at random/non-informative censoring. Observed data model in terms of a full-data model and censoring model.

WEEK 2 Marginal right-censored data on a survival time. Product integral relation between hazard-measure and survival function. Kaplan-Meier estimator of the survival function. Parametric models and maximum likelihood estimation.

Week 3 Parametric models, Asymptotic linearity of an estimator. Influence curve of an estimator. Construction of confidence bands based on the influence curve. Asymptotics of the maximum likelihood estimator, influence curve of the maximum likelihood estimator. Likelihood ratio test, Score test.

Week 4 Asymptotic efficiency in parametric models. Cramer-Rao lower bound. Efficiency of the maximum likelihood estimator. Kaplan-Meier estimator as Nonparametric Maximum Likelihood Estimator (NPMLE). Asymptotics of Kaplan-Meier estimator: Parametric Information matrix approach, Greenwoods formula.

Week 5 The EM-algorithm for computing the NPMLE by treating the likelihod as a multinomial likelihood: EM-algorithm corresponds with iterating Self-Consistency Equation. Asymptotics of Kaplan-Meier estimator: Functional Delta Method Approach. Influence curve of the Kaplan-Meier estimator.

Week 6-10 Multiplicative intensity models for counting processes. Counting process. History. Intensity of counting process. Multiplicative intensity model for intensity of counting process. Partial likelihood. Profile partial likelihood for regression coefficients. Brewlow-Aalen estimator. Asymptotics for profile likelihood estimator of the regression coefficients. Examples of counting processes: Survival data (Cox-proportional hazards model), recurrent event data.

Week 10-12 General monotone censored data structure. Estimating equation approach. Nonparametric full data model. Multivariate generalized linear regression full data model. Full data model estimating equations. "Inverse probability of censoring weighted mapping" from full-data estimating equations to observed data estimating equations. Mapping from all full-data estimating equations to all observed data estimating equations. Optimal estimating equation.

Week 13-15 Causal inference with point treatment studies. Causal inference with time-dependent treatment based on general monotone censored data Examples we will cover: Treatment specific survival function, causal odds-ratio of treatment, causal linear effect of treatment.

Evaluation of students: The evaluation of the students will be based on weekly homework assignment, and poster project at the end of semester.

Homework 1: Simulate n observations (T_i, C_i), constuct corresponding right-censored observations $(\tilde{T}_i = \min(T_i, C_i), \Delta_i = I(T_i \leq C_i))$, $i = 1, \ldots, n$, and compute the Kaplan-Meier estimator of survival function $S(t) = P(T > t)$. Do this for a variety of data generating distributions for (T, C), including the case that T is independent of C, and data generating distributions in which C and T are dependent. Plot the Kaplan-Meier estimators and the true survival function of T and comment on the observed bias.

Homework 2: Suppose we observe n i.i.d copies of W, A, Y, W baseline covariates, A treatment, Y final outcome. We assume the so called consistency assumption $(W, A, Y) = (W, A, Y(A))$, where $(Y(a) : a)$ are treatment specific counterfactual outcomes one would have observed if the subject follows treatment $A = a$. This consistency assumption states that the observed data (W, A, Y) is a missing data structure on $X = ((Y(a) : a), W)$. We also assume the randomization assumption, which states that A is independent of $(Y(0), Y(1))$, given W: $P(A = a|X) = P(A = a|W)$. We wish to estimate the $\mu(a) = EY(a)$. In this homework you need to simulate from the above model, make sure you know the true $\mu(a)$, and implement three estimators of $\mu(a)$ based on n draws (W_i, A_i, Y_i), $i = 1, \ldots, n$.

For simplicity assume that A is a binary treatment with values 0 and 1. Generate n times 1) a full data $X = (W, Y(0), Y(1))$, 2) A, given W, from a logistic regression with specified parameters, and 3) construct the corresponding observation $(W, A, Y(A))$. This gives you now a data set of n observations from the above model. What is $EY(1)$ and $EY(0)$. Now, construct an estimator of the treatment mechanism $\Pi(a|W) = P(A = a|W)$ using logistic regression by regressing the binary A on W: we will refer to this estimator as Π_n. Also construct an estimator of the regression $E(Y|A = a, W)$ by regressing Y on A, W: we will refer to this estimator as $Q_n(a, W)$.

Implement the IPTW estimator of $\mu(a)$:

$$\mu_{n,IPTW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a)}{\Pi_n(a \mid W_i)} Y_i.$$

Implement the likelihood based estimator:

$$\mu_{n,Gcomp}(a) = \frac{1}{n} \sum_{i=1}^{n} Q_n(a, W_i).$$

Implement the DR-IPTW estimator:

$$\mu_{n,DR}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a)}{\Pi_n(a \mid W_i)} (Y_i - Q_n(a, W_i)) + Q_n(a, W_i).$$

In addition, implement the naive estimator

$$\mu_{n,naive}(a) = \frac{\sum_{i=1}^{n} Y_i I(A_i = a)}{\sum_{i=1}^{n} I(A_i = a)}.$$

Compute these four estimators in the case that 1) only Π_n is misspecified, 2) only Q_n is misspecified, and 3) both are correctly specified. You might also wish to play with the strength of W as a predictor of A and Y.

Show the R-code, the results, and comment.

Homework 3: Suppose we observe n i.i.d. $X_1, \ldots, X_n$ with survival function S_0. Let $S_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i > t)$ be the empirical survival function. Given a rich collection of points t, we wish to construct a simultaneous confidence interval of the vector survival function $(S_0(t) : t)$ of the form $(S_n(t) \pm q_{0.95} \sigma_n(t)/\sqrt{n} : t)$, where $\sigma_n^2(t)$ is an estimate of the variance of the influence curve $IC(X|t) \equiv I(X > t) - S_0(t)$, and $q_{0.95}$ is an estimate of the 0.95 quantile of

$\max_t \sqrt{n} \mid S_n(t) - S_0(t) \mid /\sigma_n(t)$. Here we need to use that the latter random variable behaves as the max over t of $Z(t)$, where $Z \sim N(0, \rho)$ and ρ is the correlation matrix of the vector influence curve $(IC(X|t) : t)$. Implement this as an R-function for construction of a simultaneous confidence interval, and make sure that it indeed contains the true survival function approximately 95% of the times among a set of trials (i.e., sample n $X_1, \ldots, X_n$ many times and each time compute the simultaneous confidence interval).