Lecture 1 Notes

Daniel Rubin

January 21, 2004

Logistics

Professor: Mark van der Laan
Email: laan@stat.berkeley.edu
Office: 108 Haviland Hall
Office Hours: MW 1-2
Textbook: None required, but material is covered in a handout distributed to the class. Optional reading is *The Statistical Analysis of Failure Time Data* by Kalbfleisch and Prentice.
Grading: Based on attendance, class notes (students write their notes for one lecture as a LaTex document), occasional homework, and a group poster project at the end of the term.

Survival Analysis

Survival analysis is the branch of statistics concerning time-to-event data. Specialized techniques exist for situations where there is only partial information about the times-to-events. For example, the data could record the time until recurrence of cancer in patients, but instead of having times for each patient we may only know that some patients had no recurrence for at least five years.

Notation

- T = *Survival time.* T is a nonnegative random variable.
- F(t) = P(T $\leq$ t). F is the *cumulative distribution function* (cdf).
  F(t)=0 for t $\leq$ 0, F(t) $\rightarrow$ 1 as t $\rightarrow \infty$, F is nondecreasing, and F is right continuous.
- S(t) = 1-F(t) = *Survival function.*
- dF(t) = P(t-dt $\leq$ T $\leq$ t), for dt infintesimal. dF(t) = f(t)dt if F continuous with Lebesgue density f, and dF(t) = F(t) - F(t_-) if F is a discrete cdf.
- If $T_1, ..., T_n$ are iid observations of T, then $F_n(t) = \frac{1}{n}\Sigma_{i=1}^{n} I(T_i \leq t)$ and $S_n(t) = 1 - F_n(t)$ are the *empirical cdf* and the *empirical survival function.* If all observations are unique and $T_{(1)}, ..., T_{(n)}$ denote the sorted observations in increasing order, then $F_n(t) = 0$ for $t <= T_{(1)}$, $F_n(t) = \frac{i}{n}$ for $T_{(i)} \leq t < T_{(i+1)}$, and $F_n(t) = 1$ for $t \geq T_{(n)}$. When the observations are unique, we also have that $dF_n(t) = 1/n$ if $t \in \{T_1, ..., T_n\}$ and $dF_n(t) = 0$ zero otherwise.

- $\lambda(t)$ = *Hazard function.*
  If T is continuous, the Lebesgue hazard is $\lambda(t) = \frac{f(t)}{S(t)} = \lim_{\delta \to 0} \frac{P(T \in (t, t+\delta])/\delta}{P(T \geq t)}$
  $= lim_{\delta \to 0} \frac{(F(t)-F(t-\delta))/\delta}{S(t)} = lim_{\delta \to 0} P(T \in (t, t+\delta] | T \geq t)/\delta$
  If F is discrete, the discrete hazard is $\lambda(t) = P(T = t | T \geq t)$.
  When T represents the time until a failure, $\lambda(t)$ represents the probability of instantaneous failure given that no failure has occured in [0, t).

- $\Lambda(t)$ = *Cumulative hazard function.*
  $\Lambda(t) = \int_0^t \frac{1}{S(s_-)} dF(s)$, and $d\Lambda(s) = \frac{1}{S(s_-)} dF(s) = P(T \in [s - ds, s])/P(T \geq s)$.

Relationship between the hazard and survival functions

If T is discrete at $0 \leq t_1 < t_2 < ... < t_m < \infty$, then S(t) = $\prod_{\{j:t_j \leq t\}}$ (1-d$\Lambda(t_j)$).
In general, S(t) = $\prod_{(0,t]}(1 - d\Lambda(s))$. Here $\prod_{(0,t]}$ denotes a *product integral*, defined as the limiting product $\prod_{\{j:t_j \leq t\}}$ (1-d$\Lambda(t_j)$) for a sequence of partition intervals of (0, t] of the form $\{(0, t_1], (t_1, t_2], ..., (t_{m-1}, t_m]\}$ such that $max_{\{j:1<j\leq m\}}(t_j - t_{j-1}) \to 0$.

Lecture 2, Notes by Daniel Rubin
January 23, 2004

The hazard and survival functions in the discrete setting

We claimed in the last lecture that S(t) = $\prod_{\{j:t_j \leq t\}}(1 - d\Lambda(t))$ if T is discrete with support on $t_1 < t_2 < ... < t_m < \infty$. Below we prove this assertion.

Recall that $d\Lambda(s) = \frac{dF(s)}{S(s_-)}$. In the discrete setting $S(t_{j-}) = S(t_{j-1})$, $F(t_{j-}) = F(t_{j-1})$, and $dF(t_j) = F(t_j) - F(t_{j-}) = F(t_j) - F(t_{j-1})$. This gives that $(1 - d\Lambda(t)) = 1 - \frac{F(t_j)-F(t_{j-1})}{S(t_{j-1})} = \frac{S(t_{j-1})-F(t_j)+F(t_{j-1})}{S(t_{j-1})} = \frac{1-F(t_{j-1})-F(t_j)+F(t_{j-1})}{S(t_{j-1})} = \frac{1-F(t_j)}{S(t_{j-1})} = \frac{S(t_j)}{S(t_{j-1})}$. Now let t_i be the largest of the $t_1, ..., t_m$ that does not exceed t, and note that $S(t_i) = S(t)$ because T is discrete. Also let $t_0 = 0$ and observe that $S(t_0) = 1$. Finally, we see that $\prod_{\{j:t_j \leq t\}}(1 - d\Lambda(t)) = \prod_{\{j:1\leq j\leq i\}}(1 - d\Lambda(t)) = \prod_{\{j:1\leq j\leq i\}} \frac{S(t_j)}{S(t_{j-1})} = \frac{S(t_i)}{S(t_0)} = S(t_i) = S(t)$.

The hazard and survival functions in the continuous setting

If T is continuous then S(t) = $\exp(-\Lambda(t)) = \exp(-\int_0^t \frac{dF(s)}{S(s)})$. We prove this assertion below.

$\frac{d}{ds} log(S(s)) = -\frac{f(s)}{S(s)} = -\lambda(s)$
$\Longrightarrow log(S(t)) = log(S(t)) - log(1) = log(S(t)) - log(S(0)) = \int_0^t \frac{d}{ds} log(S(s)) ds = \int_0^t -\lambda(s) ds = -\Lambda(t)$
$\Longrightarrow S(t) = \exp(-\Lambda(t))$

Estimation of the hazard

Suppose $T_1, ..., T_n$ are iid and continuous, there is no censoring, and we want to estimate $\lambda(t)$ nonparametrically. We recall that $\lambda(t) = \frac{f(t)}{S(t)}$ because T is continuous, and then use nonparametric density estimation for the numerator and the empirical survival function for the denominator.

There are many ways to perform density estimation, but a popular method is *kernel density estimation*. For this technique we choose a positive number h to be the *bandwidth*, a density K(•) centered at zero to be the *kernel*, and estimate $f(t)$ by $f_{n,h}(t) = \frac{1}{nh}\Sigma_{i=1}^{n} K(\frac{T_i - t}{h})$. Typically the density estimates have less bias but more variance as h is decreased toward zero.

Choice of bandwidth in density estimation

The accuracy of kernel density estimation will depend on the choice of bandwidth. Whether it is preferable use a small or large bandwidth depends on the true density, and *likelihood cross-validation* is a method for adaptively choosing a good bandwidth. Before discussing this further, we must give an aside about inequalities in probability.

Let f_0 denote the true density function of T, and f denote some estimate of this function. We first introduce *Shannon's Inequality*, which states that $\int log(\frac{f(t)}{f_0(t)}) f_0(t)dt \leq log \int \frac{f(t)}{f_0(t)} f_0(t)dt$. This is a special case of *Jensen's Inequality*, which states that if g is concave then $E_{f_0} g(T) = \int g(t) f_0(t)dt \leq g(\int t f_0(t)) = g(E_{f_0} T)$. Shannon's Inequality comes from letting g(t) = log(t), and noting that g(t) is concave.

We next define the *risk* of f as $R(f) = -E_{f_0} log(f(T))$. Now, $R(f_0) - R(f) = \int [log(f(t)) - log(f_0(t))] f_0(t)dt = \int log(\frac{f(t)}{f_0(t)}) f_0(t)dt \leq log \int \frac{f(t)}{f_0(t)} f_0(t)dt = log \int f(t)dt = log(1) = 0$, by Shannon's Inequality, implying that $R(f) \geq R(f_0)$ for any f.

Because the risk is minimized at the true density, our intuition is that the kernel density estimate $f_{n,h}$ best approximating f_0 will be the one with the smallest risk: that is, we would like to choose as bandwidth

$$h_0 = \arg\min_h -int \log f_{n,h}(T) dF_0(T).$$

Since the expectation is unknown, this bandwidth is unknown as well. A seemingly natural way to estimate h_0 is to replace the expectation by the empirical mean:

$$h_{n,naive} = \arg\min -\frac{1}{n} \sum_{i=1}^{n} \log f_{n,h}(T_i).$$

However, it follows that this maximum is achieved at $h = 0$, which is clearly not a good bandwidth. Apparently, estimating the risk of a candidate density estimator $f_{n,h}$ with the empirical risk estimate is not a good idea: this is due to the fact that $f_{n,h}$ is itself already a function of the data.

Likelihood cross-validation deals with this problem by estimating the conditional risk $\int \log f_{n,h}(T) dF_0(T)$ by splitting the sample in a training and validation sample, applying the kernel denisty estimator to the training sample, and estimating its conditional risk only with the validation sample. This so called cross-validated risk estimate provides now a sensible criteria for bandwidht selection, and more general, for selection among any set of candidate density estimators such as model-specific maximum likelihood estimators.

Lecture 3 Notes

Huaxia Qin
January 26, 2004

Data Adaptive Estimation of a Density

Let $T_1, \cdots, T_n$ be i.i.d. observations from distribution density f_0. Given a kernel K, such as $K(x) = I(-1 \leq x \leq 1)/2$, we have as candidate density estimators:

$$f_{n,h}(t) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{T_i - t}{h})$$

Let P_n denote the empirical distribution, i.e. the probability distribution which puts probability $\frac{1}{n}$ on each $T_i, i = 1, \cdots, n$. Estimators can always be viewed as a function of the empirical distribution P_n. In particular, the kernel density estimator can be viewed as the following function of the empirical distribution:

$$\Psi_h(P_n) = \int \frac{1}{h} K(\frac{T - t}{h}) dP_n(T) = f_{n,h}(t).$$

Here, we note that for any function $h(T)$

$$\int h(T) dP_n(T) = \sum_{i=1}^{n} h(T_i) \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} h(T_i)).$$

The estimators $P_n \to \Psi_h(P_n)$, indexed by a bandwidth choice h, are candidate density estimators of f_0.

Definition of the cross-validation selector h_n
Let $B_n \in \{0,1\}^n$ be a random n-dimensional vector.

$$B_n(i) = \begin{cases} 1 & T_i \text{ is put in validation sample} \\ 0 & T_i \text{ is put in training sample} \end{cases}$$

In other words, $\{T_i : B_n(i) = 1\}$ is validation sample and $\{T_i : B_n(i) = 0\}$ is training sample. Let

$$p = \frac{\sum_{i=1}^{n} B_n(i)}{n}$$

denote the proportion of the n observations which are in the validation sample. Let P^0_{n,B_n} be the empirical distribution of training sample, and P^1_{n,B_n} be the empirical distribution of validation sample. The cross-validation selector of h can now be defined as:

$$\begin{aligned} h_n &\equiv arg \min_h E_{B_n} \int -\log[\Psi_h(P^0_{n,B_n})](T) dP^1_{n,B_n}(T) \\ &= arg \min_h \frac{1}{V} \sum_{B \in \{B: B_1, \cdots, B_V\}} -\frac{1}{np} \sum_{i: B_i = 1} \log \Psi_h(P^0_{n,B_n})(T_i) \end{aligned}$$

The idea behind this selector is driven by the fact that

$$f_0 = arg \min_f - \int \log f(T) dF_0(T) = E_{f_0} - log f(T)$$

where $f \rightarrow -\int \log f(T) dF_0(T)$ is called a risk function w.r.t. to the loss function $L(T, f) = -\log f(T)$. Given the bandwidth selection h_n, we estimate f_0 with the corresponding kernel density estimator $\hat{\Psi}_{h_n}(P_n)$, or equivalently,

$$f_{n,h_n}(T) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{T_i - t}{h_n})$$

Some Parametric Models for f_0

(1) $\log(T) \sim N(\mu, \sigma^2)$ μ, σ^2 unknown, log normal model
(2) $T \sim \lambda \exp(-\lambda T)$ λ unknown, exponential model
(3) $P(\log T \leq t) = \frac{1}{1+e^{-(\frac{t-\mu}{\sigma})}}$ μ, σ unknown, log-logistic model. That is, model (1) assumes that $\log T$ is normally distributed, while model (3) assumes that $\log T$ has a logistic distribution.
Models (1) and (3) are very similar, since the maximal difference between a normal density and corresponding logistic density is smaller than 0.02.

Maximum Likelihood Estimator for Parametric Models

Let $x_1, \cdots, x_n$ be i.i.d. observations of $X \sim f_\theta \in \{f_\theta : \theta \in \Theta \subset R^K\}$. $\{f_\theta : \theta \in \Theta \subset R^K\}$ is the parametric model, $\theta = (\theta_1, \cdots \theta_K)$ is a finite (K) dimensional parameter, and Θ is the parameter space.

The maximum likelihood estimator is defined as follows:

$$\theta_n = arg \max_\theta \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(x_i)$$

If the log-likeihood is a differentiable function of θ at the MLE θ_n, then θ_n solves the so called K score equations:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_j} \frac{1}{n} \sum_{i=1}^{n} \log f_{\theta_n + (\delta_1, \cdots, \delta_K)}(x_i)|_{\delta=0} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log f_\theta(x_i)|_{\theta=\theta_n} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \ f_\theta(x_i)|_{\theta=\theta_n} / f_{\theta_n}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} U_j(\theta_n)(x_i) \end{aligned}$$

$U_j(\theta)(x) = \frac{\partial}{\partial \theta_j} \log f_\theta(x)$, $j = 1, \cdots, K$. The K-variate function

$$U(\theta)(x) = (U_1(\theta)(x), \cdots, U_K(\theta)(x))$$

is called the score function. Newton-Raphson Algorithm
Define $H(\theta) = (H_1(\theta), \cdots, H_K(\theta))$, and $\theta = (\theta_1, \cdots, \theta_K)$.

The Newton-Raphson algorithm is an iterative algorithm which aims to solve the K-dimensional equation $H(\theta) = 0$. For example, in order to solve the score equations $1/n \sum_i U(\theta)(X_i) = 0$, one sets $H(\theta) = 1/n \sum_i U(\theta)(X_i)$. The k-th Newton-Raphson step is defined by:

$$\theta^{k+1} = \theta^k - (\frac{\partial}{\partial\theta} H(\theta)|_{\theta=\theta^k})^{-1} H(\theta^k)$$

One starts this algorithm with a choice θ^0, and iterates the above equation till convergence.

For example, in the special case that $H(\theta) = 1/n \sum_i U(\theta)(X_i)$, we have

$$\theta^{k+1} = \theta^k - (\frac{\partial}{\partial\theta} \frac{1}{n} \sum_{i=1}^{n} U(\theta)(x_i)|_{\theta=\theta^k})^{-1} \frac{1}{n} \sum_{i=1}^{n} U(\theta^k)(x_i)$$

It is general practice, to modify the NR algorithm for solving equations $H(\theta) = 0$ with the so called line-search ingredient w.r.t. to a criteria (in the case of score equations, this would be the log-likelihood) one aims to maximize. Namely, one does not necessarily accept the update θ^{k+1}, but first verifies if the update increases the wished criteria. For example, in the case of solving the score equations, we do the following: If $loglik(\theta^{k+1}) \not\leq loglik(\theta^k)$, then take $\alpha\theta^k + (1-\alpha)\theta^{k+1}$ for an α so that $loglik(\alpha\theta^k + (1-\alpha)\theta^{k+1}) > loglik(\theta^k)$.

Lecture 4, January 28, 2004

Properties of ML estimators

Nisha Mulakken

$X \sim f_\theta$, $\theta \subset \Theta \subset R^K$ where θ is $(\theta_1, \theta_2, ..., \theta_k)$

Score function:
$U(\theta)(X) = \frac{d}{d\theta} log f_\theta(X) = (\frac{d}{d\theta_1} log f_\theta(X), ..., \frac{d}{d\theta_k} log f_\theta(X))^t$

Property of Score function:
$E_{f_\theta}(\theta)(X) = 0$

proof:
Let $k = 1$
$\int U(\theta)(X) f_\theta(X) dx$
$= \frac{\int \frac{d}{d\theta} f_\theta(X) dx}{f_\theta(X)} f_\theta(X) dx$
$= \int \frac{d}{d\theta} f_\theta(X) dx$
$= \frac{d}{d\theta} \int f_\theta(X) dx$
$= \frac{d}{d\theta} 1 = 0$

Information Matrix:
$I(\theta)_{kxk} = E_{f_\theta}[U(\theta)(X) U(\theta)(X)^t]$

Thus, $I(\theta)(j,l) = E[U_j(\theta)(X) U_l(\theta)(X)] = \mathrm{COV}_{f_\theta}(U_j(\theta)(X), U_l(\theta)(X))$
We also have that $I(\theta) = -E \frac{d}{d\theta} U(\theta)(X)$

This follows from the identity:
$E_{f_\theta} \frac{d}{d\theta_j} U_l(\theta)(X) = -E_{f_\theta}[U_j(\theta)(X) U_l(\theta)(X)]$

$$\frac{d}{d\theta}[U_1(\theta)...U_k(\theta)]^t = \begin{bmatrix} \frac{d}{d\theta_1}U_1(\theta) & \frac{d}{d\theta_2}U_1(\theta) & \cdots & \frac{d}{d\theta_k}U_1(\theta) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{d}{d\theta_1}U_k(\theta) & \frac{d}{d\theta_2}U_k(\theta) & \cdots & \frac{d}{d\theta_k}U_k(\theta) \end{bmatrix}$$

In other words:
$-E_{f_\theta}[\frac{d^2}{d\theta_{j d\theta_l}} log f_\theta(X)] = I(\theta)(j,l)$

Cramer-Rao Lower Bound for assymptotic variance:

Let θ_{jn} be an uniformly unbiased estimator of θ_j
$E_\theta \theta_{jn} = \theta_j$ for all $\theta \in \Theta$
Then $var(\theta_{jn}) \geq (I(\theta)^{-1}(j,j))/n$ for any uniformly unbiased estimator. A generalization of this statement is the following: $var(a^t\theta_n) \geq a^t I(\theta)^{-1} a$, where $a^\top \theta_n \equiv a_1\theta_{1n} + a_2\theta_{2n} + ... + a_t\theta_{tn}$ denotes a linear combination of the j-specific estimators.

Example estimators:
Consider the parametric family $N(\mu, \sigma^2)$. A possible estimator of the true μ is given by $\mu_n = 2$. This is not a uniformly unbiased estimator
Another estimator is $\mu_n = \overline{X} = \frac{1}{n}\sum_{i=1}^n x_i$, which is a uniformly unbiased estimator:
$E_{\mu,\sigma^2}\overline{X} = \mu$ under all normal densities f_{μ,σ^2}.
Example:
Consider the exponential family of densities $f(x) = \lambda exp(-\lambda x)$ indexed by λ. Given i.i.d. data $x_1, ..., x_n$, the maximum likelihood estimator of λ is given by $\lambda_{MLE} = \frac{1}{\frac{1}{n}\sum_{i=1}^n x_i}$
This is shown as follows: $f_\lambda(x) = \lambda exp(-\lambda x)$
$log f_\lambda(x) = log\lambda - \lambda x$
$U_\lambda(x) = \frac{d}{d\lambda} log f_\lambda(x) = \frac{1}{\lambda} - x$
Score Equation
$\sum_{i=1}^n U_\lambda(x_i) = 0$ is solved by $\lambda_n = \frac{1}{\overline{x}}$, which is thus the ML estimator

Do we have: $E_{f_\lambda}\lambda_n = \lambda$?
$E_{f_\lambda}(\frac{1}{\overline{X}}) = \frac{1}{\lambda}$?
Most likely not, since identities such as $E1/Y = 1/EY$ for a given random variable Y are rare.

Most ML estimators (with one well known exception the MLE of μ in the Normal family) are not unbiased. What we want is for the estimator to be close to the truth. To do this, we want to minimize the mean squared error (MSE). This means both bias and variance are important to consider.

$MSE(\theta_{jn})$
$= E_{f_\theta}(\theta_{jn} - E\theta_j)^2 + E_{f_\theta}(\theta_{jn} - \theta_j)^2$
$= var(\theta_{jn}) + bias^2(\theta_{jn})$

Asymptotically Linear Estimator

data: $x_1, ..., x_n iid \sim f_{\theta,n}$

An estimator θ_n is an asymptotically linear estimator with influence curve $IC(X|\theta)$ where $E_\theta[IC(X|\theta)] = 0$, and $IC(X|\theta) = IC_1(X|\theta), ..., IC_k(X|\theta)$ if :
$\theta_n - \theta = -\frac{1}{n}\sum_{i=1}^n IC[x_i|\theta] + o_p(\frac{1}{\sqrt{n}})$

Definition:
$R_n = o_p(\frac{1}{\sqrt{n}})$ if $\sqrt{n}R_n \to 0$ in probability for n converging to infinity.

In general:
$R_n = o_p(a(n))$ if $\frac{R_n}{a(n)} \to 0$ in probability for n converging to infinity.

Note: $IC(X_i|\theta)$ measures influence of observation X_i on the estimator

Thus,
$\sqrt{n}(\theta_n - \theta) \cong \frac{1}{\sqrt{n}}\sum_{i=1}^n IC[x_i|\theta]$
By the CLT, this converges to the distribution $N(0, \Sigma = E[IC(X|\theta)IC(X|\theta)^t])$
Consequently, the influence curve can be used to estimate Σ and thereby provide confidence intervals and regions for the unknown parameter vector θ.

Influence curves are very convenient for asymptotic inference of estimators. For example, given a k-dimensional vector of influence curves corresponding with a set of k estimators, the influence curve of a (e.g. non-linear) function of this set of estimators is given by the gradient of the function applied to the influence curves. In particular, the difference of two estimators of two parameters has as influence curve the difference of the two estimator-specific influence curves.

Given two asymptotically linear estimators, the relative efficiency of these two estimators is given by the ratio of the variances of their influence curves.

Lecture 5

February 2, 2004

Jessica G. Young

ASYMPTOTICALLY LINEAR ESTIMATORS (CONT'D)

Suppose we have $X_1, ..., X_n$ i.i.d observations of X $\sim f_\theta$: $\theta \in \mathbf{\Theta} \subset \mathbf{R}^k$
θ_n is an asymptotically linear estimator of θ with influence curve

$$IC(X|\theta) = (IC_1(X|\theta), ..., IC_k(X|\theta))$$

if

$$\sqrt{n}(\theta_n - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n IC(X_i|\theta) + o_p(\mathbf{I}) \stackrel{D}{\Longrightarrow} N(0, \Sigma_{kxk} = E[\mathbf{IC}(X|\theta)\mathbf{IC}(X|\theta)^T])$$

We estimate the $k \times k$-covariance matrix Σ by

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{IC}}(X_i)\hat{\mathbf{IC}}(X_i)^T$$

where $\hat{\mathbf{IC}}(X_i)$ is the estimated influence curve. A typical estimate is the substitution estimate obtained by replacing θ by θ_n: $\hat{\mathbf{IC}}(X) = \mathbf{IC}(X|\theta_n)$.

Alternatively,

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{I\hat{C}}(X_i) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{I\hat{C}}(X_i)][\mathbf{I\hat{C}}(X_i) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{I\hat{C}}(X_i)]^T$$

The asymptotic 0.95 confidence interval for θ_j is given by: $\theta_{jn} \pm 1.96\frac{\sqrt{\Sigma_n(j,j)}}{\sqrt{n}}$. That is, the probability that $\theta_j \in [\theta_{jn} - 1.96\frac{\sqrt{\Sigma_n(j,j)}}{\sqrt{n}}, \theta_{jn} + 1.96\frac{\sqrt{\Sigma_n(j,j)}}{\sqrt{n}}] \underset{n\to\infty}{\longrightarrow} 0.95$, which is equivalent with

$$P[-1.96 < \sqrt{n}\frac{(\theta_{jn} - \theta_j)}{\Sigma_n(j,j)} < 1.96] \underset{n\to\infty}{\longrightarrow} 0.95$$

As an aside, we can use the complete estimated multivariate limit distribution $N(0, \Sigma_n)$ to construct a simultaneous confidence interval of the type $\theta_{jn} \pm \delta\frac{\sqrt{\Sigma_n(j,j)}}{\sqrt{n}}$, wherer δ is chosen so that the simultaneous probability that $\theta_j \in \theta_{jn} \pm \delta\frac{\sqrt{\Sigma_n(j,j)}}{\sqrt{n}}$ converges to 0.95 for $n \to \infty$.

Hypothesis testing $\Longrightarrow H_0 : \theta_j = \theta_{j_0}$

For testing $H_0 : \theta_j = \theta_{j_0}$, reject if $\theta_{j_0} \notin \theta_{jn} \pm 1.96\frac{\sqrt{\Sigma_n(j,j)}}{\sqrt{n}}$

Relative Efficiency

Suppose we have θ_{n1}, θ_{n2}, which are two asymptotically linear estimators of $\mu \in \mathbf{R}$ with influence curves $IC_1(X), IC_2(X)$. Let $\sigma_j^2 = \mathrm{VAR}IC_j(X)$, $j = 1, 2$.

Then the asymptotic relative efficiency R of the two estimators is defined as

$$R \equiv \frac{var(IC_1(X))}{var(IC_2(X))} = \frac{\sigma_1^2}{\sigma_2^2}.$$

To interpret this relative efficiency, we consider its relation to the width of the confidence intervals for θ based on these two estimators respectively. The half-width of a 0.95-confidence interval $\theta_{nj} \pm 1.96 * \frac{\sigma_j}{\sqrt{n}}$ based on IC_j is given by $1.96\frac{\sigma_j}{\sqrt{n}}$, $j = 1, 2$. Thus, if one wishes this width to be equal to ε, then one obtains the following corresponding required sample size:

$$1.96\frac{\sigma_1}{\sqrt{n_1}} = \varepsilon \Longrightarrow n_1(\varepsilon) = (1.96\frac{\sigma_1}{\varepsilon})^2$$

$$1.96\frac{\sigma_2}{\sqrt{n_2}} = \varepsilon \Longrightarrow n_2(\varepsilon) = (1.96\frac{\sigma_2}{\varepsilon})^2$$

So the sample size needed to get precision ε will depend on σ_1 and σ_2. Specifically,

$$\frac{n_1(\varepsilon)}{n_2(\varepsilon)} = \frac{\sigma_1^2}{\sigma_2^2} = R$$

For example, if $R = 3$, you would need 3 times the sample size to get the same precision using the 'bad' versus 'good' estimator.

Examples of influence curves using standard δ-method:

Example

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an asymptotically linear estimator of $\mu = EX$ with $IC(X) = (x - \mu)$.

Why is this the influence curve for $\bar{x}$? Because if we take the average of this function over all values of X, we get $\sqrt{n}(\bar{x} - \mu)$; that is we get $\bar{x}$ minus the parameter it is estimating. Specifically,

$$\sqrt{n}(\bar{x} - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC(X_i|\mu) \underset{n\to\infty}{\overset{D}{\Longrightarrow}} N(0, \sigma^2 = var(x - \mu)); \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Example

$X_1, \ldots, X_n \sim f_\lambda(x) = \lambda e^{-\lambda x}$

$\lambda_n = \frac{1}{\bar{X}}$ and $\lambda = \frac{1}{\mu}$, where $\mu = EX$.

First, we can write $\lambda_n - \lambda = f(\bar{x}) - f(\mu) \cong f\prime(\mu) \times (\bar{x} - \mu)$ (Delta method)

The Delta method states that for a given real-valued function $f : \mathbb{R} \to \mathbb{R}$ $f(x+h) - f(x) \cong f\prime(x)h$ if h is small.

This allows us to write $f(x_n) - f(x) \cong f\prime(x)(x_n - x)$ if x_n is close to x. In words, if we have a function of an estimator minus a function of the parameter it's estimating, we can always approximate this difference as the derivative of the function of the parameter (at the true parameter value) multiplied by the difference between the estimator and the truth.

In our example, $\lambda_n = f(\bar{x})$, $\lambda = f(\mu)$, where $f(y) = \frac{1}{y}$. So,

$$f\prime\mu(\bar{x} - \mu) = -\frac{1}{\mu^2}(\bar{x} - \mu) = -\frac{1}{\mu^2}[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)] = \frac{1}{n}\sum_{i=1}^{n} -\frac{1}{\mu^2}(x_i - \mu)$$

We have now shown that $\lambda_n - \lambda$ is an empirical mean of i.i.d. random variables $1/\mu^2(X - \mu)$ in the first order.

Thus, the influence curve of λ_n is given by

$$IC(X|\lambda) = -\frac{1}{\mu^2}(x - \mu) = -\lambda^2(x - \frac{1}{\lambda})$$

so,

$$\sqrt{n}(\lambda_n - \lambda) \overset{D}{\Longrightarrow} N(0, \sigma^2 = var(IC(X|\lambda)))$$

The asymptotic variance of λ_n can be obtained by:

$$\begin{aligned}
\sigma_n^2 &= \frac{1}{n}\sum_{i=1}^{n} IC^2(X_i|\lambda_n) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\lambda_n^2(x_i - \frac{1}{\lambda_n}))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\lambda_n^4(x_i - \frac{1}{\lambda_n})^2
\end{aligned}$$

So $\lambda_n \pm 1.96\frac{\sigma_n}{\sqrt{n}}$ is an asymptotic 0.95 confidence interval for λ.

λ_n is a maximum likelihood estimator and should be efficient so $var(\lambda_n)$ should be equal to $I(\lambda)^{-1}$.

What if we want to estimate $\mu = E(X^k)$?

$\mu_n = \frac{1}{n}\sum_{i=1}^n x_i^k$ so $IC(X|\mu) = x_i^k - \mu$
(this is because, as in the first example, we can write: $\sqrt{n}(\mu_n - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (x_i^k - \mu)$).

So, in general, if $\mu = E[h(X)]$, for some h, then $\mu_n = \frac{1}{n}\sum_{i=1}^n h(x_i)$ is asymptotically linear and $IC(X|\mu) = h(x) - \mu$.

Example

$X_1, \ldots . X_n$ i.i.d X, $\sigma^2 = var(X) = EX^2 - (EX)^2 = \mu_2 - (\mu_1)^2$ where $\mu_j = EX^j, j = 1, 2$

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \mu_{2n} - \mu_{1n}^2$$

where

$$\mu_{jn} = \frac{1}{n}\sum_{i=1}^n x_i^j \qquad \text{for } j = 1, 2$$

Again by the Delta method,

$$\begin{aligned} f(\mathbf{z}+\mathbf{h}) - f(\mathbf{z}) &= \dot{f}(\mathbf{z}) \cdot \mathbf{h} \\ &= \frac{\partial f}{\partial z_1}h_1 + \frac{\partial f}{\partial z_2}h_2 + \ldots + \frac{\partial f}{\partial z_k}h_k \end{aligned}$$

For this example it follows,

$$f(\mu_1, \mu_2) = \mu_2 - \mu_1^2$$

$$\dot{f} = (\frac{\partial f}{\partial \mu_1}, \frac{\partial f}{\partial \mu_2}) = (-2\mu, 1)$$

We can further write,

$$\mu_{1n} - \mu_1 = \frac{1}{n}\sum_{i=1}^n x_i - \mu_1$$

$$\mu_{2n} - \mu_2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \mu_2$$

Thus,

$$\begin{aligned} \sigma_n^2 - \sigma^2 &= f(\mu_{1n}, \mu_{2n}) - f(\mu_1, \mu_2) \\ &= \dot{f}(\mu_{1n} - \mu_1, \mu_{2n} - \mu_2) \\ &= -2\mu_1(\mu_{1n} - \mu_1) + (\mu_{2n} - \mu_2) \\ &= \frac{1}{n}\sum_{i=1}^n [-2\mu_1(x_i - \mu_1) + x_i^2 - \mu_2] \end{aligned}$$

By the above, σ_n^2 is an asymptotically linear estimator of σ^2 with influence curve $IC(X|\mu_1, \mu_2) = -2\mu_1(x - \mu_1) + x^2 - \mu^2$.

We can then get the variance of σ^2 using the techniques illustrated above.

Lecture 6, Sunduz Keles

Xin Zhao, Feb 4,2004

1 Censored Data and Model Selection

1.1 Background

Survival data Survival analysis is the collection of statistical procedures for data analysis for which the outcome variable of interest is *time until an event occurs*. We have see the following type of survival data. $T_1, T_2, ..., T_n$ are i.i.d. observatios of T. T is typically time to occurrence of some event. Events include death, disease, relapse, recovery. Time≡ Survival time. Event≡ failure

In practice, we may not observe each of the T_i's. For example, at the end of the study we might only know that a subject is still alive. That is, the subject's survival time is right-censored. In generally, there are three causes for censoring: 1) a person does not experience the event before *the study ends*; 2) a person is *lost to follow-up* during the study period; 3) a person *withdraws from the study* because of death or some other reason (e.g., adverse drug reaction, car accident). Note that the complete survival time interval has been cut off at the right side, although data can also be left-censored. Most survival data is right-censored. We will consider *right-censored* data in this class.

Right censored data structure Observed Data structure: $O = (\widetilde{T} = \min(T, C), \Delta)$ We observe n i.i.d. observations $(\widetilde{T}_1, \Delta_1), ..., (\widetilde{T}_n, \Delta_n)$ of O. We note that the observed data structure is a function of (T, C). We will denote the marginal distribution functions of T and C with F and G, respectively.

1.2 Model

Full Data Model: Model for F. is called a *full data model.*
Censoring Mechanism model: Model for conditional distribution G of $C \mid T$ is called the *censoring mechanism* or *conditional censoring distribution.*

Note: You can imagine that assumptions on full data model and censoring mechanism have a tremendous effect on the inference problems that we face.

Example 1, Exponential density Assume that T follows an exponential distribution, and that C is independent of T with unspecified marginal distribution. We will now compute the likelihood for one observation:

$$P(\widetilde{T} = t, \Delta = 1) \quad = \quad P(T = t, C > t) = P(T = t)P(c > t) = f_\lambda(t)\overline{G}(t),$$

where $f_\lambda(t) = P(T = t)$, and $\overline{G}(t) = 1 - G(t)$. Similarly,

$$P(\widetilde{T} = t, \Delta = 0) = P(T > t, C = t) = g(t)S_\lambda(t).$$

Thus the likelihood for n observations is given by:

$$L(\lambda, G \mid (\widetilde{T}_i, \Delta_i), i = 1, ..., n) = \Pi_{i=1}^n \left[f_\lambda(\widetilde{T}_i)\overline{G}(\widetilde{T}_i)\right]^{\Delta_i} \left[S_\lambda(\widetilde{T}_i)g(\widetilde{T}_i)\right]^{1-\Delta_i}.$$

Thus, the loglikelihood for n observations is given by:

$$\log L(\lambda, G \mid (\widetilde{T}_i, \Delta_i), i = 1, ..., n) = \sum_{i=1}^n \Delta_i \log f_\lambda(\widetilde{T}_i) + \sum_{i=1}^n (1-\Delta_i) \log f_\lambda(\widetilde{T}_i) + \sum_{i=1}^n (1-\Delta_i) \log g(\widetilde{T}_i) + \sum_{i=1}^n \Delta_i \log \overline{G}(\widetilde{T}_i).$$

The goal is to estimate the exponential parameter λ. We need to solve $\frac{d}{d\lambda}\log L(\lambda, G \mid (\widetilde{T}_i, \Delta_i), i = 1, ..., n) = 0$ w.r.t. λ. Recall $f_\lambda(\widetilde{T}_i) = \lambda \exp(-\lambda \widetilde{T}_i)$, $S_\lambda(\widetilde{T}_i) = \exp(-\lambda \widetilde{T}_i)$. Thus, the solution of the score equations is given by: $\lambda_n = \frac{\sum_i \Delta_i}{\sum_i \widetilde{T}_i}$.

We also note that the information is given by $I(\lambda) = \frac{P(\Delta=1)}{\lambda^2}$. Thus, the Cramer-Rao lower bound is given by: $I(\lambda)^{-1} = \frac{\lambda^2}{P(\Delta=1)}$. A simple estimate of the information matrix is given by:

$$i(\lambda n) = \frac{\frac{1}{n}\sum_i \Delta_i}{\lambda n^2}.$$

Example, Histogram regression with uncensored survival data The full data structure is now $X = (T, W)$, where W is a vector of baseline covariates. The parameter of interest is the conditional expectation:

$$\Psi_0 = E[\log T \mid W], \text{ or } \Psi_0 = E[T \mid W].$$

Suppose that we observe n i.i.d. observations of X.

Construction of histogram regression estimators of Ψ_0: Let K be the number of bins. For each bin, the estimator of the mean survival time is the empirical mean of the survival times with covariates W in that bin. Give the number of bins k, we have now a defined a histogram regression estimator of ψ_0 which we will denote with $\Psi_k(P_n)$, $k = 1, ..., K$.

How do we choose the number of bins? Choose among estimators $\Psi_1(P_n), ..., \Psi_k(P_n)$?

Cross-validation. Let $L(X, \Psi) = [T - \Psi(w]^2$. Define $B_n \in (0, 1)^n$ as an n dimensional random vector. Let $\{i : \ B_n(i) = 1\}$ be the validation sample, and let $\{i : \ B_n(i) = 0\}$ be the training sample. Let P^0_{n,B_n}, P^1_{n,B_n} be the empirical distributions of the training sample and validation sample, respectively.

If one would observe $X_1, \ldots, X_n$, then a cross-validated conditional risk estimate of the estimator $P_n \to \Psi_k(P_n)$ is defined as:

$$\widehat{\theta}_{n(1-p)}(k) = E_{B_n}[\frac{1}{np}\sum_i I(B_n(i) = 1)L(X_i, \Psi_k(\cdot \mid P^0_n B_n))].$$

PH 240B Notes For Feb 9th 2004, Srikesh G. Arunajadai

Theorem: Asymptotic Linearity of Maximum Likelihood Estimate (MLE)

Let $X_1, \ldots, X_n$ be i.i.d sample from the distribution $X \sim f_{\theta_0}$, $\theta \in \Theta \subset R^k$ where Θ is the parameter space and θ_0 denotes the true parameter value.
Let θ_n be the MLE:

$$\theta_n = argmax_\theta \sum_{i=1}^{n} log f_\theta(x)$$

The influence curve of θ_n is given by

$$I(\theta_0)^{-1} U(\theta_0)(X_i)$$

where

$$\begin{aligned} I(\theta_0)(x) &= E\left[\frac{\partial^2}{\partial\theta^2} log f_\theta(x)\right]\Big|_{\theta=\theta_0} && (1) \\ &= E\left[U(\theta)(X)U(\theta)(X)^T\right] && (2) \end{aligned}$$

is the information matrix of dimension $k \times k$ and

$$U(\theta)(X) = \frac{\partial}{\partial \theta} log f_\theta(X)$$

Under regularity conditions

$$\theta_n - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} I(\theta_0)^{-1} U(\theta_0)(X_i) + o_p\left(\frac{1}{\sqrt{n}}\right)$$

Hence

$$\sqrt{n}(\theta_n - \theta_0) \Rightarrow N\left(0, \sum = I(\theta)^{-1}\right)$$

for $n \to \infty$: that is, the $\sqrt{n}$-standardized difference $\theta_n - \theta_0$ converges in distribution to the normal distribution with mean zero and variance equal to the variance of the influence curve $I(\theta_0)^{-1} U(\theta_0)(X)$ of θ_n.

Note

$$\begin{aligned} E\left[I(\theta_0)^{-1} U(\theta_0)(X)\right]\left[I(\theta_0)^{-1} U(\theta_0)(X)\right]^T &= I(\theta_0)^{-1} E\left[U(\theta_0)(X) U(\theta_0)(X)^T\right] I(\theta_0)^{-1} && (3) \\ &= I(\theta_0)^{-1} I(\theta_0) I(\theta_0)^{-1} && (4) \\ &= I(\theta_0)^{-1} && (5) \end{aligned}$$

Proof

Consistency of θ_n

$$\begin{aligned} 0 &\leq \int log\left(\frac{f_{\theta_0}(x)}{f_{\theta_n}(x)}\right) dP_{\theta_0}(x) && (6) \\ 0 &\leq \int log\left(\frac{f_{\theta_0}(x)}{f_{\theta_n}(x)}\right) d(P_{\theta_0} - P_n)(x) + \int log\left(\frac{f_{\theta_0}(x)}{f_{\theta_n}(x)}\right) dP_n(x), && (7) \end{aligned}$$

where P_θ denotes the probability distribution corresponding with the density f_θ so that $dP_\theta(x) = f_\theta(x) d\lambda(x)$ in the case that f_θ denotes a density of P_θ w.r.t. a measure λ.

The function $\log\left(\frac{f_{\theta_0}(x)}{f_{\theta_n}(X)}\right)$ is not a fixed function of X but a random function of X. Hence we cannot use the law of large numbers to prove that the first term converges to zero in probability. However, we can bound the first term by the supremum over the collection of functions $\{X \to \log\left(\frac{f_{\theta_0}(X)}{f_\theta(X)}\right) : \theta\}$, and then employ a uniform law of large numbers, as established in empirical process theory (e.g., van der Vaart, Wellner, 1996). Also the second term is less then or equal to zero as it can be written as follows.

$$\int log f_{\theta_0}(x) dP_n(x) - \int log f_{\theta_n}(x) dP_n(x) \leq 0$$

as the second integral uses the maximum likelihood estimate. Hence, we have

$$\begin{aligned} 0 &\leq \int log\left(\frac{f_{\theta_0}(x)}{f_{\theta_n}(x)}\right) d(P_{\theta_0} - P_n)(x) + \int log\left(\frac{f_{\theta_0}(x)}{f_{\theta_n}(x)}\right) dP_n(x) && (8) \\ 0 &\leq Sup_{\theta \in \Theta} \left| \int log\left(\frac{f_{\theta_0}(x)}{f_\theta(x)}\right) d(P_{\theta_0} - P_n)(x) \right|. && (9) \end{aligned}$$

The latter term can be handled by empirical process theory.

Definition: A class $F = \{f : x \to \mathrm{R}\}$ of real valued functions of X is called a uniform Glivenco-Catelli (GC) class if

$$Sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} [f(x_i) - Ef(x)] \right| \to 0$$

converges in probability to 0 when $n \to \infty$
Empirical process theory provides many examples of such classes. Thus if

$$\mathcal{F} \equiv \left\{ x \to log \left(\frac{f_{\theta_0}(x)}{f_{\theta_n(x)}} \right) : \theta \in \Theta \right\}$$

is a GC class then we have shown that

$$d_{KL}\left(f_{\theta_n}, f_{\theta_0}\right) \equiv \int \log \frac{f_{\theta_n}(X)}{f_{\theta_0}(X)} dP_{\theta_0}(X) \to 0$$

when

$$n \to \infty$$

.

Example Given a function $f : \mathrm{R}^2 \to \mathrm{R}$ we define the uniform sectional variation norm as follows

$$\parallel f \parallel = Max \left[sup_{X_2} \int |df(\partial X_1, X_2)|, sup_{X1} \int |df(X_1, \partial X_2)|, \int |f(X_1, X_2)| \right].$$

If there exists a M such that

$$\mathcal{G} = \{f : \parallel f \parallel_v \leq M < \infty\}$$

then $\mathcal{G}$ is a GC class.

Thus, for example, if each of the functions in $\mathcal{F}$ has uniform sectional variation norm bounded by a universal constant $M < \infty$, then we have proved $d_{KL}(f_{\theta_n}, f_{\theta_0}) \to 0$ in probability for $n \to \infty$. Since the $L_1(P_{\theta_0})$-norm $\|f_{\theta_n} - f_{\theta_0}\|_{\theta_0,1} \equiv \int \mid f_{\theta_n} - f_{\theta_0} \mid (x) dP_{\theta_0}(x)$, and $L_2(P_{\theta_0})$-norm $\|f_{\theta_n} - f_{\theta_0}\|_{\theta_0,2} \equiv \sqrt{\int (f_{\theta_n} - f_{\theta_0})^2 dP_{\theta_0}}$ can be bounded by the Kullback-Leibler divergence $d_{KL}(f_{\theta_n}, f_{\theta_0})$, under the condition that f_{θ_0} is uniformly bounded away from zero, the converges in Kullback-Leibler divergence implies the convergence L_1 and L_2 norm of f_{θ_n} to f_{θ_0}. This proves consistency of f_{θ_n} to f_{θ_0}. In order to prove consistency of θ_n to θ_0 one needs to be able to write θ as a continuous function of f_θ.

Lecture of February 11, 2004, Oliver Bembom

Asymptotic Linearity of the MLE θ_n

Let $X_1, ..., X_n$ be i.i.d. with $X \sim P_{\theta_0}$. Let $H(\theta, P) \equiv E_p[U(\theta)(X)] \in \mathbb{R}^k$, with $U(\theta)(X) = \frac{d}{d\theta} \log f_\theta(X) \in \mathbb{R}^k$. Then we know:

$$(1) \qquad H(\theta_0, P_{\theta_0}) = E_{\theta_0}[U(\theta_0)(X)] = \int \frac{d}{d\theta_0} \log f_{\theta_0}(X) dF_\theta(X) = \int \frac{\frac{d}{d\theta_0} f_{\theta_0}(X)}{f_{\theta_0}(X)} f_{\theta_0}(X) dX$$

$$= \frac{d}{d\theta_0} \int f_{\theta_0}(X) dX = \frac{d}{d\theta_0} 1 = 0$$

$$(2) \qquad H(\theta_n, P_n) = \frac{1}{n} \sum_{i=1}^{n} U(\theta_n)(X_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\theta_n} \log f_{\theta_n}(X_i) = \frac{1}{n} \left[\frac{d}{d\theta_n} \sum_{i=1}^{n} \log f_{\theta_n}(X_i) \right] = 0$$

where (2) follows since θ_n is defined as that $\theta \in \Theta$ that maximizes $\sum_{i=1}^{n} \log f_\theta(X_i)$. An estimator that is defined as the solution under the empirical distribution of an equation that holds for the true θ_0 under f_{θ_0} is called an M-estimator. MLE are thus part of this class of M-estimators. The proof of asymptotic linearity given here applies in fact to any M-estimator.Using (1) and (2), we get

$$(3) \qquad H(\theta_n, P_{\theta_0}) - H(\theta_0, P_{\theta_0}) = - \left[H(\theta_n, P_n) - H(\theta_n, P_{\theta_0}) \right]$$

We will use a first-order Taylor series expansion for the term on the left, and results from empirical process theory for the term on the right. Since $H : \mathbb{R}^k \to \mathbb{R}^k$, we review the definition of the derivative in this setting: We say that $g : \mathbb{R}^k \to \mathbb{R}^k$ is differentiable at $a \in \mathbb{R}^k$ if there exists a linear transformation $\dot{g} : \mathbb{R}^k \to \mathbb{R}^k$ such that

$$\lim_{x \to a} \frac{\|g(x) - g(a) - \dot{g}(a)(x - a)\|}{\|x - a\|} = 0$$

where $\|x\| = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$ is the Euclidean norm. Note that $\dot{g}$ can represented as a $k \times k$ matrix. Note also that this definition implies that

$$g(x) - g(a) = \dot{g}(a)(x - a) + o(\|x - a\|)$$

i.e. $g(x) - g(a)$ can be written as a linear approximation plus a remainder term which tends to 0 even when divided by $\|x - a\|$.In our case, we have $H : \mathbb{R}^k \to \mathbb{R}^k$ with

$$\frac{d}{d\theta} H(\theta, P_{\theta_0}) \Big|_{\theta=\theta_0} = \frac{d}{d\theta} E_{P_{\theta_0}} \left[U(\theta)(X) \right] \Big|_{\theta=\theta_0} = E_{P_{\theta_0}} \left[\frac{d}{d\theta} U(\theta)(X) \right] \Big|_{\theta=\theta_0} = -I(\theta_0)$$

Using that $\theta \to H(\theta, P_{\theta_0})$ is thus differentiable at θ_0 we can write

$$H(\theta_n, P_{\theta_0}) - H(\theta_0, P_{\theta_0}) = \frac{d}{d\theta} H(\theta, P_{\theta_0})\Big|_{\theta=\theta_0} (\theta_n - \theta_0) + o(\|\theta_n - \theta_0\|) = -I(\theta_0)(\theta_n - \theta_0) + o(\|\theta_n - \theta_0\|)$$

Substituting this into the left-hand side of (3) and writing out the right-hand side, we get

$$-I(\theta_0)(\theta_n - \theta_0) = -\frac{1}{n} \sum_{i=1}^{n} \left[U(\theta_n)(X_i) - \int U(\theta_n)(X) dP_{\theta_0}(X) \right] + o(\|\theta_n - \theta_0\|)$$

Assuming that $I^{-1}(\theta_0)$ exists, we have that

$$\theta_n - \theta_0 = I^{-1}(\theta_0) \left[\frac{1}{n} \sum_{i=1}^{n} \left[U(\theta_n)(X_i) - \int U(\theta_n)(X) dP_{\theta_0}(X) \right] \right] + o(\|\theta_n - \theta_0\|)$$

For the term inside the outer square brackets, we use the following result from empirical process theory:

If $\mathcal{F} = \{X \to U(\theta)(X) : \theta \in \mathbf{\Theta}\}$ is a so-called Donsker class (this is for example true if $\exists\, M < \infty$ such that $\forall\, \theta \in \mathbf{\Theta} :\ \|U(\theta)\|_v \leq M$) and

$$\int \left[U(\theta_n)(X) - U(\theta_0)(X) \right]^2 dP_{\theta_0}(X) \longrightarrow 0 \text{ in probability}$$

then

$$\frac{1}{n} \sum_{i=1}^{n} \left[U(\theta_n)(X_i) - \int U(\theta_n)(X) dP_{\theta_0}(X) \right] = \frac{1}{n} \sum_{i=1}^{n} \left[U(\theta_0)(X_i) - \int U(\theta_0)(X) dP_{\theta_0}(X) \right] + o_p\left(\frac{1}{\sqrt{n}} \right)$$

Note that the proof of consistency of the MLE from last lecture implies the second assumption of this result. Hence under the assumption that $\mathcal{F}$ is a Donsker class, we have

$$\theta_n - \theta_0 = I^{-1}(\theta_0) \left[\frac{1}{n} \sum_{i=1}^{n} \left[U(\theta_0)(X_i) - \int U(\theta_0)(X) dP_{\theta_0}(X) \right] \right] + o_p\left(\frac{1}{\sqrt{n}} \right) + o(\|\theta_n - \theta_0\|)$$

$$= \frac{1}{n}\sum_{i=1}^{n} I^{-1}(\theta_0)\left[U(\theta_0)(X_i) - E_{P_{\theta_0}}U(\theta_0)(X)\right] + o_p\left(\frac{1}{\sqrt{n}}\right) + o(\|\theta_n - \theta_0\|)$$

$$= \frac{1}{n}\sum_{i=1}^{n} I^{-1}(\theta_0)U(\theta_0)(X_i) + o_p\left(\frac{1}{\sqrt{n}}\right) + o(\|\theta_n - \theta_0\|)$$

In order to prove asymptotic linearity of the MLE, we need to show that the last term $o(\|\theta_n - \theta_0\|)$ is also $o_p\left(\frac{1}{\sqrt{n}}\right)$. Note that the above implies that

$$\theta_n - \theta_0 = O_p\left(\frac{1}{\sqrt{n}}\right) + o_p\left(\frac{1}{\sqrt{n}}\right) + o(\|\theta_n - \theta_0\|) = O_p\left(\frac{1}{\sqrt{n}}\right) + o(\|\theta_n - \theta_0\|)$$

since the asymptotic normal distribution of

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} I^{-1}(\theta_0)U(\theta_0)(X_i)\right]$$

is bounded in probability. This implies that

$$\theta_n - \theta_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$$

and hence that

$$o(\|\theta_n - \theta_0\|) = o_p\left(\frac{1}{\sqrt{n}}\right)$$

Thus we have proved that

$$\theta_n - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} I^{-1}(\theta_0)U(\theta_0)(X_i) + o_p\left(\frac{1}{\sqrt{n}}\right)$$

Hence the MLE is asymptotically linear with influence curve given by

$$IC(X) = I^{-1}(\theta_0)U(\theta_0)(X)$$

Explanation of o_p and O_p notation

1. We say that a sequence of real numbers $f(n)$ is $o(1)$ if $lim_{n\to\infty} f(n) = 0$, i.e. if

$$\forall\, \epsilon > 0 \quad \exists\, N :\, \forall n \geq N \quad |f(n)| < \epsilon$$

   We say that a sequence of real numbers $f(n)$ is $o\left(g(n)\right)$ if the sequence $\frac{f(n)}{g(n)}$ is $o(1)$.

2. We say that a sequence of real numbers $f(n)$ is $O(1)$ if $f(n)$ is bounded, i.e. if

$$\exists\, M < \infty : \quad \forall n \quad |f(n)| \leq M$$

   We say that a sequence of real numbers $f(n)$ is $O\left(g(n)\right)$ if the sequence $\frac{f(n)}{g(n)}$ is $O(1)$.

3. We say that a sequence of random variables $R(n)$ is $o_p(1)$ if $R(n)$ tends to 0 in probability as $n \to \infty$, i.e. if

$$\forall\, \delta \quad \forall\, \epsilon \quad \exists\, N : \quad \forall\, n \geq N \quad P(|R(n)| > \epsilon) < \delta$$

   A sequence of random variables $R(n)$ is $o_p\left(h(n)\right)$ if the sequence $\frac{R(n)}{h(n)}$ is $o_p(1)$.

4. We say that a sequence of random variables $R(n)$ is $O_p(1)$ if it is bounded in probability, i.e. if

$$\limsup_n P(|R(n)| > M) \to 0 \text{ as } M \to \infty$$

   We say that a sequence of random variables $R(n)$ is $O_p\left(h(n)\right)$ if the sequence $\frac{R(n)}{h(n)}$ is $O_p(1)$.

Simultaneous confidence regions for θ_0

Definition: Let $A = Q\Lambda Q^T$ be the spectral decomposition of an n × n matrix A (Q is orthonormal and Λ diagonal). Then define the square-root of A as

$$A^{\frac{1}{2}} = Q\Lambda^{\frac{1}{2}}Q^T$$

Note:

$$A^{\frac{1}{2}}A^{\frac{1}{2}} = Q\Lambda^{\frac{1}{2}}Q^TQ\Lambda^{\frac{1}{2}}Q^T = Q\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}Q^T = Q\Lambda Q^T = A$$

Suppose $\sqrt{n}(\theta_n - \theta_0) \Longrightarrow N(\mathbf{0}, \mathbf{\Sigma})$ in distribution. This implies that

$$\frac{\sqrt{n}(\theta_n - \theta_0)}{\sigma_n} \Longrightarrow N(\mathbf{0}, \rho) \text{ in distribution}$$

where σ_n is an estimate of $\sqrt{\Sigma(j,j)}$ for $j = 1, ..., n$, and ρ is the correlation matrix corresponding to $\mathbf{\Sigma}$. Let $\rho^{-\frac{1}{2}}$ be the square-root of ρ^{-1}. Then

$$\rho^{-\frac{1}{2}}\left(\frac{\sqrt{(}n)(\theta_n - \theta_0)}{\sigma_n}\right) \Longrightarrow N(\mathbf{0}, \mathbf{I}) \text{ in distribution}$$

Proof: Let $\mathbf{Z}$ be a random n-vector with covariance matrix $\mathbf{\Sigma}$ and $\mathbf{A}$ be an $n \times n$ matrix of real numbers. Then

$$cov(\mathbf{AZ}) = E\left[\mathbf{AZZ}^T\mathbf{A}^T\right] = \mathbf{A\Sigma A}^T$$

In particular, if $\mathbf{Z} \sim N(\mathbf{0}, \rho)$ then $\mathbf{AZ} \sim N(\mathbf{0}, \mathbf{A}\rho\mathbf{A}^T)$. For $\mathbf{A} = \rho^{-\frac{1}{2}}$ we get

$$cov(\rho^{-\frac{1}{2}}\mathbf{Z}) = \rho^{-\frac{1}{2}}\rho(\rho^{-\frac{1}{2}})^T = \rho^{-\frac{1}{2}}\rho^{\frac{1}{2}}\rho^{\frac{1}{2}}\rho^{-\frac{1}{2}} = \mathbf{I}$$

Thus $\rho^{-\frac{1}{2}}\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$.

Class Notes: February 18, 2004, Merrill Birkner

$X_1, \ldots, X_n$ i.i.d. $X \sim f_{\theta_0}$ $\theta_0 \in \theta \subset R^k$
$\sqrt{n}(\theta_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IC(X_i|\theta_0) + o_p(1) \overset{D}{\Rightarrow} N(0, \Sigma_0 = E[IC(X|\theta_0)IC(X|\theta_0)^T])$
$\frac{\sqrt{n}(\theta_n - \theta_0)}{\sigma_n} \overset{D}{\Rightarrow} N(0, \rho_0 = correlation(\Sigma))$ where $\sigma_n = v\hat{a}r(IC(X|\theta_0))$

The above is a standardization, by dividing by the standard error.

Elliptical Confidence Region

Let ρ_0^{-1} be the inverse of ρ_0

Let $\rho_0^{-1} = TDT^T$ be the eigenvalue decomposition:

where T is a matrix of eigenvectors and D is a diagonal matrix.

Define $\rho_0^{-1/2} = T\sqrt{D}T$

** It is the square root of ρ_0^{-1}. Then, $\rho_0^{-1/2}(\frac{\sqrt{n}(\theta_n - \theta_0)}{\sigma_n}) \overset{D}{\Rightarrow} N(0, I)$. This implies that:

$\Pr(||\rho_0^{-1/2}(\frac{sqrtn(\theta_n - \theta_0)}{\sigma_n})||^2 \leq \mathcal{X}^2_{k,0.95}) \rightarrow 0.95$,

where $\mathcal{X}^2_{k,0.95}$ is the 0.95 quantile of a $\mathcal{X}^2_k$.

**and the Euclidian norm is defined as follows: $||x|| = \sqrt{\sum_{j=1}^{k} x_j^2}$

Rectangular 0.95 Confidence Region

After standardization, we are looking for a constant **a** such that:

$Pr(Max_j|\frac{\sqrt{n}(\theta_n - \theta_0)}{\sigma_n}| < a) \overset{n \rightarrow \infty}{\rightarrow} 0.95$. Then, $Pr(\theta_{0j} \in (\theta_{nj} \pm \frac{a\sigma_{nj}}{\sqrt{n}}) \forall j = 1...k) \rightarrow 0.95$.

Therefore, components with large standard errors will have wider confidence intervals.

Thus,

$\theta : \theta_j \in (\theta_{nj} \pm \frac{a\sigma_{nj}}{\sqrt{n}}), j = 1...k \rightarrow 0.95$.

This is a 0.95 Confidence Region for θ_0

How can we get this? We can simulate from $N(0, \rho_0)$ distribution:

By Simulation:

Choose **a** equal to the 0.95 quantile ($q_{0.95}$) of the $E \equiv Max_{j=1..k}|W_j|$, where $W \sim N(0, \rho_0)$

Simulate 20,000 vectors (size k) from $N(0, \rho_0)$

Take the maximum value of each vector; therefore you end up with 20,000 maximum values.

Now you want to take the 0.95 quantile of this vector, length 20,000, of maximums.

We can do this by:

1. Simulate N=20,000 $\vec{W}_1, \vec{W}_2, ..., \vec{W}_N \sim N(0, \rho_0)$
2. Create N: $E_1 = Max_j|W_{1j}|, E_2 = Max_j|W_{2j}|, ...E_N = Max_j|W_{Nj}|$
3. Compute 0.95 quantile of $E_1....E_N$. This is $q_{0.95}$

How to Simulate from a MVN distribution.

Simulating from $N(0, \rho_0)$. Let U be such that $U^T U = \rho_0$. U can be chosen to be equal to the Choleski decomposition.

**Note: In R (chol(ρ_0))

Then $UZ \sim N(0, UU^T = \rho_0)$ where $Z \sim N(0, I)$

An Application:

Suppose you want to find the confidence region of a survival function with probability 0.95. You want a simultaneous confidence region. With probability 0.95 you want the entire survival curve in that 'envelope'

Example: $T_1, ..., T_n$ i.i.d. $T \sim f_0$

$S_0(t) = P(T > t)$

Let $\theta_0 = (S_0(t_1)....S_0(t_k))$ be the parameter vector of interest.

How to estimate this?
$\theta_n = (S_0(t_1)....S_0(t_k))$ where $S_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t)$
$\theta_n - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} IC(T_i|\theta_0)$
where $IC(T|\theta_0) = [I(T > t_1) - S_0(t_1), ...I(T > t_k) - S_0(t_k)]^T$
note: the vector IC has no remainder.
Thus: $\sqrt{n}(\theta_n - \theta_0) \stackrel{D}{\Rightarrow} N(0, \Sigma_0 = E[IC(T|\theta_0)IC(T|\theta_0)^T])$
How to estimate the influence curve? First, estimate the influence curve: $\hat{IC}(T) = IC(T|\theta_n)$.
$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} \hat{IC}(T_i)\hat{IC}(T_i)^T$.

Remember that: $IC(T|\theta_n) = [I(T > t_1) - S_0(t_1), ...I(T > t_k) - S_0(t_k)]^T$ For every person you are going to create an IC vector and take a sample standard covariance (Σ_n)
Then you will plot θ_n: the survival curve.
To find the cut-off **a** we need to standardize (and therefore work with a common cut-off)
ρ_{0n}= correlation matrix corresponding to Σ_n
thus, $\frac{\sqrt{n}(\theta_n - \theta_0)}{\sigma_n} \sim N(0, \rho_{0n})$ where $\sigma_n^2 = var(\hat{IC}(T))$ or the diagonal of Σ_n
In order to find the constant, we need the 0.95 quantile ($\hat{q}_{0.95}$) of:
$Max_j|W_j|$ where $W \sim N(0, \rho_{0n})$

We then want to use the Cholesky decomposition of ρ_{0n} (described earlier in notes)
$UZ \sim N(0, \rho_{0n})$: we want to find U
You will come up with a number for **a**. For example $\hat{q}_{0.95} = 3$
In a point wise calculation, this value would be equal to 1.96 (since that is the 0.95 quantile of a N(0,I).

Thus, $\theta_{nj} \pm \hat{q}_{0.95}\frac{\sigma_{nj}}{\sqrt{n}}$ is a simultaneous confidence region. (j=1...k)
One can plot the survival function (Survival versus Time) and then plot the simultaneous confidence 'envelope' around the estimated survival function (θ_n or S_n). The confidence region should be closer to θ_n on the two extremes of the plot and further away in the middle of the survival function. Since it is Bernoulli, it will not be equally placed and therefore thinner on the ends. The survival curve will lay in the envelope with probability 0.95.

Lecture 14, Kathryn Steiger, February 27, 2004.

General Approach for Construction of an Estimator

Our Model

We observe $X_1, \ldots, X_n$ i.i.d. observations of $X \sim P \in \mathcal{M}$.

The parameter of interest is a function of the actual distribution of the data. In other words, a parameter of the distribution of the data mapped from the model to the outcome space.

$\theta \circ \mathcal{M} \to D$ (e.g., D might be the euclidean space)

So in particular, we can ask – what is the parameter value of the Truth?

$$\theta_0 = \theta(P_0) \tag{10}$$

But we already have a mapping from the data generating distribution to the truth so we can use the empirical distribution.

Let P_n be the empirical distribution. Let ϕ be a mapping defined on any P, in particular P_n s.t.

$$\phi(P) = \theta(p), \qquad P \in \mathcal{M}$$

We wish to find an extension of that parameter s.t. it applies to the empirical distribution. Now we can estimate θ_0 with substitution estimator

$$\theta_n = \phi(P_n) \tag{11}$$

Example

Suppose $X \sim f_{\theta_0}$, $\theta_0 \in \Theta \subset \mathbb{R}^K$. For model identifiability, let θ_0 be a parameter of interest of f_{θ_0}. Define $\theta(P) = \arg\max_{\theta\in\Theta} \int Log f_\theta(x) dP(x)$.

Note: if we plug in

$$\phi(P_{\theta_0}) = \arg\min_{\theta\in\Theta} \int Log f_\theta(x) dP_{\theta_0}(x) \tag{12}$$

this goes to $f_{\theta_0} = \Theta_0$ (Recall $P(A) = P(X \in \mathcal{A}) = \int_{\mathcal{A}} f(x)dx$)

We can also apply to the empirical

$$\phi(P_n) = \arg\max_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} Log f_\theta(X_i) \tag{13}$$

Log Likelihood – Maximum Likelihood

There are usually many possible *phi*'s outside of the model. The moment we put P into the model it's just the mean.

Example

$X \sim P_0 \in \mathcal{M}$. $\theta(P_0) = E_{P_0}\phi(P) = \int x dP(x)$, $\theta_n = \phi(P_n) = \int X\, dP_n(x)$, $\theta_n = \bar{X}$.

Right Censored Data: Kaplan-Meier as a substitute estimator

We observe:

$$O_i = (min(T, W)), \Delta_i = I(T \le C)); \qquad T \perp C, \quad \tilde{T} \sim F_0, \quad C \sim G_0 \tag{14}$$

We wish to estimate:

$$\theta_0 = S_0(t) = 1 - F_0(t) \tag{15}$$

We want to find a ϕ s.t. if applied to the data we get back the survival function.

$$\phi(P_{F_0,G_0}) = S_0(t) \tag{16}$$

[Question: Is this a parameter at this point? – No – that would be true only in the Full data model. Here we have Right Censored data.]

We start with the following:

$$S_0(t) \quad = \quad \prod_{s\in(0,t)} \left(\frac{1 - P(T \in [s, s + ds])}{P(T \ge s, C \ge s)} \right) \quad By : C \perp T$$

$$
\begin{aligned}
S_0(t) &= \prod_{s\in(0,t)} \left(\frac{1 - P(T \in [s, s+ds], C \geq s), c \geq s}{P(T \geq s, c \geq s)} \right) \\
S_0(t) &= \prod_{s\in(0,t)} \left(\frac{1 - P_{F_0G_0}(\tilde{T} \in [s, s+ds], \Delta = 1),}{P_{F_0G_0}(\tilde{T} \geq s)} \right)
\end{aligned}
$$

Now we have written the survival function in terms of the observed data. Next we plug in the empirical by defining some subdistributions.

$$
\begin{aligned}
P_1(t) &= P(\tilde{T} \leq t, \Delta = 1) \\
P_0(t) &= P(\tilde{T} \leq t, \Delta = 0)
\end{aligned}
$$

So we have:

$$
\begin{aligned}
S_0(t) &= \prod_{s\in(0,t)} \left(1 - \frac{P_1(ds)}{1 - [(P_0 + P_1)(s)]} \right) \\
S_0(t) &= \phi(P_0, P_1) \\
&\quad \text{Survival function as } \phi \text{ of distribution of the data.} \\
P_{1_n}(t) &= \frac{1}{n} \sum_{i=1}^{n} I(\tilde{T}_i \leq t, \Delta_i = 1) \\
P_{0_n}(t) &= \frac{1}{n} \sum_{i=1}^{n} I(\tilde{T}_i \leq t, \Delta_i = 0)
\end{aligned}
$$

Now our estimator is :

$$
\begin{aligned}
S_n(t) &= \phi(P_{0_n}, P_{1_n}) \\
S_n(t) &= \prod_{s(0,t)} \left(\frac{1 - [\frac{1}{n} \sum_{i=1}^{n} I(\tilde{T}_i = s, \Delta_i = 1)]}{\frac{1}{n} \sum_{i=1}^{n} I(\tilde{T}_i \geq s)} \right) \\
S_n(t) &= \prod_{\{j:t_j \leq t\}} \left(\frac{1 - [\frac{1}{n} \sum_{i=1}^{n} I(\tilde{T}_j = t_j, \Delta_i = 1]}{\sum_{i=1}^{n} I(\tilde{T}_i \geq t_j)} \right) \\
S_n(t) &= \prod_{\{j:t_j \leq t\}} \left(1 - \frac{d_j}{n_j} \right)
\end{aligned}
$$

This is the Kaplan-Meier estimator.

For any such 2 distributions – any distribution of the data can be indexed. So the unnivariate Right-censored data model is a nonn-parametric model. Whenever a model is locally nonn-parametric, any consistent estimator is asymptotically efficient.

Another estimator uses the general trick - inverse probability censored data mapping (IPCD).

$$
\begin{aligned}
S_0(t) &= E_0\, I(T \leq t) = E_0 \frac{I(\tilde{T} \leq t)\Delta}{\bar{G}_0(\tilde{T})} \\
\bar{G}_0(T) &= 1 - G_0 = P(C > t) \\
S_0(t) &= \frac{\frac{1}{n} \sum_{i+1}^{n} I(\tilde{T}_i \leq t)}{\bar{G}_{KM}(\tilde{T}_i)}
\end{aligned}
$$

$\bar{G}_{KM}$ is Kaplan-Meier estimator based on $\tilde{T}_i, (1-\Delta_i)$, $i = 1, \ldots, n$.

IPCD will be discussed further next lecture.

Raul Aguilar Schall

General approach for constructing an estimator

Consider the following Observations

$$O_i = (min(T_i, C_i), \Delta_i = I(T_i \leq C_i), W_i) \qquad i = 1, \ldots, n \qquad i.i.d.$$

Full data: $\widetilde{T}_i = min(T_i, C_i)$
$X_i = (T_i, W_i) \qquad i = 1, \ldots, n$
$(T, W) \sim F_x$
$C \mid X \sim G(\cdot \mid X) \equiv P(C \leq \cdot \mid X)$
$O \equiv P_{F_x, G}$

Now, the parameter of interest is $\varphi_0 = E[T \mid W]$ or equivalent $E[\log T \mid W]$

Coarsening at random on the censoring mechanism:$T \perp C \mid W$

Set k : # of bins. From previous results we have

$$\varphi_k(w \mid l_n) = \sum_{n=1}^{k} \left\{ \frac{I(w \text{ in } h^{th} \text{ bin})}{\sum_{i=1}^{n} I(w_i \text{ in } h^{th} \text{ bin})} \left[\sum_{i=1}^{n} I(w \text{ in } h^{th} \text{ bin}) T_i \right] \right\}$$

Notice that if $k = 1$ then the previous equation reduces to $\frac{1}{n} \sum_{i=1}^{n} T_i$

There is a problem with the used model. It only considers observed dat and acctually that is not the case we want to approach, since we also have censored data. It turns out that we do not have an estimator if we want to do histogram regression.

Assume for a second that we do have such estimators and want to choose among them using crossvalidation as we did before.

Choosing among estimators .
Consider the following set up
$\varphi_1(\cdot \mid P_n), \ldots, \varphi_k(\cdot \mid P_n) \qquad 1, \ldots, k \qquad$ bins
$L(X \mid \varphi) = [T - \varphi(w)]^2$
$\theta(k) = E_{F_X} L(X \mid \varphi)$
$B_n \in \{0, 1\}^n$ random vector
$i : B_n(i) = 1$ validation sample
$i : B_n(i) = 0$ training sample

$P^0_{n, B_n}, p^1_{n, B_n}$

The estimators would be of the form

$$\widehat{\theta}_{n(1-p)}(k) = E_{B_n}\left[\frac{1}{np}\sum_{i=1}^{n} I(B_n(i)=1)L(X_i, \varphi_k(\cdot \mid P^0_{n,B_n}))\right]$$

$$L\left(X_i, \varphi_k(\cdot \mid P^0_{n,B_n})\right) = \left[T_i - \varphi_k(w_i, P^0_{n,b_n})\right]^2$$

This does not work for censoring data because the equation needs all T_i, which we don't observe. The problems arrise within the Loss function.

Consider the following set up
$X_1, \ldots, X_n \qquad i.i.d. \qquad X \sim F_x$
Then $\mu \equiv \mu(F_x)$ and $\eta \equiv \eta(F_x)$

Definition: $D(X \mid \mu\eta)$
$E_{F_x} D(X \mid \mu(F_x), \eta(F_x)) = 0$

Then the estimating equation looks like:
$\frac{1}{n}\sum_{i=1}^{n} D(X_i \mid \mu\widehat{\eta}) = 0 \qquad \Rightarrow \widehat{\mu}$

example 1 :
$T_1, \ldots, T_n \qquad T \sim F_X \qquad i.i.d.$
$\mu = S(t) = P(T > t)$
$D(X \mid \mu) = I(T > t) - \mu$

this is the estimating function we are claiming and it clearly depends on the observed values and the parameter of interest. Now we need to prove that its expected value equals zero

$$\begin{aligned} E_{F_X} D(X \mid \mu) &= E[I(T\ t) - \mu] \\ &= E[I(T\ t)] - \mu \\ &= P(T\ t) - \mu \\ &= 0 \end{aligned}$$

example 2 :
$\mu = E[T]$
$D(X \mid \mu) = T - \mu$

$$\begin{aligned} E_{F_X} D(X \mid \mu) &= E[T] - E[\mu] \\ &= \mu - \mu \\ &= 0 \end{aligned}$$

Thus, $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} T_i$.

$D(X \mid \mu)$ is known as the full data estimating function.

Now let us move to right censored data
Observed data : $O_1, \ldots, O_n \qquad i.i.d \qquad O_i \equiv \widetilde{T}_i = \min(T_i, C_i)$
$\Delta_1, \ldots, \Delta_n \qquad \Delta_i = I(T_i \leq C_i)$
$T \sim F_X$
$C \sim G$
$O \sim P_{F_X,G}$
$\mu = I(T > t)$
$\overline{G}(\cdot \mid X) = P(C > \cdot \mid X) D(X \mid \mu) = I(T > t) - \mu$

Inverse probability of censoring weighted mapping (IPCW mapping) $D(X \mid \mu) = I(Tt) - \mu$
$IC(O \mid D, \mu, G) = \frac{D(X|\mu)\cdot\Delta}{\overline{G}(T|X)}$ which is the observed data estimating function
This estimating function equals zero if censoring occured, and, it equals $D(X \mid \mu)/\bar{G}(T \mid X)$ if censoring did not occur.

Now we have to see if this estimating function has $E[IC] = 0$ under the true distribution of the data

$$\begin{aligned} E_{P_{F_X,G}} IC(O \mid D, \mu, G) &= E_{P_{F_X,G}} \left[\frac{D(X \mid \mu) \cdot \Delta}{\overline{G}(T \mid X)} \right] \\ &= E_{F_X} \left[E_G \left[\frac{D(X \mid \mu) \cdot \Delta}{\overline{G}(T \mid X)} | X \right] \right] \\ &= E_{F_X} \left[\frac{D(X \mid \mu) \cdot \Delta}{\overline{G}(T \mid X)} E(\Delta \mid X) \right] \end{aligned}$$

Notice $E(\Delta \mid X) = 0 \cdot P(\Delta = 0 \mid X) + 1 \cdot P(\Delta = 1 \mid X)$
$= P(C \geq T \mid X) = \overline{G}(T \mid X)$

Thus the last term of the previous expectation reduces to $E_{F_X}[D(X \mid \mu)] = 0$. One regularity condition we need for all the previous work is

$$\overline{G}(\cdot \mid X) > \delta > 0 \qquad \text{in the support of } X.$$

Now, obtaining an estimator for $\overline{G}(T \mid X)$ we get the following

$$\frac{1}{n} \sum_{i=1}^{n} \frac{[I(T_i > t) - \mu]\Delta_i}{\overline{G}_n(T_i \mid X)} = 0$$

we will get a consistent estimator as long as we get a consistent estimator $\overline{G}_n$.

The new histogram regression estimator for censored data looks like

$$\varphi_k(w \mid P_n) = \sum_{n=1}^{k} \left\{ \frac{I(w \text{ in } h^{th} \text{ bin})}{\sum_{i=1}^{n} I(w_i \text{ in } h^{th} \text{ bin})} \left[\sum_{i=1}^{n} I(w \text{ in } h^{th} \text{ bin}) \cdot \frac{T_i \cdot \Delta_i}{\overline{G}_n(T_i \mid X)} \right] \right\}$$

$\varphi_0 = E\left[T \mid w\epsilon[w_i, w_{i+1}]\right]$
$D(X) = I(w\epsilon[w_i, w_{i+1}])T - \varphi_0$

$$\begin{aligned} E[D(X)] &= E[I(w\epsilon[w_i, w_{i+1}])T] - \varphi_0 \\ &= E[T \mid w\epsilon[w_i, w_{i+1}]] - \varphi_0 \\ &= 0 \end{aligned}$$

The questions that arrieses naturally is if we can get a cross-valitadion risk estimator with censored data. In order to do so, we need to find a new Loss function.
Remember that $L(X, \varphi) = [T - \varphi(w)]^2$ is the full data Loss function. Now we need the corresponding to the observed data

$IC(O \mid G, L(\cdot, \varphi)) = \frac{L(X,\varphi)\Delta}{\overline{G}(T|X)}$
Then

$$\widehat{\theta}_{n(1-p)}(k) = E_{B_n}\left[\frac{1}{np}\sum_{i=1}^{n} I(B_n(i) = 1)\frac{L(X_i, \varphi_k(\cdot \mid P^0_{n,B_n}))\Delta_i}{\overline{G}_n(T_i \mid X_i)}\right]$$

notice that this new cross-validation risk depends again on $\overline{G}$.

Lecture Notes
February 23, 2004
Scribe: Ritu Roydasgupta

Let $X_1, X_2, ..., X_n$ be i.i.d. observations having distribution f_θ where $\theta \in \Theta \subset R^K$

$H_0 : \theta = \theta_0$

I. Likelihood Ratio Test:

$$L_n(\theta) = L(\theta|X_1, X_2, ..., X_n) = \prod_{i=1}^{n} f_\theta(X_i) = \overset{argmax}{\theta \in \Theta} \sum_{i=1}^{n} log\ f_\theta(X_i)$$

Let θ_n be the maximum likelihood estimator for this parametric model.
Then the test statistic is given by

$$-2\ log\frac{L_n(\theta_0)}{L_n(\theta_n)} = 2[logL_n(\theta_n) - logL_n(\theta_0)]$$

Under regularity conditions,

$$-2\ log\frac{L_n(\theta_0)}{L_n(\theta_n)} \overset{H_0}{\sim} \chi^2_K$$

Proof:
Second order Taylor series expansion is given by

$$f(x) = f(x_0) + \frac{d}{dx}f(x)|_{x=x_0}(x - x_0) + (x - x_0)^t\frac{d^2}{dx^2}f(x)|_{x=x_0}(x - x_0)$$

where $f : R^K \rightarrow R$
We apply this to $f(\theta) = log\ L_n(\theta)$ with $x \Leftrightarrow \theta_0$, $x_0 \Leftrightarrow \theta_n$. Thus,

$$\begin{aligned} log\ L_n(\theta_0) - log\ L_n(\theta_n) &= \frac{d}{d\theta} log\ L_n(\theta)|_{\theta=\theta_n}(\theta_0 - \theta_n) \quad + \quad \sqrt{n}(\theta_0 - \theta_n)^t \frac{\frac{d^2}{d\theta^2} log\ L_n(\theta)|_{\theta=\theta_n}}{2n} \sqrt{n}(\theta_0 - \theta_n) \\ &= 0 \quad - \quad \sqrt{n}(\theta_0 - \theta_n)^t I(\theta_0) \sqrt{n}(\theta_0 - \theta_n)/2 \overset{D}{=} -\chi^2_K \ / \ 2, \end{aligned}$$

$$\begin{aligned} since \qquad & -\frac{1}{n}\frac{d^2}{d\theta^2} log\ L_n(\theta)|_{\theta=\theta_n} \ is\ an\ estimate\ of\ I(\theta_0), \\ & Z_{K\times 1} \sim N(0_{K\times 1}, \sum_{K\times K}) \ so\ that\ Z^t \sum^{-1} Z \sim \chi^2_K \ and \\ & \sqrt{n}(\theta_0 - \theta_n) = I(\theta_0) \end{aligned}$$

II. Score Test:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U(\theta_0)(X_i) \overset{H_0}{\sim} N(0, I(\theta_0))$$

$$U(\theta)(X) = \frac{d}{d\theta} log\ f_\theta(X)$$

so that

$$\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U(\theta_0)(X_i)\right]^t I^{-1}(\theta_0) \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U(\theta_0)(X_i)\right] \overset{H_0:\ \theta=\theta_0}{\sim} \chi^2_K$$

III. Chi Square Test:

$$\sqrt{n}(\theta_0 - \theta_n)^t I(\theta_0) \sqrt{n}(\theta_0 - \theta_n) \overset{H_0:\ \theta=\theta_0}{\sim} \chi^2_K$$

as $\sqrt{n}(\theta_0 - \theta_n) \overset{H_0}{\sim} N(0, I^{-1}(\theta_0))$
We could use $I(\theta_n)$ or another estimate of $I(\theta)$ (the true information matrix).

E.g.: $H_0 : P = P_0$
Test Statistic usually used

$$\frac{\sqrt{n}(\hat{P} - P_0)}{P_0(1 - P_0)} \sim N(0, 1)$$

estimates standard error incorrectly.
Test Statistic

$$\frac{\sqrt{n}(\hat{P} - P_0)}{\hat{P}(1 - \hat{P})}$$

estimates standard error correctly.

Which has more power ?

Right Censored Data

Let $T_1, T_2, ..., T_n$ be i.i.d. observations of $T \sim F_0$ (cdf) and censoring times $C_1, C_2, ..., C_n$ be i.i.d. observations of $C \sim G_0$

$$\begin{aligned} F_0(t) &= P(T \leq t) \\ S_0(t) &= 1 - F_0(t) \\ G_0(t) &= P(C \leq t) \\ \overline{G_0}(t) &= 1 - G_0(t) \end{aligned}$$

Only assumption: C and T are independent.
We are interested in estimating S_0.
We observe that

$$\left(\widetilde{T}_i = min(T_i, C_i), \Delta_i = I(T_i \leq C_i)\right) \sim P_{F_0,G_0}, \quad i = 1, 2, ..., n$$

That is, distribution of $(\widetilde{T}, \Delta)$ is indexed by $F_0,\ G_0$
This is a semi-parametric model as there are some assumptions like C and T are independent but F_0 and G_0 can be any cdfs.

$$\begin{aligned} P_{F,G}(t, \Delta = 1) &= P(\widetilde{T} = t, \Delta = 1) = P(T = t, T \leq c) = P(T = t, C > t) = dF(t)\overline{G}(t-) \\ P_{F,G}(t, \Delta = 0) &= P(\widetilde{T} = t, \Delta = 0) = P(C = t, T > t) = dG(t)S(t) \end{aligned}$$

Likelihood:

$$\begin{aligned} L(F, G|(\widetilde{T}_i, \Delta_i);\ i = 1, 2, ..., n) &= \prod_{i=1}^{n} P_{F,G}(\widetilde{T}_i, \Delta_i) \\ &= \prod_{i=1}^{n} \left[dF(\widetilde{T}_i)\overline{G}(\widetilde{T}_i)\right]^{\Delta_i} \left[S(\widetilde{T}_i)dG(\widetilde{T}_i)\right]^{1-\Delta_i} \\ &= \prod_{i=1}^{n} \left[dF(\widetilde{T}_i)^{\Delta_i} S(\widetilde{T}_i)^{1-\Delta_i}\right] \prod_{i=1}^{n} \left[\overline{G}(\widetilde{T}_i)^{\Delta_i} dG(\widetilde{T}_i)^{1-\Delta_i}\right] \end{aligned}$$

This is the factorization of likelihood into relevant F-part and irrelevant G-part.

$$log\ L(F, G|(\widetilde{T}_i, \Delta_i), i = 1, 2, ..., n) = \sum_{i=1}^{n} \Delta_i log\ dF(\widetilde{T}_i) + \sum_{i=1}^{n} (1 - \Delta_i) log\ S(\widetilde{T}_i) + G - part$$

Maximization of F does not depend on G-part.
Relevant log likelihood is given by the function

$$l_n(F) = \sum_{i=1}^{n} \Delta_i log\ dF(\widetilde{T}_i) + \sum_{i=1}^{n} (1 - \Delta_i) log\ S(\widetilde{T}_i)$$

Let the non-parametric maximum likelihood estimate,

$$NPMLE \;=\; F_n \;\equiv\; \overset{argmax}{F}\; l_n(F),$$

F is any cdf. Can we calculate this in closed form ?

Step 1: Show that $F \rightarrow l_n(F)$ is maximized at a discrete F with support $t_1 < t_2 < ... < t_m$, $m \leq n$, where the $t_j's$ are the distinct observed failure times.

Step 2:

$$dF(\widetilde{T}_i) \;\equiv\; d\Lambda(\widetilde{T}_i)S(\widetilde{T}_i-)$$

Re-parameterize $l_n(F)$ in terms of hazard function:
The number of failures at t_j is given by

$$\begin{aligned} d_j &= \sum_{i=1}^{n} I(\widetilde{T}_i = t_j)\Delta_i \\ dF(\widetilde{t}_i) &= d\Lambda(\widetilde{t}_i)S(\widetilde{t}_i-) \\ S(t) &= \prod_{s\ in\ (0,t]} (1 - d\Delta(s)) \end{aligned}$$

So,

$$\begin{aligned} l_n(F) &= l_n(\lambda_1, \lambda_2, ..., \lambda_m) \\ &= \sum_{i=1}^{n} \Delta_i log\ dF(\widetilde{T}_i) \;+\; \sum_{i=1}^{n}(1-\Delta_i) log\ S(\widetilde{T}_i-) \\ &= \sum_{i=1}^{n} \Delta_i log(d\Lambda(\widetilde{T}_i)S(\widetilde{T}_i-)) \;+\; \sum_{i=1}^{n}(1-\Delta_i) log\ S(\widetilde{T}_i-) \\ &= \sum_{i=1}^{n} \Delta_i log(d\Lambda(\widetilde{T}_i)) \;+\; \sum_{i=1}^{n} \Delta_i log(S(\widetilde{T}_i-)) \;+\; \sum_{i=1}^{n} log\ S(\widetilde{T}_i-) \;-\; \sum_{i=1}^{n} \Delta_i log\ S(\widetilde{T}_i-) \\ &= \sum_{i=1}^{n} \Delta_i log(d\Lambda(\widetilde{T}_i)) \;+\; \sum_{i=1}^{n} log\ S(\widetilde{T}_i-) \\ &= \sum_{j=1}^{m} d_j log\lambda_j \;+\; \sum_{i=1}^{n} log \left[\prod_{\{j:t_j<\widetilde{T}_i\}} (1 - d\Delta(t_j)) \right] \\ &= \sum_{j=1}^{m} d_j log\lambda_j \;+\; \sum_{i=1}^{n} \sum_{\{j:t_j<\widetilde{T}_i\}} log(1-\lambda_j) \end{aligned}$$

Let $\vec{\lambda_n} = (\lambda_{1n}, \lambda_{2n}, ..., \lambda_{mn})$ be the set of λ's that gives the maximum of $l_n(\lambda_1, \lambda_2, ..., \lambda_m)$.

$$\frac{d}{d\lambda_k} l(\lambda_1, \lambda_2, ..., \lambda_m) = 0$$

$$\Rightarrow \quad \frac{d}{d\lambda_k}\left[\sum_{j=1}^{m} d_j log\lambda_j \ + \ \sum_{i=1}^{n}\sum_{j:t_j<\widetilde{T_i}} log(1-\lambda_j)\right] \ = \ 0$$

$$\Rightarrow \quad \frac{d_k}{\lambda_k} - \frac{1}{1-\lambda_k}\sum_{i=1}^{n} I(\widetilde{T_i} > t_k) \ = \ 0$$

$$\Rightarrow \quad \frac{d_k}{\lambda_k} \ = \ \frac{1}{1-\lambda_k}\sum_{i=1}^{n} I(\widetilde{T_i} > t_k)$$

$$\Rightarrow \quad \frac{1}{\lambda_k} - 1 \ = \ \frac{1}{d_k}\sum_{i=1}^{n} I(\widetilde{T_i} > t_k)$$

$$\Rightarrow \quad \frac{1}{\lambda_k} \ = \ \frac{1}{d_k}\sum_{i=1}^{n} I(\widetilde{T_i} > t_k) \ + \ \frac{d_k}{d_k}$$

$$\Rightarrow \quad \lambda_{kn} \ = \ \frac{d_k}{d_k + \sum_{i=1}^{n} I(\widetilde{T_i} > t_k)}, \quad k = 1, 2, ..., m$$

This is the MLE of $l_n(\lambda_1, \lambda_2, ..., \lambda_m)$.

Class Notes: February 25, 2004

$(\tilde{T_i} = min(T_i, C_i), \triangle_i = I(T_i \leq C_i))$ i=1...n
$T \sim F_o$
$S_o(t) = 1 - F_o(t) = P(T > t)$

Log likelihood:
$ln(F) = \sum_{i=1}^{n} log[dF(\tilde{T_i})]\triangle_i + log(S(\tilde{T_i})(1 - \triangle_i)$

F is discrete on $t_i < \ldots < t_n$, these represent failure times
$dF(\tilde{T_i}) = S(\tilde{T_i^-})\lambda(\tilde{T_i})$, *every F has a corresponding hazard*

Note:

$$\lambda(t) = \frac{dF(t)}{S(t)} = \begin{cases} 0 & \text{if } t \in \{t_i, \ldots, t_n\} \\ \frac{dF(t)}{S(t_j^-)} & \text{if } t = t_j \end{cases}$$

$dF(\tilde{T_i}) = S(\tilde{T_i^-})\lambda(\tilde{T_i}) = \Pi(1 - \lambda_j)$
$S(t) = \Pi(1 - d\Lambda(s))$, from before

Thus, we have
$ln(F) = \sum_{i=1}^{n} log(\lambda(\tilde{T_i}))\triangle_i + log[\Pi(1 - \lambda_i)_{\{j:t_j<\tilde{T_j}\}}$

$= \sum_{j=1}^{m} log(\lambda_j)d_j + \sum_{i=1}^{n}\sum_{j:t_j<\tilde{T_j}} log(1 - \lambda_j)$
$= ln(\lambda_1 \ldots \lambda_m)$

We can find the maximum likelihood estimator by maximizing:
$(\lambda_1 \ldots \lambda_m) \rightarrow ln(\lambda_1 \ldots \lambda_m)$

Setting $\frac{d}{d\lambda_j} ln(\lambda_1 \ldots \lambda_m) = 0$ j = 1 to m

give us the closed form solutions (score equations)

$\lambda_{jn} = \frac{d_j}{d_j + \sum_{i=1}^{m} I(\tilde{T}_i > t_j)} = \frac{d_j}{n_j}$
represents number of people at risk at time t_j

The corresponding MLE of $S_o(t)$: $S_n(t) = \Pi_{j:t_j \leq t}(1 - \lambda_{jn}) = \Pi_{j:t_j \leq t}(1 - \frac{d_j}{n_j})$
- the Kaplan-Meier estimate

Use Greenwoods Formula to find estimate of variance -
Derivation uses the delta-method:

Working model: F is discrete on fixed points $t_1 < \ldots < t_m$
$\lambda_n = (\lambda_{1n} \ldots \lambda_{mn})$ is the MLE of λ in this parametric model

$S_n(t) = \Pi_{j:t_j < t}(1 - \lambda_{jn}) = g(\lambda_n)$
$S_n(t) = \Pi_{j:t_j < t}(1 - \lambda_{jo}) = g(\lambda_o)$

Delta-method - use Taylor expansion:
$S_n(t) - S_o(t) = g(\lambda_n) - g(\lambda_o) = a(\lambda_o)^T(\lambda_n - \lambda_o)_{m \times 1}$

$a(\lambda_o)$ = gradient of g at λ_o
$a(\lambda_o)^T = \frac{d}{d\lambda_1} g(\lambda), \ldots, \frac{d}{d\lambda_m} g(\lambda)|_{\lambda = \lambda_o}$

In this working model: $\sqrt{n}(\lambda_n - \lambda_o) \to N(0, I(\lambda_o)^{-1})$

Now use: var $a(\lambda_o)^T(\sqrt{n}(\lambda_n - \lambda_o))$

$= a(\lambda_o)^T I(\lambda_o)^{-1} a(\lambda_o)$

So, var $\sqrt{n}(S_n(t) - S_o(t)) \approx a(\lambda_o)^T I(\lambda_o)^{-1} a(\lambda_o)$
–Greenwood's Formula–

Note: We needed the gradient and the information matrix.
Step 1: Find $a(\lambda_o)^T$
Step 2: Find $I(\lambda_o)$
$\log P(\tilde{T} = t, \triangle = \delta) = \log (d\Lambda(t))\delta + \log S(t^-)$
$= \sum_{j=1}^{m} I(t = t_j) \log (\lambda_j)\delta + \sum_{j:t_j < t} \log (1 - \lambda_j)$
Another derivative or cov of scores gives the information matrix.
Note: the information matrix is diagonal, all cross derivatives are 0. i.e. $\frac{d}{d\lambda_1}\frac{d}{\lambda_2} = 0$
$U_j(\lambda) = \frac{d}{d\lambda_j} \log P_\lambda(\tilde{T}, \triangle) = \frac{I(\tilde{T} = t_j)\triangle}{\lambda_j} + \frac{I(\tilde{T} > t_j)}{1 - \lambda_j}$. Additionally, for j≠k $\frac{d^2}{d\lambda_j d\lambda_k} \log P_\lambda(\tilde{T}, \triangle) = 0$.
$I_{jj}(\lambda) = -E(\frac{d}{d\lambda_j} U_j(\lambda)(\tilde{T}, \triangle)) = E(\frac{I(\tilde{T} = t_j)\triangle}{\lambda_j^2} + \frac{I(\tilde{T} > t_j)}{(1 - \lambda_j)^2})$

Thus, $I(\lambda) = \text{diag } (I_{11}(\lambda) \ldots I_{mm}(\lambda))$
or $I^{-1} = \frac{1}{diags}$, i.e. inverse of each diagonal element

$\frac{dg}{d\lambda_j} = -I(t_j \leq t)\frac{S(t)}{1-\lambda_j}$

Gradient,
$a(\lambda_o)^T = -(I(t_1 \leq t)\frac{S(t)}{1-\lambda_1}, \ldots, I(t_m \leq t)\frac{S(t)}{1-\lambda_m})$

Some algebra gives,
$\sigma^2 = S^2(t)\sum_{t_j \leq t} I_{jj}^{-1}(\lambda_o)\frac{1}{(1-\lambda_{oj})^2}$

To estimate σ^2 use Kaplan-Meier,
$\hat{\sigma^2} = n^2(S_n(t)^2)\sum_{t_j \leq t}\frac{d_j}{n_j(n_j-d_j)}$

var $S_n(t) \approx nS_n(t)\sum_{t_j \leq t}\frac{d_j}{n_j(n_j-d_j)}$

This suggests that C.I.'s will be 'wider at the tails'.

C.I. $= S_n(t) \pm 1.96\frac{\hat{\sigma}}{\sqrt{n}}$ is aymptotic 0.95 C.I. for $S_o(t)$

Need Influence Curve to get simultaneous C.I.'s
- more about this next time!

Locally efficient estimation when death is reported with delay, Alan Hubbard

March 3, 2004
Notes by Kelly Moore

Description

CDC and California State Office of AIDS collect data on date of AIDS diagnosis. This data set is cross-referenced with hospital registry data on mortality. The result is data with some demographic variables, a little direct health information (CD4 count for some) and date of death for some subjects. Goal: Want to estimate survival by cohorts (year of diagnosis) to see trends in survival over time.

Data Structure

- Let T be the failure time of interest, e.g., time from AIDS diagnosis to death.
- Let C be censoring time, e.g., time from AIDS diagnosis to date of data collection.
- Let V be the time of failure reporting, e.g., time at which death of subject is reported.
- Let $X(u) = [W, R(u), R(u) * I(T \leq t)]$, where $R(u) = I(V \leq u)$ and W is a vector of baseline covariates.
- $\mu = F(t) = P(T \leq t)$.
- Full Data is: $\bar{X}(V)$
- Observed Data is: $Y = (\widetilde{T} = min(V, C), \Delta = I(\widetilde{T} = V), \bar{X}(\widetilde{T}))$

- Note, this is the same as the right-censoring data structure already discussed in class.

Ignoring Delay

If one ignores delay and treats the time of analysis as the censoring time say by performing Kaplan-Meier on the data:
$Y = (\widetilde{T} = min(T, C), \Delta = I(\widetilde{T} = T))$
then some subjects will be treated as if $T > C$ when in fact, $V > C$ and $T < C$.
Thus, the Kaplan-Meier estimator for $F(t) = I(T \leq t)$ will be biased low (or S(t) biased high).

Assumptions

There are 2 assumptions on censoring for the estimators:
1) $G(c|X) \equiv P(C < c|X) = G(c|W)$
2) $\frac{I(T \leq t)}{\bar{G}(V|W)} > 0$
F_X almost everywhere, where $\bar{G}(c|W) \equiv P(C \geq c|W)$
The second assumption implies that, given V, there has to be positive probability $C > V$ if $I(T \leq t)$.

Simple estimator in case of no covariates (W)

The nonparametric full-data estimating equation is: $I(T \leq t) - \mu$
Use the same, old IPCW trick:

$$E\left(\frac{I(T \leq t)\Delta}{\bar{G}(\widetilde{T}|X)}\right) = F(t)$$

In this case $E[\Delta|X] = P(C \geq V|X) = \bar{G}(V)$

IPCW Estimator

This leads to the following estimator:

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(T_i \leq t)\Delta_i}{\bar{G}_n(\widetilde{T}_i)} \qquad (*)$$

where $\bar{G}_n(\widetilde{T}_i)$ is the Kaplan-Meier estimator of the censoring distribution based on n observations of:

$$(\widetilde{T} = min((V, C), 1 - \Delta))$$

and V now plays the role of the censoring variable.
If there is no delay (so $V = T$), then $(*)$ reduces to the Kaplan-Meier estimator.
Like the Kaplan-Meier estimator, this IPCW estimator (in absence of covariates) is efficient.
Heuristic proof is that if G is estimated efficiently assuming only CAR, then the IPCW estimator is efficient.
In this case, the Kaplan-Meier is the NPMLE estimator of censoring assumes only $CAR[G(c|X) = G(c)]$.
More detailed proof in paper.

Inference - Influence curve

Robins and Rotnizky (1992) show that the influence curve for the IPCW estimator is:

$$IC(Y|F_X, F(t), G) = IC_0(Y|F(t), G) - \Pi(IC_0(Y|F(t), G)|T_2)$$

where

$$IC_0(Y|F(t), G) = \frac{I(T \leq t)\Delta}{\bar{G}(V)} - F(t)$$

and T_2 is the tangent space of the scores for estimation of G under independent censoring.
Robins (1996) showed that T_2 for Kaplan-Meier is:

$$\int h(u)dM(u) : h$$

Estimating the influence curve

The influence curve for the IPCW estimator is:

$$IC(Y|F_X, F(t), G) = \frac{I(T \leq t)\Delta}{\bar{G}(V)} - F(t) + \int F(t|\widetilde{T} > u)\frac{dM(u)}{\bar{G}(u)}$$

where

$$dM(u) = I(C \in du, \Delta = 0) - \int \Lambda_c(du)I(\widetilde{T} > u)$$

This can be estimated as:

$$IC(Y_i|F_{X,n}, F_n(t), G_n) = \frac{I(T_i \leq t)\Delta_i}{\bar{G}_n(V_i)} - F_n(t) + \frac{(1-\Delta)F_n(t|\widetilde{T} > C_i)}{\bar{G}_n(C_i)} - \sum_{u_j < \widetilde{T}_i} F_n(t|\widetilde{T} > u_j)\frac{\lambda_n(u_j)}{\bar{G}_n(u_j)}$$

Finally, estimate $F_n(t|\widetilde{T} > u)$ be repeating the IPCW sample for each censoring time (u) using only those subjects for which $\widetilde{T} > u$
The variance of the IPCW estimator can be estimated as:

$$\hat{\sigma}^2(t) = \frac{1}{n}\sum_{i=1}^{n} IC^2(Y_i|F_{X,n}, F_n(t), G_n)$$

Of course, this whole procedure is typically repeated for many t.

Long Delays

For more recent cohorts, there is the real possibility that the delay in reporting will be much greater than the censoring support. Because chronological censoring time is fixed in this case, it is easy to define the maximum possible censoring time, say τ.
If there is a chance of having $T < t$ but $V > \tau$, then the IPCW estimator is no longer consistent.

$$E\left(\frac{I(T \leq t\Delta)}{\bar{G}(V)}\right) = E\left(\frac{I(T \leq t)}{\bar{G}(V)}E[\Delta|V]\right)$$

$$E[\Delta|V] = \int I(C > V) dG(c) = I(V < \tau)\bar{G}(V)$$

so,

$$E\left(\frac{I(T \leq t\Delta)}{\bar{G}(V)}\right) = P(T \leq t, V < \tau)$$

Note:

$$F(t) = \frac{P(T \leq t, V < \tau)}{P(V < \tau | T \leq t)}$$

so if one knew (or could estimate) $P(V < \tau | T \leq t)$, then consistent estimation could be salvaged. In the AIDS survival case, earlier cohorts can be used to estimate this probability for later cohorts (assuming reporting mechanism does not change over time).

Regression with Delay of Reporting

Now consider a regression problem of the form:

$$\log T = \beta_0 + \beta_1 X + e$$

with $E(e|X) = 0$ and X a covariate of interest.
With uncensored data, a consistent estimating equation is based on simple least-squares, or:

$$\min_{\hat{\beta}} \sum_{i=1}^{n} (\log T - \beta_0 - \beta_1 X)^2$$

which is solved by:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where $Y = \log T$.

Weighted Regression

Now, must alter estimating equation to be consistent under censoring with reporting delay.
As it was in estimating the marginal survival distribution, the additional weight is:

$$\frac{\Delta_i}{\bar{G}_n(\widetilde{T}_i)}$$

the weights and estimator are:

$$W_i = \frac{\Delta_i}{G_n(\widetilde{T}_i)} \qquad \hat{\beta}_W = (X^T W X)^{-1} X^T W Y$$

Simulation

Below is simulation with:

1. Analysis time fixed at 8 months
2. AIDS dx is exponential, $\lambda = 0.25$

3. Covariate, $X \sim U(0,1)$
4. Death, $T = exp(X + e)$, $e \sim N(0, 0.5)$
5. $V = T + U$, $U \sim$ exponential, $\lambda = 0.5$

Code used to simulate data

```
x<-runif(1000,0,1)
T<-exp(x+rnorm(1000,0,0.5))
V<-T+pmin(rexp(1000,2),0.5)
cc<-rexp(1000,rate=1/4)
CC<-ta-cc
# Get rid of people whose dx is > analysis time
x<-x[CC>0]
V<-V[CC>0]
T<-T[CC>0]
CC<-CC[CC>0]
ttilde<-pmin(V,CC)
censor<-as.numeric(V<CC)
invcens<-1-censor
```

Code used to analyze simulated data

```
## Survival Time
surv.cens<-survfit(Surv(ttilde,invcens)~1)

## Get 1-Gn(V)
cens.prob<-get.at.surv.times(surv.cens,ttilde)

## Make weights
cwts<-censor/cens.prob

## Linear regression of logT on x
logt<-log(T)
init.lm<-lm(logt~x,weights=cwts)
```

Function to get censoring survival distribution at all

```
get.at.surv.times<-function(surv.cens, times)
{
#
# surv.cens is an object created by survfit
# times is a vector of times at which you want
# an estimate of the survival function
#
        nt <- length(times)
        outs <- rep(0, nt)
        survv <- summary(surv.cens)$surv
        ns <- length(survv)
        timev <- summary(surv.cens)$time
```

```
for(i in 1:nt) {
  if(times[i] < timev[1]) {
     outs[i] <- 1
  }
  else if(times[i] >= timev[ns]) {
     outs[i] <- survv[ns]
  }
  else {
     outs[i] <- survv[timev == max(timev[timev <= times[i]])][1]
  }
}
no <- length(outs[outs == 0])
outs[outs == 0] <- rep(survv[ns - 1], no)
return(outs)
}
```

Lecture notes for March 10, 2004
Hideaki Nakamura
hideakin demog.berkeley.edu

2 Functional Derivative

$(\widetilde{T} = Min(T, C), \Delta = I(T \leq C))$

$$S(t) = P(T > t) = \prod_{s \in (0,t]} \Big(1 - \frac{P_1(ds)}{\overline{P}(s-)}\Big), \quad \text{where}$$

$$P_1(t) = P(\widetilde{T} \leq t, \Delta = 1), \quad P(t) = P(\widetilde{T} \leq t), \quad \overline{P}(t) = P(\widetilde{T} > t)$$

$$S_n(t_0) = \prod_{s \in (0,t_0]} \Big(1 - \frac{P_{1n}(ds)}{\overline{P}_n(s-)}\Big)$$

$$P_{1n}(t) = \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{T}_i \leq t, \Delta_i = 1), \quad \overline{P}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{T}_i > t)$$

$$\text{Define} \quad \varphi(P_1, \overline{P}) = \prod_{s \in (0,t_0]} \Big(1 - \frac{P_1(ds)}{\overline{P}(s-)}\Big)$$

Then, $S(t_0) = \varphi(P_1, \overline{P}), \quad S_n(t_0) = \varphi(P_{1n}, \overline{P}_n)$

$$S_n(t_0) - S(t_0) = \varphi(P_{1n}, \overline{P}_n) - \varphi(P_1, \overline{P})$$

$$= \dot{\varphi}_1(P_1, \overline{P})(P_{1n} - P_1) + \dot{\varphi}_2(P_1, \overline{P})(\overline{P}_n - \overline{P})$$

Here, we use functional delta-method, which is a generalization of delta-method.

Review: Delta-method: This is a first-order approximation.

$$f(\theta_n) - f(\theta_0) = \frac{d}{d\theta} f(\theta) \cdot (\theta_n - \theta_0) + O(||\theta_n - \theta_0||)$$

where

$$\frac{d}{d\theta} f(\theta) = \left(\frac{d}{d\theta_1} f, \frac{d}{d\theta_2} f, \ldots, \frac{d}{d\theta_k} f \right)$$

We would like to define a derivative of φ in the direction h.
Example1: $f : \mathbf{R} \to \mathbf{R}$

$$\frac{d}{d\varepsilon} f(x + \varepsilon h)|_{\varepsilon=0} = f'(x) \cdot h$$

This is the simplest case (Linear mapping).
Example2: 2-dim case. $f : \mathbf{R^2} \to \mathbf{R}$

$$f(\theta, \theta_i) \quad \vec{h} = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

$$\frac{d}{d\varepsilon} f(\vec{\theta} + \varepsilon h)|_{\varepsilon=0} = \frac{d}{d\theta_1} f(\theta) h_1 + \frac{d}{d\theta_2} f(\theta) h_2 = \left(\frac{d}{d\theta_1} f(\theta), \frac{d}{d\theta_2} f(\theta) \right) \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

Define

$$\dot{\varphi_1}(P_1, \overline{P})(h_1) = \frac{d}{d\varepsilon} \varphi(P_1 + \varepsilon h_1, \overline{P})|_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{\varphi(P_1 + \varepsilon h_1, \overline{P}) - \varphi(P_1, \overline{P})}{\varepsilon}$$

Similarly,

$$\dot{\varphi_2}(P_1, \overline{P})(\overline{h}) = \frac{d}{d\varepsilon} \varphi(P_1, \overline{P} + \varepsilon \overline{h})|_{\varepsilon=0}$$

$$(P_{1n} - P_1)(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \left[I(\widetilde{T}_i \le t, \Delta_i = 1) - P_1(\cdot) \right] = \frac{1}{n} \sum_{i=1}^{n} f_1(\cdot | \widetilde{T}_i, \Delta_i)$$

Thus, by linearity of $\dot{\varphi_1}(P_1, \overline{P})$,

$$\dot{\varphi_1}(P_1, \overline{P})(P_{1n} - P_1) = \dot{\varphi_1}(P_1, \overline{P}) \left(\frac{1}{n} \sum_{i=1}^{n} f_1(\cdot | \widetilde{T}_i, \Delta_i) \right) = \frac{1}{n} \sum_{i=1}^{n} \dot{\varphi_1}(P_1, \overline{P}) \left[f_1(\cdot | \widetilde{T}_i, \Delta_i) \right]$$

A mapping $\dot{\varphi}$ is linear :

$$\dot{\varphi}(\alpha h_1 + \beta h_2) = \alpha \cdot \dot{\varphi}(h_1) + \beta \cdot \dot{\varphi}(h_2)$$

For example, Is $\dot{\varphi}(h)(t) = \int_0^t h(s) s^2 ds$ linear? ——— Yes, it is.

$$\dot{\varphi}(\alpha h_1 + \beta h_2) = \int_0^{\cdot} (\alpha(h_1)(s) + \beta(h_2)(s)) s^2 ds =$$

$$\alpha \int_0^{\cdot} (h_1)(s) s^2 ds + \beta \int_0^{\cdot} (h_2)(s) s^2 ds = \alpha \cdot \dot{\varphi}(h_1) + \beta \cdot \dot{\varphi}(h_2)$$

Similarly,

$$\dot{\varphi_2}(P_1, \overline{P})(\overline{P}_n - \overline{P}) = \dot{\varphi_2}(P_1, \overline{P}) \left(\frac{1}{n} \sum_{i=1}^{n} f(\cdot | \widetilde{T}_i) \right) = \frac{1}{n} \sum_{i=1}^{n} \dot{\varphi_2}(P_1, \overline{P}) \left(f(\cdot | \widetilde{T}_i) \right)$$

$$\text{where} \qquad (\overline{P}_n - \overline{P})(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\left(I(\widetilde{T}_i > \cdot) - \overline{P}(\cdot)\right) = \frac{1}{n}\sum_{i=1}^{n} f(\cdot|\widetilde{T}_i)$$

Thus,

$$S_n(t_0) - S(t_0) \cong \frac{1}{n}\sum_{i=1}^{n}\left[\dot{\varphi_1}(P_1, \overline{P})\left(f_1(\cdot|\widetilde{T}_i, \Delta_i)\right) + \dot{\varphi_2}(P_1, \overline{P})\left(f(\cdot|\widetilde{T}_i)\right)\right] = \frac{1}{n}\sum_{i=1}^{n} IC(\widetilde{T}_i, \Delta_i|P_1, \overline{P})$$

In this way, we can get IC. Once we get IC, we can compute 95-confidence interval, simultaneous confidence region, and so on. Now, we would like to find $\dot{\varphi_1}$ and $\dot{\varphi_2}$.
Example: $(\widetilde{T}_i, \Delta_i), i = 1, 2, \ldots, n$

$$\Lambda(t) = \int_0^t \Lambda(ds) = \int_0^t \frac{P_1(ds)}{\overline{P}(s-)} = \varphi^*(P_1, \overline{P})$$

$$\Lambda_n(t) = \int_0^t \frac{P_{1n}(ds)}{\overline{P}_n(s-)} = \varphi^*(P_{1n}, \overline{P}_n)$$

$$\Lambda_n(t) - \Lambda(t) = \varphi^*(P_{1n}, \overline{P}_n) - \varphi^*(P_1, \overline{P}) = \dot{\varphi_1^*}(P_{1n} - P_1) + \dot{\varphi_2^*}(\overline{P}_n - \overline{P}), \quad \text{where}$$

$$\dot{\varphi_1^*}(h_1) = \frac{d}{d\varepsilon}\varphi^*(P_1 + \varepsilon h_1, \overline{P})|_{\varepsilon=0}$$

$$\dot{\varphi_2^*}(\overline{h}) = \frac{d}{d\varepsilon}\varphi^*(P_1, \overline{P} + \varepsilon\overline{h})|_{\varepsilon=0}$$

$$\frac{d}{d\varepsilon}\varphi^*(P_1 + \varepsilon h_1, \overline{P}) = \frac{d}{d\varepsilon}\int_0^t \frac{(P_1 + \varepsilon h_1)(ds)}{\overline{P}(s-)} = \int_0^t \frac{h_1(ds)}{\overline{P}(s-)}$$

$$\frac{d}{d\varepsilon}\varphi^*(P_1, \overline{P} + \varepsilon\overline{h})|_{\varepsilon=0} = \frac{d}{d\varepsilon}\int_0^t \frac{P_1(ds)}{(\overline{P} + \varepsilon\overline{h})(s-)}|_{\varepsilon=0}$$

$$= -\int_0^t \frac{P_1(ds)}{\{(\overline{P} + \varepsilon\overline{h})(s-)\}^2} \cdot \overline{h}(s-)|_{\varepsilon=0} = -\int_0^t \frac{\overline{h}(s-)P_1(ds)}{\{\overline{P}(s-)\}^2}$$

$$\dot{\varphi_1^*}(P_{1n} - P_1) = \int_0^t \frac{(P_{1n} - P_1)(ds)}{\overline{P}(s-)} = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{I(\widetilde{T}_i \le ds, \Delta_i = 1)}{\overline{P}(s-)} - \int_0^t \frac{P_1(ds)}{\overline{P}(s-)}$$

$$\dot{\varphi_2^*}(\overline{P}_n - \overline{P}) = -\int_0^t (\overline{P}_n - \overline{P})(s-)\frac{P_1(ds)}{\{\overline{P}(s-)\}^2} = -\frac{1}{n}\sum_{i=1}^{n}\int_0^t I(\widetilde{T}_i \ge s)\frac{P_1(ds)}{\{\overline{P}(s-)\}^2} + \int_0^t \frac{P_1(ds)}{\overline{P}(s-)}$$

Thus,

$$\Lambda_n(t) - \Lambda(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I(\widetilde{T}_i \le t, \Delta_i = 1)}{\overline{P}(\widetilde{T}_i-)} - \int_0^{Min(t,\widetilde{T}_i)} \frac{P_1(ds)}{\{\overline{P}(s-)\}^2}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} IC^*(\widetilde{T}_i, \Delta_i | P_1, \overline{P})$$

Now, we could get IC of cumulative hazard. Next time, we will apply chain rule for functional derivative, and get influence curve of survival function.

Estimation of Influence Curve of Kaplan Meier using functional derivative
3/15/04

Joseph Poj Gavinlertvatana

Objective:

Derive the influence curve for Kaplan-Meier estimate by writing Kaplan-Meier minus truth as functional derivative of empiral minus truth in first order

Recall setup:

Observe: $(\tilde{T} = min(T, c), \Delta = I(T \le c))$, assume $T \perp c$

Parameter of interest:

$S(t) = P(T \ge t) = \prod_{s\in(0,t)}(1 - \frac{\partial p_1(s)}{\bar{p}(s^-)})$

where $p_1(t) = P(\tilde{T} \le t, \Delta = 1)$ and $\bar{p}(t) = P(\tilde{T} > t)$

$S(t) = \Phi(p_1, p_n)$ estimated by $S_n(t_o) = \Phi(p_{1n}, \bar{p}_n)$

Recall when we did directional derivative last time:

$\Phi(p_1, p_n) = \Phi_2[\Phi_1(p_1, p_n)]$

where $\Phi_1(p_1, p_n) = \int_0^{\cdot} \frac{\partial p_1(s)}{\bar{p}(s^-)} = \Lambda(\cdot)$

and $\Phi_2(\Lambda) = \prod_{s\in(0,t_0)}(1 - \partial\Lambda(s))$

$\Phi(p_1, p_n)$ maps data into cumulative hazard

$\Phi_2(\Lambda)$ maps cumulative hazard into survival

For influence curve, we need to find the directional deriv:

$\dot{\Phi}(h_1, \bar{h}) = \frac{\partial}{\partial\epsilon}\Phi(p_1 + \epsilon h_1, \bar{p} + \epsilon\bar{h})$

Since $\Phi(p_1, p_n)$ is a composite function, we use the chain rule:

$\dot{\Phi}(h_1, \bar{h}) = \dot{\Phi}_2[\dot{\Phi}_1(h_1, \bar{h})]$

where $\dot{\Phi}_1(h_1, \bar{h}) = \frac{\partial}{\partial\epsilon}\Phi_1(p_1 + \epsilon h_1, \bar{p} + \epsilon\bar{h}) \mid_{\epsilon=0} = \int_0^t \frac{h_1(\partial s)}{\bar{p}(s-)} - \int_0^t \frac{\bar{h}(s-)p_1(\partial s)}{[\bar{p}(s-)]^2}$

and $\dot{\Phi}_2(g) = \frac{\partial}{\partial\epsilon}\Phi_2(\Lambda + \epsilon g) \mid_{\epsilon=0}$

To get $\dot{\Phi}_2(g)$, we use a telescoping trick

e.g.

$a_1a_2 - b_1b_2 = (a_1 - b_1)b_2 + a_1(a_2 - b_2)$

$a_1a_2a_3 - b_1b_2b_3 = (a_1 - b_1)b_2b_3 + a_1(a_2 - b_2)b_3 + a_1a_2(a_3 - b_3)$

In general:

$\prod_{j=1}^{k} a_j - \prod_{j=1}^{k} = \sum_{j=1}^{k}[(\prod_{l=1}^{j-1} a_l)(a_j - b_j)(\prod_{l=j+1}^{k} b_l)]$

Continuing, we get:

$\dot{\Phi}_2(g) = \lim_{s\to 0} \frac{\Phi_2(\Lambda+\epsilon g)-\Phi_2(\Lambda)}{\epsilon}$

$= \lim_{\epsilon\to 0} \frac{\prod_{s\epsilon(0,t_0)}[1-(\Lambda+\epsilon g)(s)]-\prod_{s\epsilon(0,t_0)}[1-\partial\Lambda(s)]}{\epsilon}$

(using telescoping)

$= \lim_{\epsilon\to 0} \frac{\int_{s\epsilon(0,t_0)}[\prod_{u\epsilon(0,s)}(1-\partial(\Lambda+\epsilon g)(u))][1-\partial(\Lambda+\epsilon g)(s)-1+\partial(\Lambda)(s)][\prod_{u\epsilon(s,t_0)}(1-\partial\Lambda(u))]}{\epsilon}$

$= \lim_{\epsilon\to 0} -\int_{s\epsilon(0,t_0)}[\prod_{u\epsilon(0,s)}(1-\partial(\Lambda+\epsilon g)(u))][\partial g(s)][\prod_{u\epsilon(s,t_0)}(1-\partial\Lambda(u))]$

$= -\int_{s\epsilon(0,t_0)}[\prod_{u\epsilon(0,s)}(1-\partial(\Lambda)(u))][\partial g(s)][\prod_{u\epsilon(s,t_0)}(1-\partial\Lambda(u))]$

$= -\int_{s\epsilon(0,t_0)} \prod_{u\epsilon(0,t_0)} \frac{(1-\partial(\Lambda)(u))}{(1-\partial(\Lambda)(s))}\partial g(s)$

$= -\prod_{u\epsilon(0,t_0)}(1-\partial(\Lambda)(u)) \int_0^{t_0} \frac{\partial g(s)}{(1-\partial(\Lambda)(s))}$

$= -S(t_0) \int_0^{t_0} \frac{\partial g(s)}{(1-\partial(\Lambda)(s))}$

If $\Lambda(s)$ is continuous, then $\partial\Lambda(s) = 0$, so

$= -S(t_0)g(t_0)$

So by plugging in $g = \dot{\Phi}_1(h_1, \bar{h})$

$\dot{\Phi}(h_1, \bar{h}) = -S(t_0) \int_0^{t_0} \frac{\partial\dot{\Phi}_1(h_1,\bar{h})}{1-\partial\Lambda(s)}$

$= S(t_0) \int_0^{t_0} \frac{1}{1-\partial\Lambda(s)}[\frac{h_1(\partial s)}{\bar{p}(s-)} - \frac{\bar{h}(s-)}{(p)^2(s-)}\partial p_1(s)]$

So, finally, we can write Kaplan-Meier minus truth in first order:

$S_n(t_0) - S(t_0) \cong \dot{\Phi}(p_{1n} - p_1, \bar{(p)}_n - \bar{p})$

$= -S(t_0)[\int_0^{t_0} \frac{1}{\partial\Lambda(s)} \frac{\partial p_{1n}}{\bar{p}(s-)} - \int_0^{t_0} \frac{\bar{p}_n}{\bar{p}^2(s-)}\partial p_1(s)]$

$= -S(t_0)[\int_0^{t_0} \frac{1}{\partial\Lambda(s)} \frac{\partial(p_{1n}-p_1)}{\bar{p}(s-)} - \int_0^{t_0} \frac{1}{\partial\Lambda(s)} \frac{(\bar{p}_n-\bar{p})(s-)}{\bar{p}^2(s-)}\partial p_1(s)]$

And we make this into a sample average:

$= \frac{1}{n}\sum_{i=1}^{n} -S(t_0)[\int_0^{t_0} \frac{1}{1-\partial\Lambda(s)} \frac{\partial I(\tilde{T}_i\epsilon\partial s,\Delta_i=1)}{\bar{p}(s-)} - \int_0^{t_0} \frac{1}{1-\partial\Lambda(s)} \frac{I(T_i\geq s)}{\bar{p}^2(s-)}\partial p_1(s)]$

Finally, we get the influence curve:

$S_n(t_0) - S(t_0) \cong \frac{1}{n}\sum_{i=1}^{n} -S(t_0)[\frac{I(\tilde{T}_i\leq t_0,\Delta_i=1)}{(1-\partial\Lambda(\tilde{(T)}_i))\bar{p}(\tilde{T}_i-)} - \int_0^{min(t_0,\tilde{T}_i)} \frac{1}{1-\partial\Lambda(s)} \frac{\partial p_1(s)}{\bar{p}^2(s-)}]$

recall that $\partial\Lambda(\tilde{T}_i) = \frac{\partial p_1(\tilde{T})}{\bar{p}}$

Lecture notes
March 29, 2004
Vera Klimkovsky
klimkovsky@yahoo.com

QUANTILE ESTIMATION BASED ON RIGHT-CENSORED DATA

Data:

$$(\widetilde{T} = \min(T, C), \Delta = I(T \leq C))$$

where $C \perp T$ (C and T are independent), $C \sim G$, $T \sim F$

Parameter of interest:

$$\theta(F) = F^{-1}(p), \text{ where } p \in (0,1) \text{ and } F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

Estimation:

$$F_n = 1 - S_n, \text{ and } S_n\text{is Kaplan-Meier estimator. } \theta_n = \theta(F_n) = F_n^{-1}(p)$$

Now, how do we do the inference?
We need to estimate IC.
$\theta_n - \theta = \theta(F_n) - \theta(F) \cong \dot{\theta}(F_n - F)$ (using Delta Method)
Take a directional derivative:
$\dot{\theta} = \frac{d}{d\varepsilon}\theta(F + \varepsilon \cdot h)\big|_{\varepsilon=0}$
For simplicity, let $F_\varepsilon = F + \varepsilon h$, $\theta_\varepsilon = \theta(F_\varepsilon)$, and $y = F(\theta_\varepsilon)$. Also, note that $F_\varepsilon(\theta_\varepsilon) = F(\theta_\varepsilon) + \varepsilon h(\theta_\varepsilon)$
Then,

$$\dot{\theta} = \frac{d}{d\varepsilon}\theta(F + \varepsilon \cdot h)\big|_{\varepsilon=0}$$

$$\dot{\theta} = \lim_{\varepsilon \to 0} \frac{\theta(F_\varepsilon) - \theta(F)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{F_\varepsilon^{-1}(p) - F^{-1}(p)}{\varepsilon} = \lim_{\varepsilon \to 0}[-\frac{F^{-1}F_\varepsilon(\theta_\varepsilon) - F^{-1}F(\theta_\varepsilon)}{\varepsilon}]$$

Substituting y into the expression of limit, and using the fact that

$$F^{-1}(y + \varepsilon h) - F^{-1}(y) \cong \frac{d}{dy}F^{-1}(y) \cdot \varepsilon h$$

, we get

$$\dot{\theta}(h) = \lim_{\varepsilon \to 0}\left[-\frac{F^{-1}(y + \varepsilon h(\theta_\varepsilon)) - F^{-1}(y)}{\varepsilon}\right] = \lim_{\varepsilon \to 0}\left[-\frac{\frac{d}{dy}F^{-1}(y) \cdot \varepsilon h(\theta_\varepsilon)}{\varepsilon}\right]$$

$$= -\lim_{\varepsilon \to 0}\frac{d}{dy}F^{-1}(y)\Big|_{y=F(\theta_\varepsilon)} \cdot h(\theta_\varepsilon)$$

$$= -\frac{d}{dy}F^{-1}(y)\Big|_{y=F(\theta)} \cdot h(\theta)$$

$$= -\frac{1}{f(\theta)}h(\theta) = -\frac{1}{f(F^{-1}(p))}h(F^{-1}(p))$$

Note: Above we differentiated the inverse function.
What is the derivative of F^{-1}?
Recall:

$$F^{-1}(F(y)) = y$$

Differentiate both sides with respect to y:

$$\frac{d}{dy}F^{-1}F(y) = 1$$

Applying Chain Rule:

$$(F^{-1})'(F(y)) \cdot F'(y) = 1, \text{ note: } F'(y) \equiv f(y)$$

So,

$$(F^{-1})'(F(y)) = \frac{1}{f(y)}$$

Thus,

$$(F^{-1})'(z) = \frac{1}{f(F^{-1}(z))}$$

Now we are ready for the influence curve.
Applying the first order Taylor expansion:

$$\theta_n - \theta \cong \frac{-1}{f(\theta)}(F_n - F)(\theta) = \frac{1}{f(\theta)}(S_n - S)(\theta)$$

$$\cong \frac{1}{f(\theta)} \frac{1}{n} \sum_{i=1}^{n} IC_{KM}(\widetilde{T}_i, \Delta_i | \theta)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{f(\theta)} IC_{KM}(\widetilde{T}_i, \Delta_i | \theta)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{-1}{f(\theta)} S(\theta) \left[\frac{I(\widetilde{T}_i, \Delta_i = 1)}{\bar{p}(\theta)} - \int_0^{\min(\theta, \widetilde{T}_i)} \frac{dP_1(S)}{\bar{P}^2(S)} \right]$$

$$f_n(\theta) = \frac{F_n(\theta + h) - F_n(\theta - h)}{2h}$$

(Another way to construct confidence interval is to use bootstrap method. However, there are situations when the bootstrap method fails.)

SELF-CONSISTENCY EQUATION FOR CENSORED DATA

$X \sim F_n, C|X \sim G(\cdot|X)$ where distribution function G is called a censoring mechanism.
We observe: $Y = \varphi(C, X) \sim P_{F_X, G}$
A random set $C(Y)$ is called a coarsening of X if $Pr(X \in C(Y)) = 1$
(1)

$$Y = (\widetilde{T}, \Delta)$$

$$C(\widetilde{T}, \Delta) = \begin{cases} \{\widetilde{T}\}, \text{if } \Delta = 1 \\ (\widetilde{T}, \infty), \text{if } \Delta = 0 \end{cases}$$

(2)

$$Y = (C, I(T < C))$$

$$C(\widetilde{T}, \Delta) = \begin{cases} (-\infty, C), \text{if } \Delta = 1 \\ (C, \infty), \text{if } \Delta = 0 \end{cases}$$

We say $C(Y)$ satisfies *coarsening at random* (CAR) if $Pr(X = x|Y = y) = Pr(X = x|X \in C(y))$.

Equivalenty, $Pr(Y = y|X = x)$ is constant for $x \in C(y)$.

Suppose we have n iid: $Y_1, \ldots, Y_n$. We want to estimate $F_X(A) = Pr(X \in A)$.

Full data:

$$F_{X,n} = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in A)$$

Censored data:

$$F_{X,n}(A) = \frac{1}{n} \sum_{i=1}^{n} E\left[I(X_i \in A|Y_i)\right]$$

Or

$$F_{X,n}(A) = \frac{1}{n} \sum_{i=1}^{n} E_{F_{X,n}}\left[I(X_i \in A|X_i \in C(Y_i))\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} P_{F_{X,n}}(X_i \in A \big| X_i \in C(Y_i))$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{P_{X,n}(X_i \in A \cap C(Y_i))}{P_{F_{X,n}}(X \in C(Y_i))}$$

Thus,

$$\underbrace{F_{X,n}(A) = \frac{1}{n} \sum_{i=1}^{n} \frac{P_{X,n}(X_i \in A \cap C(Y_i))}{P_{F_{X,n}}(X \in C(Y_i))}}_{\text{This is called a self-consistency equation}}$$

If we enforce $F_{X,n}$ to be discrete on $\{X_1, \ldots, X_m\}$ with point masses $p_{1n} \ldots p_{mn}$, then $p_{j,n} = P_{F_{X,n}}(X = x_j)$

$$p_{j,n} = \frac{1}{n} \sum_{i=1}^{n} \frac{p_{j,n} I(x_i \in C(Y_i))}{\sum_{x_l \in C(Y_i)} p_{l,n}}$$

for $j = 1, \ldots, m$

The following algorithm can be used to solve this equation:

$$p_{j,n}^{k+1} = \frac{1}{n} \sum_{i=1}^{n} \frac{p_{j,n}^{k} I(x_i \in C(Y_i))}{\sum_{x_l \in C(Y_i)} p_{l,n}^{k}}$$

where the iteration start with $p^0 = (p_{1,n}^0, \ldots, p_{m,n}^0)$.

Truncation

Notes by Keith Betts

March 17, 2004

Situation

Survival time T has distribution F
Truncation time C^* has distribution G^*
T and C^* are independent
Observe n i.i.d. observations of (T', $C^{*\prime}$) with distribution (T, C^*) | $T > C^*$
We want to estimate $S(t) = Pr(T > t)$.

Example

Population of HIV infected patients
$T = T_{Death} - T_{AIDS}$
$C^* = 1988 - T_{AIDS}$ where $C^* = 0$ if $T_{AIDS} > 1988$
Sample is from (T, C^*) | $T > C^*$

Define $P(s) = Pr(T' \geq s)$
Define $\bar{P}(s) = Pr(T' \geq s, C^{*\prime} < s)$

$$\begin{aligned}
S(t) = Pr(T > t) &= \prod_{s\in(0,t)} 1 - \frac{Pr(T \in (s, s+ds))}{Pr(T \geq s)} \\
&= \prod_{s\in(0,t)} 1 - \frac{Pr(T \in (s, s+ds), C^* < s)}{Pr(T \geq s, C^* < s)} \\
&= \prod_{s\in(0,t)} 1 - \frac{Pr(T \in (s, s+ds), C^* < s \mid T > C^*)}{Pr(T \geq s, C^* < s \mid C^*)} \\
&= \prod_{s\in(0,t)} 1 - \frac{Pr(T' \in (s, s+ds), C^{*\prime} < s)}{Pr(T' \geq s, C^{*\prime} < s)} \\
&= \prod_{s\in(0,t)} 1 - \frac{Pr(T' \in (s, s+ds))}{Pr(T' \geq s, C^{*\prime} < s)} \\
&= \prod_{s\in(0,t)} 1 - \frac{P(s)}{\bar{P}(s)} \\
&\equiv \varphi(P(s), \bar{P}(s))
\end{aligned}$$

Essientially, we have proved Identifiability.

Trouble arises if $\bar{P}(s) = Pr(T' \geq s, C^{*\prime} < s) = 0$
This occurs if $C^* < s$ is never true.

Assuming that $Pr(C^* < s) > 0$ for $s \in (0, t)$, Identifiability is shown.

We now focus on the task of constructing estimators.
Define $P_n(s) = \frac{1}{n}\sum_{i=1}^{n} I(T_i' \leq s)$
Define $\bar{P}_n(s) = \frac{1}{n}\sum_{i=1}^{n} I(T_i' \geq s, C_i^{*\prime} < s)$

We have shown that $S(t) = \varphi(P, \bar{P})$
Therefore, $S_n(t) = \varphi(P_n, \bar{P}_n)$
This is the Product Limit Estimator for truncated data.

We will now turn our attention to obtaining the Influence Curve.
Using the Functional Delta Method:

$$\begin{aligned}
S_n(t) - S(t) &\approx \dot{\varphi}(P_n - P, \bar{P}_n - \bar{P}) \\
\text{where } \dot{\varphi}(h, \bar{h}) &= \frac{d}{d\epsilon}\varphi(P + \epsilon h, \bar{P} + \epsilon \bar{h})|_{\epsilon=0} = -S(t)[\int_0^t \frac{h(ds)}{\bar{P}(s)} - \int_0^t \frac{\bar{h}(s)dP(s)}{\bar{P}^2(s)}] \\
&\quad \text{Thus, } h = P_n - P, \bar{h} = \bar{P}_n - \bar{P} \\
S_n(t) - S(t) &\approx \frac{1}{n}\sum_{i=1}^{n} -S(t)[\int_0^t I(T_i' \in (s, s+ds) - \int_0^t I(T_i' \geq s, C_i^{*\prime} < s)\frac{dP(s)}{\bar{P}^2(s)} \\
&\approx \frac{1}{n}\sum_{i=1}^{n} \underbrace{-S(t)[\frac{I(T_i' \leq t)}{\bar{P}(T_i')} - \int_0^t I(T_i' \geq s, C_i^{*\prime} < s)\frac{dP(s)}{\bar{P}^2(s)}}_{IC(T_i', C_i^{*\prime}, \mid P, \bar{P}, t)}
\end{aligned}$$

Truncation with Right Censoring

Assume:
$T \perp (C, C^*)$
T $\sim$ F
$C^* \sim G^*$
C $\sim$ G
We observe $(\tilde{T} = min(T, C), \triangle = I(T \leq C), C^*)) \mid T > C^*, C > C^*$
Where $(T', C', C^{*\prime})$ has distribution $(T, C, C^*) \mid T > C^*, C > C^*$
Define $\tilde{T}' = min(T', C')$ and $\triangle' = I(T' \leq C')$

We want to estimate S(t) = Pr(T > t)
Express S(t) as distribution of data based on n i.i.d. $(\tilde{T}_i', \triangle_i', C_i^{*\prime})$ for i = $1, \ldots, n$

$$\begin{aligned}
S(t) = Pr(T > t) &= \prod_{s \in (0,t)} (1 - \frac{Pr(T = s)}{Pr(T \geq s)}) \\
&= \prod_{s \in (0,t)} (1 - \frac{Pr(T = s, C^* < s, C > s)}{Pr(T \geq s, C^* < s, C > s)})
\end{aligned}$$

$$
\begin{aligned}
&= \prod_{s\in(0,t)} (1 - \frac{Pr(T = s, C^* < s, C > s | T > C^*, C > C^*)}{Pr(T \geq s, C^* < s, C > s | T > C^*, C > C^*)}) \\
&= \prod_{s\in(0,t)} (1 - \frac{Pr(\tilde{T}' = s, \triangle' = 1, C^{*\prime} < s)}{Pr(\tilde{T}' \geq s, C^{*\prime} < s)}) \\
&\approx \varphi(P_1, \bar{P}) \\
\text{where } P_1(s) &= Pr(\tilde{T}' \leq s, \triangle' = 1) \\
\bar{P}(s) &= Pr(\tilde{T}' \geq s, C^{*\prime} < s)
\end{aligned}
$$

$$
\begin{aligned}
\text{Thus, } S(t) &= \varphi(P_1, \bar{P}) \\
S_n(t) &= \varphi(P_{1n}, \bar{P}_n) \\
\text{where } P_{1n}(s) &= \frac{1}{n}\sum_{i=1}^{n} I(\tilde{T}_i' \leq s, \triangle_i' = 1) \\
\text{and } \bar{P}_n(s) &= \frac{1}{n}\sum_{i=1}^{n} I(\tilde{T}_i' \leq s, C_i^{*\prime} < s)
\end{aligned}
$$

Using the Functional Delta Method

$$
\begin{aligned}
S_n(t) - S(t) &\approx \dot{\varphi}(P_{1n} - P_1, \bar{P}_{1n} - \bar{P}_1) \\
\text{where} \qquad & \dot{\varphi}(h_1, \bar{h}) = -S(t)[\int_0^t \frac{h_1(ds)}{\bar{P}(s)} - \int_0^t \frac{\bar{h}(s)dP(s)}{\bar{P}^2(s)}] \\
S_n(t) - S(t) &\approx \frac{1}{n}\sum_{i=1}^{n} -S(t)[\int_0^t \frac{I(\tilde{T}_i' = s, \triangle_i' = 1)}{\bar{P}(s)} - \int_0^t I(\tilde{T}_i' \geq s, C_i^{*\prime} < s)\frac{dP_1(s)}{\bar{P}^2(s)}] \\
&\approx \frac{1}{n}\sum_{i=1}^{n} \underbrace{-S(t)[\int_0^t \frac{I(\tilde{T}_i' = s, \triangle_i' = 1)}{\bar{P}(T_i')} - \int_0^t I(\tilde{T}_i' \geq s, C_i^{*\prime} < s)\frac{dP_1(s)}{\bar{P}^2(s)}]}_{IC(T_i', C_i^{*\prime} \mid P_1, \bar{P}, t)}
\end{aligned}
$$

Quantiles of F

With right censored data, the mean is difficult to estimate.
A natural method of comparison is using Quantiles.

Assume:
T $\sim$ F
$\theta = F^{-1}(p)$, $p \in (0, 1)$
$\theta_n = F_n^{-1}(p)$, where $F_n = 1 - S_n$
Notice:
$\theta = \varphi(F) \equiv F^{-1}(p)$
$\theta_n = \varphi(F_n)$

Use the functional Delta method:

$$\begin{aligned}
\theta_n - \theta &\approx \dot{\varphi}(F_n - F) \\
\text{where} \qquad & \dot{\varphi}(n) = \frac{d}{d\epsilon}\varphi(F + \epsilon h)|_{\epsilon=0} \\
\theta_n - \theta &\approx \dot{\varphi}(\frac{1}{n}\sum_{i=1}^{n} IC_i(.)) \\
&\approx \frac{1}{n}\sum_{i=1}^{n} \underbrace{\dot{\varphi}(IC_i(.))}_{IC\ of\ \theta_n}
\end{aligned}$$

Lecture notes, Peter Dimitrov, March 31

3 The EM-Algorithm For Computing The MLE of a Full Data Model Based on Censored Data

Let $X_1, X_2, \ldots, X_n$ be n *i.i.d.* observations drawn from a full data distribution F_X with density f_θ. Let the observed data $Y_1 = \phi(C_1, X_1), Y_2 = \phi(C_2, X_2), \ldots, Y_n = \phi(C_2, X_n)$ be n *i.i.d.* observations $Y = \phi(C, X) \sim P_{f_\theta, G} \equiv P_{\theta,G}$, where the random variable C represents the censoring on X (note that C and X may not necessarily be univariate R.V.). Let $G \equiv G_{C|X}$ denote the censoring mechanism.

Let $C(Y)$ be the *coarsening* of X implied by Y: $C_Y \equiv C(Y)$ is a subset of domain (X) such that $\Pr(X \in C(Y)) = 1$. We'll assume for the rest of this lecture that C_Y satisfies the so-called *Coarsening At Random* (CAR) condition: $\Pr(Y = y \mid X = x)$ is constant for $x \in C(Y)$. Alternatively the condition can be expressed as that the coarsening mechanism (i.e. the conditional density g of the conditional distribution $G'(\cdot|X)$ of the observed data Y given X) is CAR if it's a function of Y only. In other words knowing what the value of the random variable X is, does not in any way provide more info on the observed data Y, including the censoring mechanism G.

Assuming CAR, the likelihood for an individual observation factorizes into a part that depends on the censoring mechanism G and a part that doesn't: $P_{\theta,G}(Y = y) = P_{f_\theta}(X \in C(Y)) P_G(Y = y \mid X = x)$. Hence, the log-likelihood of $Y_1, Y_2, \ldots, Y_n$ under CAR simplifies to:

$$Loglik(Y_1, Y_2, \ldots, Y_n \mid \theta, G) = \sum_{i=1}^{n} \log F_\theta(C(Y_i)) + \sum_{i=1}^{n} \log P_G(Y_i|X_i),$$

which is maximized by the MLE estimator:

$$\theta_n = \arg\max_\theta \sum_{i=1}^{n} \log F_\theta(C(Y_i)).$$

Note that $F_\theta(C(Y_i))$ is an alternative form for $\int_{C(Y)} f_\theta(x)\,dx$.

In the most general case, the model for F_θ is non-parametric , and θ_n is found using the *Expectation-Maximization algorithm*, or *EM-alg* for short, which is described below:

1. Initialize θ with some value θ^0;

2. For $k = 1, 2, \ldots$ repeat until convergence:

$$\theta^{k+1} = \arg\max_{\theta} E_{\theta^k} \left[\log f_\theta \left(X_1, \ldots, X_n \right) \mid Y_1, \ldots, Y_n \right].$$

The EM algorithm performs repeatedly two steps: first, for θ^k - fixed, the conditional expectation of the the observed data log-likelihood is calculated, which simultaneously imputes the missing/censored data as well. This is followed by finding which value of θ maximizes the full-data likelihood using the newly imputed values of X and setting the θ^{k+1} to that value.

The EM-alg always converges to a local maximum. In the case of X being distributed with density from the exponential family of distributions:

$$f_\theta(x) = \exp\left[p(\theta) K(x) + S(x) + q(\theta) \right]$$

the EM-alg reduces to:

$$\begin{aligned}
\theta^{k+1} &= \arg\max_{\theta} \sum_{i=1}^{n} E_{\theta^k} \left[\log f_\theta(X_i) \mid X_i \in C(Y_i) \right] \\
&= \arg\max_{\theta} \sum_{i=1}^{n} E_{\theta^k} \left[\log f_\theta(X_i) \mid Y_i \right] \\
&= \arg\max_{\theta} \sum_{i=1}^{n} E_{\theta^k} \left[p(\theta) K(X_i) + S(X_i) + q(\theta) \mid Y_i \right] \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \left\{ p(\theta) E_{\theta^k} \left[K(X_i) \mid Y_i \right] + E_{\theta^k} \left[S(X_i) \mid Y_i \right] + q(\theta) \right\} \\
&= \arg\max_{\theta} \left\{ p(\theta) \sum_{i=1}^{n} E_{\theta^k} \left[K(X_i) \mid Y_i \right] + \sum_{i=1}^{n} E_{\theta^k} \left[S(X_i) \mid Y_i \right] + nq(\theta) \right\} \\
&= \arg\max_{\theta} \left\{ p(\theta) K_n(Y_1, \ldots, Y_n) + S_n(Y_1, \ldots, Y_n) + nq(\theta) \right\},
\end{aligned}$$

where $K_n(Y_1, \ldots, Y_n) \equiv \sum_{i=1}^{n} E_{\theta^k} \left[K(X_i) \mid Y_i \right]$ and $S_n(Y_1, \ldots, Y_n) \equiv \sum_{i=1}^{n} E_{\theta^k} \left[S(X_i) \mid Y_i \right]$.

Now, compare the above with the full data MLE based on $X_1, \ldots, X_n$:

$$\begin{aligned}
\theta^{\text{Full}} &= \arg\max_{\theta} \sum_{i=1}^{n} \log f_\theta(X_i) \\
&= \arg\max_{\theta} \left\{ p(\theta) \sum_{i=1}^{n} K(X_i) + \sum_{i=1}^{n} S(X_i) + nq(\theta) \right\} \\
&= \arg\max_{\theta} \left\{ p(\theta) K_n(X_1, \ldots, X_n) + S_n(X_1, \ldots, X_n) + nq(\theta) \right\},
\end{aligned}$$

where $K_n(X_1, \ldots, X_n) \equiv \sum_{i=1}^{n} K(X_i)$ and $S_n(X_1, \ldots, X_n) \equiv \sum_{i=1}^{n} S(X_i)$. Hence, in the case of distribution from the exponential family, knowing how to find the full data MLE directly translates in knowing how to find the observed data MLE: replace $K(X_i)$ with the conditional expectation $E_{\theta^k} \left[K(X_i) \mid Y_i \right]$ and iterate.

NPMLE based on censored data

Let $X_1, X_2, \ldots, X_n$ be n *i.i.d.* $X \sim F_X \in \mathcal{M}$, where $\mathcal{M}$ is non-parametric, and $Y_1, Y_2, \ldots, Y_n$ be n *i.i.d.* observations $Y = \phi(C, X) \sim P_{F_X, G}$. Assume CAR on G. As in the case when we were dealing

with finite dimensional real-valued parameter, we maximize the log-likelihood ratio of the observed data:

$$F_{X,n} = \arg\max_{F_X} \sum_{i=1}^{n} \log F_X\left(C\left(Y_i\right)\right).$$

Suppose we restrict to or know that F_X domain is a finite and discrete set of points $\{x_1, \ldots, x_m\}$: $F_X \in D = \left\{p = (p_1, \ldots, p_m) : \sum_{j=1}^{m} p_j = 1\right\}$. Hence, F_X is a multinomial distribution $M\left(n, p_1, \ldots, p_m\right)$. Let $n_j = \sum_{i=1}^{n} I\left(X_i = x_j\right)$. Thus we can use the general EM-alg, because the multinomial distribution is a member of the exponential family of distributions.

1. In step $k = 0$ initialize $p^0 = (1/m, \ldots, 1/m)$;
2. For $k = 1, 2, \ldots$ until convergence:

$$\begin{aligned} p^{k+1} &= \arg\max_{p} E_{p^k}\left[\log f_p\left(n_1, \ldots, n_m\right) \mid Y_1, \ldots, Y_n\right] \\ &= \arg\max_{p} E_{p^k}\left[\sum_{j=1}^{m} n_j \log p_j \mid Y_1, \ldots, Y_n\right] \\ &= \arg\max_{p} \sum_{j=1}^{m} E_{p^k}\left[n_j \mid Y_1, \ldots, Y_n\right] \log p_j. \end{aligned}$$

Setting

$$\frac{\partial}{\partial p_j}\left(\sum_{j=1}^{m} E_{p^k}\left[n_j \mid Y_1, \ldots, Y_n\right] \log p_j\right) = 0,$$

and using the constraint $\sum p_j = 1$ leads to the following result:

$$\begin{aligned} p_j^{k+1} &= \frac{1}{n} E_{p^k}\left[n_j \mid Y_1, \ldots, Y_n\right] \\ &= \frac{1}{n} E_{p^k}\left[\sum_{i=1}^{n} I\left(X_i = x_j\right) \mid Y_1, \ldots, Y_n\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} P_{p^k}\left[X_i = x_j | Y_i\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{P_{p^k}\left(X_i = x_j\right) I\left(x_j \in C\left(Y_i\right)\right)}{\sum_l p_l^k I\left(x_l \in C\left(Y_i\right)\right)},\ j = 1, \ldots, m \end{aligned}$$

which is an example of a *self-consistency equation* introduced in the previous lecture.

NPMLE found using EM-alg is not necessarily consistent. One such interesting example is the NPLME of the bivariate survival function subject to censoring. Typically, this kind of problems arise in studies of twins with a certain disease. Let $T = (T_1, T_2)$ be the bivariate survival time of a randomly drawn twin pair from a population with unknown survival distribution S_0 to be estimated. Assume that each pair is subject to right censoring which will be denoted with $C = (C_1, C_2)$. For each twin we observe the minimum of the censoring and survival time, as well as if the observation is censored or not:

$$Y_i \equiv \left(\tilde{T}_i, D_i\right),\ \tilde{T}_i = \left(\tilde{T}_{i1}, \tilde{T}_{i2}\right)^T, D_i = I\left(T_{i1} \leq C_{i1}, T_{i2} \leq C_{i2}\right),\ i = 1, \ldots, n.$$

The data can be thought of as occupying points, half-lines and quadrants in the plane as shown in fig. 1 below.

Assuming that the model for bivariate survival function is non-parametric, the NPMLE satisfies the self-consistency equations solved by the EM-alg. In the initialization step, the algorithm assigns mass of $1/n$ for each observation. Then for all censored pairs, i.e. those that are represented as half-lines and quadrants, their mass is redistributed over the associated region of coarsening $C(Y_i)$, according to an estimate of the conditional distribution which is obtained over all uncensored observations that fall into $C(Y_i)$. This is repeated with the new masses until the algorithm converges. If the time is on a continuous scale then the half-lines a.s. do not contain any (uncensored) observations, hence the conditional distribution for singly-censored observations can not be estimated properly, and as a consequence the NPMLE for bivariate right-censored data is inconsistent.

In series of papers Van der Laan[1994, 1995] proposed a way to repair the consistency of the NPMLE, by employing the following strategy: for each of the singly-censored observations replace the half-line with (half-)strip of width h parallel to the half-line (as shown in the picture) and estimate the conditional distribution from all uncensored observations with bigger time-to-event than that of the singly-censored observation, hence allowing the EM-alg to redistribute the mass $1/n$ over such a half-strip. This estimator is asymptotically efficient when $h \to \infty$ and is asymptotically unbiased even if h is fixed.

References

[1994] M.J. van der Laan (1994), Modified EM-estimator of the Bivariate Survival Function. **Mathematical Methods of Statistics**: 3, pp. 213-243.

[1995] M.J. van der Laan (1996), Efficient Estimation of the Bivariate Censoring Model and Repairing NPMLE. **Annals of Statistics**: 24(2), pp. 596-627.

Notes on Estimating Functions Daniel Rubin, April 15, 2004

What is an Estimating Function

Suppose that $X_1, ..., X_n \sim F_X \in M$ iid, that $\mu(F_X) \in R^k$ is a euclidean parameter of interest, and that $\eta(F_X)$ is a nuisance parameter. Then $D(X|\mu, \eta) \in R^k$ is an estimating function if $E_{F_X}[D(X|\mu(F_X), \eta(F_X)] = 0$. That is, if the function evaluated at the true parameters has expectation zero. If $\hat{\eta}$ is an estimate of η, then the estimating function estimate of $\mu(F_X)$ is the μ_n solving $\frac{1}{n}\sum_{i=1}^{n} D(X_i|\mu, \hat{\eta}) = 0$.

Background on Tangent Spaces

Let $\epsilon \to F_\epsilon$ be a one-dimensional parametric submodel of M, going through F_X at $\epsilon = 0$. Recall that the score vector of this submodel is defined as $s(X) = \frac{d}{d\epsilon} log f_\epsilon(X) \in L_0^2(F_X)$. Here $L_0^2(F_X)$ is the Hilbert space of all functions of X with mean zero and finite variance under F_X, and the tangent space $T(F_X)$ is defined as the linear closure of all possible scores in this Hilbert space. A very important space in efficiency theory is $T_{NUIS} \subset T$, called the nuisance tangent space, which is the linear closure in $L_0^2(F_X)$ of all scores of submodels F_ϵ such that $\frac{d}{d\epsilon}\mu(F_\epsilon) = 0$. The interpretation of such submodels is that local fluctuations from the truth only change η and not μ. Finally, we define the orthogonal complement of the nuisance tangent space as $T_{NUIS}^{\perp}(F_X) = \{h \in L_0^2(F_X) : \langle h(X), s(X) \rangle = 0$

$\forall s(X) \in T_{NUIS}(F_X)\}$, where $\langle h(X), s(X)\rangle = E_{F_X}[h(X)^T s(X)]$ is the inner product of h(X) and s(X) in $L_0^2(F_X)$.

Relationship Between Estimating Functions and Tangent Spaces

It has been shown by van der Laan and Robbins that in a very strong sense, the only estimating functions of interest are those that are members of $T_{NUIS}^{\perp}$. For such estimating functions $D(X|\mu,\eta)$, it can be shown that if F_ϵ is a submodel whose score is in the nuisance tangent space, then $\frac{d}{d\epsilon}E_{F_X}[D(X|\mu(F_\epsilon),\eta(F_\epsilon))] = 0$ at $\epsilon = 0$. This property can be used to show that if $\hat{\eta}$ is consistent for η, then under regularity conditions the solution μ_n of $\frac{1}{n}\sum_{i=1}^n D(X_i|\mu,\hat{\eta}) = 0$ is asymptotically linear for $\mu(F_X)$ with influence curve IC $= -[\frac{d}{d\epsilon}E_{F_X}[D(X|\mu(F_X),\eta(F_X)]^{-1}D(X|\mu(F_X),\eta(F_X))$.

Another property of the nuisance tangent space $T_{NUIS}^{\perp}$ is that it contains the linear span of all components of all gradients. Here a gradient is a random variable l(X) such that if F_ϵ has score s(X) then $lim_{\epsilon\to 0}\frac{1}{\epsilon}(\mu(F_\epsilon) - \mu(F_X)) = \langle l(X), s(X)\rangle$. In general there can be many different gradients, but the unique gradient in the tangent space $T(F_X)$ is called the canonical gradient. If an asymptotically linear estimator has an influence curve equal to the canonical gradient, then the estimator is efficient, meaning that it has better asymptotic performance than any other regular estimator.

Example: Survival Analysis with no Censoring

Let $T \sim F$ be a survival time, where M is a completely nonparametric model. Here F can be any cdf of a nonnegative random variable. Let $\mu(F) = F(t)$ be the parameter of interest, for fixed t. In this example there is no nuisance parameter η. Consider the parametric submodel $f_\epsilon(T) = (1 + \epsilon h(T))f(T)$, where $\int h(T)dF(T) = 0$ but h is otherwise arbitrary. Note that the score of this submodel is h(T), so the tangent space is all of $L_0^2(T)$, which immediately tells us that any asymptotically linear estimate is efficient because its influence curve must be in this tangent space. We now formally calculate $T_{NUIS}^{\perp}$, the orthogonal complement of the nuisance tangent space.

$\frac{d}{d\epsilon}\mu(F_\epsilon)|_{\epsilon=0} = lim_{\epsilon\to 0}\frac{1}{\epsilon}(\mu(F_\epsilon) - \mu(F)) = lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_0^t (f_\epsilon(T) - f(T))dT$
$= lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_0^t \epsilon h(T)f(T)dT = \int_0^t h(T)dF(T) = \int (I(T \le t) - F(t))h(T)dF(T)$
$= \langle I(T \le t) - F(t), h(T)\rangle$

Setting this derivative to zero and solving for h gives the nuisance tangent space. Thus, $T_{NUIS} = \{h(T) \in L_0^2(T) : h(T) \perp I(T \le t) - F(t)\}$. Taking the orthogonal complement of this space gives that $T_{NUIS}^{\perp} = [I(T \le t) - F(t)]$, where $[\bullet]$ denotes the linear span. Hence, $D(T|\mu) = I(T \le t) - \mu$ is in $T_{NUIS}^{\perp}$, and it is an estimating function because it clearly has mean zero under the truth. Because $T_{NUIS}^{\perp}$ is the span of a single random variable, we can check that using any estimating function from this space gives the same estimator, which is just $\mu_n = \frac{1}{n}\sum_{i=1}^n I(T_i \le t)$. We conclude from our previous results that μ_n is asymptotically linear (which we already could see without any efficiency theory), and it is efficient because it is the unique estimating function estimate over all estimating functions in $T_{NUIS}^{\perp}$.

Notes by Melinda Teng
April 14, 2004

Estimating functions in regression

Suppose

$$Y = m(Z \mid B) + \epsilon$$

where Y denote the outcome, Z is the vector of covariates, and B denote the regression parameters. Rewriting as $\epsilon(B) = Y - m(Z \mid B)$, we assume that $E[K(\epsilon \mid B) \mid Z] = 0$ for a given monotone increasing function $x \longrightarrow K(x)$. For example, when $K(\epsilon) = \epsilon$, we have $m(Z \mid B_0) = E[Y \mid Z]$; when $K(\epsilon) = I(\epsilon > 0) - 1/2$, we have $m(Z \mid B_0) = Med[Y \mid Z]$; when $K(\epsilon) = I(\epsilon > 0) - (1-p)$, we have $m(Z \mid B_0) = p - th$ quantile of $Y \mid Z$.

Let us now define

$$D_h(Y, Z, B) = h(Z)K(\epsilon(B))$$

to be the class of all estimating functions for B, and $B_n(h)$ be a solution of

$$\begin{aligned} 0 &= \sum_{i=1}^{n} D_h(Y_i, Z_i, B) \\ &= \sum_{i=1}^{n} h(Z_i)K(Y_i - m(Z_i \mid B)) \end{aligned}$$

where $h(.)$ can be any vector function.

Now,

$$B_n(h) - B_0 \approx \frac{1}{n} \sum_{i=1}^{n} C^{-1}(h) D_h(Y_i, Z_i, B_0)$$

where $C(h) = -\left[\frac{\delta}{\delta B} E_0[h(Z)K(\epsilon(B))]\right]$.

Let

$$h_{opt}(Z) = \frac{\frac{\delta}{\delta B} m(Z \mid B)}{Var[K(\epsilon) \mid Z]},$$

and h_n be an estimator of h_{opt} according to a (guessed) model of $E[K(\epsilon)^2 \mid Z]$. If B_n is a solution of

$$0 = \sum_{i=1}^{n} h_n(Z_i)K(\epsilon_i(B)),$$

then $B_n - B_0 \approx \frac{1}{n} \sum_{i=1}^{n} C^{-1}(h^\star) h^\star(Z_i) K(\epsilon_i(B))$ where $h^\star$ is the limit of h. In addition, note that B_n is asymptotically linear, and is an efficient estimator of B_0 if $h_n \longrightarrow h_{opt}$.

Right-censored data structure

Let $X(t)$ denote a time-dependent full data structure process of interest, and T denote the end-point of this process. A *full data structure* is defined as $X = \bar{X}(T) = (X(t) : t \leq T)$. The *observed data structure* (what we can observe) is defined as $O = (\tilde{T} = min(T, C), \triangle = I(T \leq C), \bar{X}(\tilde{T}))$ where C is a censoring variable. We denote the distribution of the observed data structure by $P_{F_X, G_{C|X}}$ where $G_{C|X}$ represents the distribution of the censoring mechanism.

Under CAR (Coarsening At Random) assumption, for $t < T$, we have

$$\lambda_C(t \mid X) = m(t, \bar{X}(t))$$

for some function m.

Lecture of April 19, 2004, Zheng Yin

4 Right Censored Data Structure (Cont'd)

Let $X(t)$ denote the time-dependent full data process of interest, and T be an endpoint of this process. Then the full data structure is defined as

$$X = \bar{X}(T) \equiv (X(t) : t \leq T) = (X(s), s \leq T) \sim F_X$$

We observe data $O = (\tilde{T} = \min(T, C), \Delta = I(T \leq C), \bar{X}(\tilde{T})) = \phi(C, X)$, where C is a censoring variable for a known ϕ. Then we have $O \sim P_{F_X, G_{C|X}}$.
The full data model: $F_X \in M^F$.

$$eg. M^F = \{F_X : E(\log T|X) = m(Z|\beta)\}, Z \in X(0)$$

Or if M^F is nonparametric, *eg.* (2). $M^F = \{F_X : \lambda_{T|Z}(t|Z) = \lambda_0(t)e^{\beta Z}\}$.

To find the full data distribution, we need the parameter of interest: $\mu : M^F \to \mathcal{R}^k$, *eg.*

$$\mu(F_X) = E_{F_X}(\log T|Z), \mu(F_X) = P_{F_X}(T > t_0), \mu(F_X) = \beta$$

CAR: for $t < T, \lambda_{C|X}(t|X) = m(t, \bar{X}(t))$ for some function m.
Let F_X be part of $f(o) = L_{F_X}(o)g(o|X)$, then the likelihood can be written as

$$L(F_X) = f_X(\bar{X}(T))^{\Delta} f_X(\bar{X}(\tilde{T}))^{1-\Delta} = [\prod_{t=0}^{T} P(X(t)|\bar{X}(t-))]^{\Delta} [\prod_{t=0}^{\tilde{T}} P(X(t)|\bar{X}(t-))]^{1-\Delta}$$

The content in the square bracket is the probability of seeing $X(t)$ given the past.

Example: Observed data $O : (\tilde{T}, \Delta, W)$, where W is baseline covariates. By CAR, $C \perp T|W$. Let $M^F = nonparametric, \mu(F_{T,W}) \equiv P(T > t_0)$. Then we can use Kaplan-Meier estimator $\hat{P}(T > t_0|W = w)$: because $E_W P(T > t_0|W) = P(T > t_0)$, then

$$\hat{P}(T > t_0) = \frac{1}{n} \sum_{i=1}^{n} \hat{P}(T > t_0|W = w_i) = \sum_{j=1}^{k} \frac{n_j}{n} \hat{P}(T > t_0|W = x_j)$$

when W is discrete with outcome $x_1, \ldots, x_k$.
However, when W is continuous or has many categories, the situation will be much more complicated. By the curse of dimensionality, we can only construct sensible estimators of $\mu(F_X)$ by making model assumptions on either $\lambda_C(t|X)$ or $f_X(x)$.
Define $A(t) = I(\tilde{T} \leq t, \Delta = 0)$ which is a counting process of censoring. Then the partial likelihood is given by

$$L_G(\tilde{T}, \Delta, \bar{X}(\tilde{T})) = \prod_{t}^{\tilde{T}} E(\partial A(t)|\bar{X}(t), \bar{A}(t-))^{\partial A(t)} (1 - E(\partial A(t)|\bar{X}(t-), \bar{A}(t-)))^{1-\partial A(t)}$$

We can model the intensity of $A(t)$ w.r.t. $(\bar{X}_A(t), \bar{A}(t-))$.
With a multiplicative intensity model:

$$E(\partial A(t)|\bar{A}(t-), \bar{X}(t)) = I(\tilde{T} \geq t)\lambda_0(t)e^{\alpha L(t)}$$

where $\lambda_0(t)e^{\alpha L(t)} = Pr(C = t|C \geq t, \bar{X}(t))$. Then $L(t) = f(\bar{X}(t))$ for some function f, $X_A(t) = X(\min(t, \tilde{T}))$.
If censoring $A(t)$ only jumps at discrete times $t_1 < t_2 < \cdots < t_M$, then using logistic model

$$E(\partial A(t)|\bar{A}(t_j-), \bar{X}(t_j)) = I(\tilde{T} \geq t)\frac{1}{1 + e^{-f(L(t_j)|\alpha)}}$$

where $-f(L(t_j)|\alpha)$ could be $-\alpha_1(t_j) + \alpha_2 L(t_j)$.
Note, by CAR, $E(\partial A(t)|\bar{A}(t) = 0, \bar{X}(t)) = P(C = t|C \geq t, \bar{X}(t)) = \lambda_{C|X}(t|X)$.

Class Notes: Merrill Birkner, April 21, 2004

Observed data:

$$O = (\tilde{T} = \min(T, C), \Delta = I(T \leq C), \bar{X}(\tilde{T})) = \Phi(C, X) \sim P_{F_x,G},$$

where $X(t) = (I(T \leq t), L(t))$.

The Full data model $\mathcal{M}^F$ and let $\mu(F_X)$ be the full data parameter.
For example:

1. $\mathcal{M}^F$ is nonparametric $\mu(F_X) = \rho(L_1, L_2)$
2. $\mathcal{M}^F = \{F_X : E_{F_X}(Y - m(Z|\beta_0)|Z) = 0\}$, $\mu(F_X) = \beta_0$
   This is equivalent to mean regression.
3. $\mathcal{M}^F = \{F_X : E_{F_X}(K(Y - m(Z|\beta_0)|Z) = 0)\}$, $\mu(F_X) = \beta_0$.
   This is equivalent, for example, to median/quantile regression[depending on K]
4. $\mathcal{M}^F = \{F_X : \lambda_{T|Z}(t|Z) = \lambda_0(t)exp(\beta_0 Z)\}$, $\mu(F_X) = \beta_0$
   This is the Cox proportional hazards model.

What is Y? It could be log survival, and therefore log(t).
In general $Y = f(\bar{X}(T))$ for some f
****

Question: If there is no censoring, what is the parameter of interest? What are full data estimating functions for the parameter of interest $\mu(F_X)$?
Full data estimating functions $D_h(X, \mu|\eta)$, where $h \in \mathcal{H}$ is an index ranging over an index set.

We are looking for the orthoganol complement of the nuisance scores in $L_0^2(P_0)$.
For example, if $\mathcal{M}^F = \{F_X : E_{F_X}(K(\epsilon(\beta_0))|Z) = 0\}$ and β_0 is the paramter of interest, then the class of estimating functions is given by

$$D_h(X, \beta) = h(Z)K(\epsilon(\beta)),$$

where h can be any function of Z.

In a nonparametric model with $\mu = P(T > t)$, there is only one estimating function $D(x, \mu) = I(T > t) - \mu$. Similarly, if in the nonparametric model the paraeter of interest is given by $\mu(F_x) = \rho(L_1, L_2) = \int E(L_1(t)L_2(t))w(t)dt$, then we can find the single estimating function as the influence curve of an ad hoc (empirical) estimator of this parameter.

$$\mathcal{M}^F = \{F_X : E(dN(t)|Z, \bar{N}(t)) = I(T > t)\lambda_0(t)exp(\beta Z)\},$$

where $N(t) = I(T \leq t)$ Now, the class of estimating functions is given by

$$D_h(X, \beta|\lambda) = \int h(t, Z)dM_{\lambda_{0,\beta}}(t),$$

where $dM(t) = dN(t) - E(dN(t)|Z, \bar{N}(t))$. This can be argued by noting that the model corresponds with a Bernoulli regression at each fixed t.

In the above equation, the nuisance parameter is the baseline hazard λ_0
Refer to Chapter 2 of van der Laan and Robins
****
What do we do when there is censoring?
*First it is important to understand the parameter of interest in the full data world.
We need the CAR assumption: $\lambda_{C|X}(t|X) = m(t, \bar{X}(t))$ for a function m.
Why don't we use this MLE based on a completely specified full data model? We are only interested in certain parameters and the MLE makes model assumptions. The person who does MLE did not model the parameter of interest. Putting down model of a small piece of data structure.
Marginal distribution of t: $\mu(F_x) = P(T > t_0)$

Notes re. Maximum Likelihood vs Previously Mentioned Method

- The Maximum Likelihood method factorizes the full data and censoring, but it does not care about censoring. This is equivalent to a Bayesian point of view where one does not care about the other part of the likelihood.
- We also know that the censoring is independent of survival and covariates. One who performs only the maximum likelihood method cannot do KM on all of those levels.
- The maximum likelihood is asymptotically efficient if it works, but the MLE in this case (when taking into account censoring and multiple covariates [with potentially many levels]) it is biased. This is refered to as the curse of dimensionality.
- The ML tries to be globally efficient and therefore it breaks down when the model becomes too high dimensional.
- KM within every cell unbiased and very variable. You increase the variance when you increase the dimensions.
- A person who performs the maximum likelihood method needs to model the error distribution. We do not have to assume a parametric model for the error distribution.
- Double Robust: If either the censoring model or likelihood is correct you get a consistent locally efficient estimator.

IPCW Estimating Functions:

$$D_h(X,\mu)\frac{\Delta}{\bar{G}(T|X)},$$

where $\bar{G}(t^-|X) = exp(-\int_0^t \lambda(s|x)ds$. We can estimate the censoring mechanism by modelling the censoring intensity $E(dA(t)|F_t)$ and fitting the corresponding partial likelihood $\Pi_t(E(dA(t)|F_t)^{dA(t)}(1-E(dA(t))|F_t)^{1-dA(t)}$.

It is a function of the observed data but after taking the conditional expectation, given X, you get back the estimating function of the full data. That is why it is unbiased, because it is unbiased in the full data world.

Estimating Functions for Right Censored Data, Keith Betts, April 23, 2004

Define the observed data structure as:

$$O = (\tilde{T} = min(T, O, \triangle = I(T \leq C), \bar{X}(\tilde{T}))) \sim P_{F_x}$$

Model is just as Model Right Censored data

$$\begin{aligned} X = (T, \bar{X}(T)) &\sim F_x \sim \mathcal{M}^F \\ C \mid X &\sim G_{C|X} \in G_{CAR} \\ \lambda_C(t \mid X) &= \mathcal{M}(t, \bar{X}(t)) \end{aligned}$$

Define $\mu(F_x)$ as a paramter defined on $\mathcal{M}^F$

IPCW-estimating functions

Define the full data estimating functions as:

$$D_h(X, \mu \mid \eta)$$

The IPCW estimate is

$$D_h(X, \mu \mid \eta)\frac{\triangle}{\bar{G}(T \mid X)}$$

where $\bar{G}(T \mid X) = \exp^{-\int_0^{\inf} \lambda_C(s|X)}$if C is continuous
and $\bar{G}(T \mid X) = \prod_{s\in(0,t)}(1 - \lambda_{C_X}(s \mid X))$ in general.

Suppose $D_h(X, \mu \mid \eta) = \int h(t, \bar{Z}(t))\partial M_{\beta,\lambda_0}$
where $\partial M_{\beta,\lambda_0} = \partial N(t) - E[\partial N(t) \mid \bar{N}(t-), \bar{Z}(t)]$

$\int h(t, \bar{Z}(t))$ is like the sum of unbiased estimates
Use Inverse Weighting

$$D_h(X, \mu \mid \eta) \int \left[h(t, \bar{Z}(t))\partial M_{\beta,\lambda_0} \frac{I(\tilde{T} \geq t)}{\bar{G}(t \mid X)} \right]$$

This corresponds to every line getting weight $\bar{G}(t \mid X)$
These weights are time dependent

In the continuous case, could use Cox PH with weights.

In discrete case (using logistic regression):
Define $N(t) = I(T \leq t)$ as the risk of jumping at time t

$$P(\partial N(t_j) = 1 \mid \bar{N}(t_j-) = 0, Z) \quad = \quad \underbrace{\frac{1}{1 + \exp^{\beta_0 + \beta_1 t_j + \beta_2 Z}}}_{\mathcal{M}(t, Z \mid \beta)}$$

The optimal estimating function for logistic regression is:

$$\sum_{t_j}^{T} \frac{\frac{\partial}{\partial \beta} \mathcal{M}(t_j, Z \mid \beta)}{\mathcal{M}(t_j, Z \mid \beta)(1 - \mathcal{M}(t_j, Z \mid \beta))} \left[\partial N(t_j) - \mathcal{M}(t_j, Z \mid \beta)\right]$$

Which can alternatively be written as:

$$\sum_{t_j}^{T} \frac{\frac{\partial}{\partial \beta} \mathcal{M}(t_j, Z \mid \beta)}{\mathcal{M}(t_j, Z \mid \beta)(1 - \mathcal{M}(t_j, Z \mid \beta))} \left[\partial N(t_j) - I(N(t_j))\right]$$

If events only happen at certain time points, glm solves this equation.
This is in the Full data world (No Censoring)
Suppose $N(t)$ is a repeated process

$$Pr(\partial N(t_j) = 1 \mid \bar{N}(t_j), Z) = \frac{1}{1 + \exp^{\beta_0 + \beta_1 t_j + \beta_2 Z + \beta_3 f(\bar{N}(t_j))}}$$

where $f(\bar{N}(t_j))$ is a function of the past
This is efficient in the full data world.

Suppose the counting process is not observed until the end of the study
That it was Censored in a random manner.
IPCW:

$$\sum_{j=0} \frac{I(\tilde{T} > t_j)}{\bar{G}(t_j \mid X)} \frac{\frac{\partial}{\partial \beta} \mathcal{M}(t_j, Z \mid \beta)}{\mathcal{M}(t_j, Z \mid \beta)(1 - \mathcal{M}(t_j, Z \mid \beta))} \left[\partial N(t_j) - I(N(t_j) \text{at risk})(\mathcal{M}(t_j, Z \mid \beta))\right]$$

Suppose the data was of the form:

$\partial N(t_j)$	t_j	Z	$W(t_j)$
0	.	.	.
0	.	.	.
1	.	.	.
0	.	.	.
$\vdots$	$\vdots$	$\vdots$	$\vdots$

In any software package, use GLM (logistic link)
Add weights to deal with censoring

Emperical mean of Estimating Equation

$$IC_0(O \mid G, D_h(. \mid \mu, \eta)) \equiv D_h(X, \mu \mid \eta)$$

This maps full data $\Rightarrow$ observed data
The mapping depends on the censoring mechanism.

Example: Suppose we wish to estimate the Marginal Survival function

$$\begin{aligned} \mu &= Pr(T > t) \\ \mu_n &= \sum_{i=1}^{n} \frac{I(T_i > t)\triangle_i}{\bar{G}(T_i \mid X)} \end{aligned}$$

There are three possible approaches

1. Use Kaplan-Meier to estimate μ
2. Use Kaplan-Meier to estimate $\bar{G}$
3. Use Cox PH to estimate $\bar{G}$

The third method is the most efficient, especially when there are covariates
Example:
Suppose age is a covariate
Censoring is completely independent
One should still include covariates important to the outcome in the model for $\bar{G}$.
Software Confidence Intervals are conservative (computed as if $\bar{G}$ known)
To get better Confidence Intervals either compute the Influence Curve or do the Bootstrap

Multiplicative Intensity Models

Matthew Sylvester

April 28, 2004

Counting Processes Data: $(\widetilde{T} = \min(T, C), \Delta = I(T \leq C), \bar{X}(\widetilde{T}))$, where T is the endpoint, C is the censoring time and $\bar{X}(\widetilde{T})$ is any data that might be collected on a person until time $\widetilde{T}$.

Suppose $X(t) = (N_1^*(t), \ldots, N_k^*(t), L(t))$,
where $N_h(t)$ is a counting process, h=1,. . . ,k denotes the index of that process and $*$ denotes that it is based on the full data. $L(t)$ is a time-dependent covariate process. Note that there might be several jumps when moving to the next state and that all counting processes stop jumping at T. That is, $N_h^*(t) = N_h^*(min(t, T))$.

Examples

1. *Consider multiple counting processes*

$$N_h^*(t) = I(T_h \leq t), h = 1, \ldots, k$$

where $T_1, \ldots, T_k$ refer to distinct counting processes (i.e. the time to certain events). For example, T_1 might refer to the time to AIDS while T_2 refers to the time to death. This might mean that the first process jumps when AIDS is contracted and that the second process jumps at death or T.

2. *Now consider just one counting process*

$$N^*(t) = \sum_{j=1}^{k} I(T_j \leq t)$$

that jumps whenever the events occur. For example, for asthmatic children, T_1 might be the time to the first attack, T_2 might be the time to the second attack, etc.

3. *Consider the Simplest Case*

$$N^*(t) = I(T \leq t)$$

that jumps at death.

Observed Data

$$N_h(t) = N_h^*(min(t, \widetilde{T})), \text{where } h = 1, \ldots, k$$

We have the history:

$$\mathcal{F}(t) = (\bar{X}(min(t, \widetilde{T})), \bar{A}(min(t, \widetilde{T})), A(t) = I(C \leq t))$$

When $t = \infty$, we see the full data past.

Multiplicative Intensity Models Suppose we have coarsening at random (CAR). We can get an intensity with respect to history when we are trying to model the probability of a counting process jumping given the past:

$N_h$with history $\mathcal{F}(t)$:

$$E(\partial N_h(t) \mid \mathcal{F}(t)) \equiv \lambda_h(t \mid \mathcal{F}(t)),$$

For $t < T$,

$$\lambda_c(t \mid X) = m(t, \bar{X}(t)),$$

which is reasonable if censoring is only determined by the past.
Then the intensity of the observed counting process is:

$$\begin{aligned} E(\partial N_h(t) \mid \mathcal{F}_t) &= I(\widetilde{T} \geq t) E(\partial N_h(t) \mid \mathcal{F}(t), \widetilde{T} \geq t) \\ &= I(\widetilde{T} \geq t) E(\partial N_h^*(t) \mid \bar{X}(t), \widetilde{T} \geq t, C \geq t) \\ &CAR \\ &= (I(\widetilde{T} \geq t) E(\partial N_h^*(t) \mid \bar{X}(t), \widetilde{T} \geq t) \end{aligned}$$

where we are conditioning on the full data past, and the censoring only depends on the past. We are assuming censoring does not depend on covariates not included, but this assumption becomes weaker as more covariates are put in the past. Note that the multiplicative intensity model can only be applied to the individual counting processes, of which every subject might have several (failure, infection, etc.)

If there are no covariates, the baseline hazard of dying now given that the person has not died yet and censoring has not occurred yet is given by:

$$P(T = t \mid T \geq t, C \geq t)$$

which is what we would estimate with the Cox proportional hazards model with no covariates. If there is independent censoring, this reduces to $P(T = t \mid T \geq t)$.

Now, if we would like to estimate the intensity with respect to a subset of the past:

$$E(\partial N_h(t) \mid \bar{N}_h(t-), Z(t-)),$$

where $(N_h(t), Z(t) \subset X(t))$.

If censoring is independent of the past, this is consistent, but inefficient. If we have CAR for the original data, we can use IPCW:

$$\int h(t), \bar{Z}(t-)) \partial M(t) \frac{I(\widetilde{T} \geq t)}{\bar{G}(t \mid X)}.$$

Or,

$$\int h(t), \bar{Z}(t-)) \partial M(t) \frac{I(\widetilde{T} \geq t) \bar{G}(t \mid \bar{Z}(t))}{\bar{G}(t \mid X)}$$

where we use cox proportional hazards to model the hazard of censoring mechanism. However, even when it is known that this is one, we should still estimate because it is more efficient.

Now, when we are adjusting for the entire past, the IPCW is not necessary. We implement it when we have just Z(t) where Z(t) is not the whole past. For example, we might just have the

treatment arm. Now, for the partial likelihood approach, we need to throw in everything that might be informative of censoring. We need to keep adjusting for confounding.

More on Multiplicative Intensity Models When the counting process is continuous(i.e. it can jump at any point in time), assume:

$$\begin{aligned}\lambda_h(t \mid \mathcal{F}(t-)) &= Y_h(t)E(\partial N_h(t) \mid \mathcal{F}(t), Y_h(t) = 1) \\ &= Y_h(t)\lambda_{0h}(t)exp(\beta_h Z_h(t)\end{aligned}$$

where $Y_h(t)$ is defined as the indicator that $N_h(t)$ is still at risk of jumping at time t, $\mathcal{F}(t)$ indicates the past, $\lambda_0 h(t)$ indicates the baseline hazard, $Z_h(t)$ is a function of $\mathcal{F}(t)$, and is composed of covariates thought predictive of the counting process jumping, and β_h denotes the regression coefficients.

For the special case of proportional hazards with one counting process and setting the history of censoring equal to $\bar{A}(min(t, \widetilde{T})$:

$$N(t) = I(\widetilde{T} \leq t, \Delta = 1)$$
$$\mathcal{F}(t) = (\bar{N}(t), W, \bar{A}(min(t, \widetilde{T}))$$

where we are only adjusting for the baseline covariates.

Then,

$$\begin{aligned}E(\partial N(t) \mid \mathcal{F}(t)) &= I(\widetilde{T} \geq t)\lambda_0(t)exp(\beta W) \\ &= I(\widetilde{T} \geq t)(\partial N(t) \mid T \geq t, C \geq t, W) \\ C \perp T \mid W & \\ &= P(T = t \mid T \geq t, W)\end{aligned}$$

which is the Cox proportional hazards model. Then, Cox proportional hazards assumes that $\lambda_{T|W}(t \mid W) = \lambda_0(t)exp(\beta W)$, a proportional hazards assumption. The ratio of two will not be dependent on time, a restrictive assumption. Our method is more general, and we do not need to make this assumption. Then W can be a function of time. This means that we can do data-adaptive work by pluggin on other basis functions for W, etc.

Maximum Likelihood Estimation We are interested in estimating the baseline hazards and coefficients β.

Give every counting process its own covariates. Then, we can create a long vector that has the real values when counting process 1 occurs and 0's when counting process 2 occurs, for example.

$$\lambda_h(t \mid \mathcal{F}(t-)) = Y_h(t)\lambda_0 h(t)exp(\beta Z_h(t)),$$

where we choose the covariate so that it reduces to what is wanted.

Example We have $\beta_1 W_1$ and $\beta_2 W_2$.
Then, set $Z_1 = (W_1, 0)$, $Z_2 = (0, W_2)$, and $\beta = (\beta_1, \beta_2)$
We then have,

$$\beta^T Z_1 = (\beta_1 \beta_2) \begin{pmatrix} W_1 \\ 0 \end{pmatrix} = \beta_1 W_1$$

$$\beta^T Z_2 = (\beta_1 \beta_2) \begin{pmatrix} 0 \\ W_2 \end{pmatrix} = \beta_2 W_2$$

The partial likelihood of $(N_1, \ldots, N_k)$ with respect to $\mathcal{F}(t-)$ is defined as:

$$L_P(\Lambda_h, \beta \mid \mathcal{F}(\inf)) = \prod_t \prod_{h=1}^{k} \lambda_h(t \mid \mathcal{F}(t))^{\partial N_h(t)} (1 - \lambda_\cdot(t \mid \mathcal{F}(t-))^{1-\partial N_\cdot(t)}$$

where:

$$N_\cdot(t) \equiv \sum_{h=1}^{k} N_h(t)$$

and

$$\lambda_\cdot(t \mid \mathcal{F}(t-)) = E(\partial N_\cdot(t) \mid \mathcal{F}(t-)) = \sum_{h=1}^{k} \lambda_h(t \mid \mathcal{F}(t-))$$

If something jumps, we use the intensity of the counting process; otherwise, we use one minus the intensity.

So, for the maximum likelihood estimate over the parameter fixing the coefficients and maximizing over the baseline hazards:

$$L_P(\Lambda_h, \beta \mid O_1, \ldots, O_n) = \prod_{i=1}^{n} \prod_t \prod_{h=1}^{k} \lambda_h(t \mid \mathcal{F}(t))^{\partial N_h(t)} (1 - \lambda_i(t \mid \mathcal{F}(t-))^{1-\partial N_i(t)}$$

and only keep track of counting processes at time t where $O_i, i = 1 \ldots n$ are the observations.