Introduction to Empirical Likelihood

Ian McKeague

November 8, 2019

P9111 Lecture Notes

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Outline

- Background on empirical likelihood (EL)
- Expanding the scope of EL
- Generalized EL Theorem
- EL with plug-in for nuisance parameters
- Background on survival analysis
- EL methods with censoring
- EL for Cox proportional hazards model

Background on Empirical likelihood

- Especially useful for finding nonparametric confidence regions without having to estimate standard errors.
- Thomas and Grunkemeier (1975) for survival function estimation. Owen (1988, 1990, . . . , 2001).
- First developed for finite-dimensional features $\theta = \theta(F)$ of a cdf (e.g., mean, median, cdf at a single point).
- Applies more generally to parameters identifiable from estimating equations.
- "Empirical likelihood" has over 100,000 Google hits.

Pros and cons of EL

Advantages	**Disadvantages**
adapts well to skewness (cf. bootstrap)	computationally more intensive than Wald type confidence regions
nonparametric	
better small sample performance than approaches based on asymptotic normality	asymptotic theory can be difficult to develop in semiparametric settings
confidence sets respect the range of the parameter	
often yields distribution-free tests (no need for simulation)	
regularity conditions are weak and natural (smoothness conditions often not needed)	
confidence regions are Bartlett correctable (unlike bootstrap) and transformation preserving	

Empirical cdf

Nonparametric likelihood

$$L(F) = \prod_{i=1}^{n} (F(X_i) - F(X_i-))$$

NPMLE $F_n = \arg\max_F L(F)$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\} = \mathbb{P}_n 1\{X \leq x\}$$

Nonparametric likelihood ratio

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i$$

where F places mass $p_i \geq 0$ on X_i, where $\sum_{i=1}^{n} p_i \leq 1$.

$\mathbb{P}_n$ is the NPMLE of P

First suppose there are no ties in the data.

$\mathbb{P}_n$ puts mass $\hat{p}_i = 1/n$ on X_i

$L(P) = 0$ unless P puts mass $p_i > 0$ on each X_i, and then

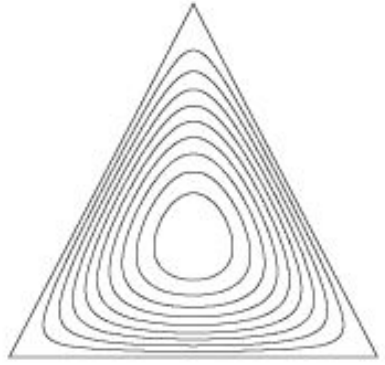
$$\log(L(P)/L(\mathbb{P}_n)) = \sum_{i=1}^{n} \log(p_i/\hat{p}_i) = n \sum_{i=1}^{n} \hat{p}_i \log(p_i/\hat{p}_i)$$

$$< n \sum_{i=1}^{n} \hat{p}_i \left(\frac{p_i}{\hat{p}_i} - 1 \right) \leq 0$$

since $\log x \leq x - 1$ with equality only when $x = 1$.

Same argument works when there are ties in the data.

Contours of likelihood ratio for $n = 3$



Picture shows the contours of $R(F)$ on the simplex (in $\mathbb{R}^3$):

$$\{(p_1, p_2, p_3) : p_i \geq 0, p_1 + p_2 + p_3 = 1\}$$

Lemma

If $R(F) \geq r_0 > 0$, then F places mass $m_n = O(1/n)$ outside $\{X_1, \ldots, X_n\}$.

Proof

$$r_0 \leq R(F) = \prod_{i=1}^{n} np_i \leq \prod_{i=1}^{n} n\left(\frac{1-m_n}{n}\right) = (1-m_n)^n$$

$$m_n \leq 1 - \exp(-n^{-1}\log(1/r_0)) \leq n^{-1}\log(1/r_0),$$

with the last inequality from $1 - x \leq e^{-x}$. □

- Justifies restricting to F supported by the data: $\sum_{i=1}^{n} p_i = 1$.

EL function

$$\mathrm{EL}_n(\theta_0) \quad = \quad \sup\{R(F) : \theta(F) = \theta_0\} = \frac{\sup\{L(F) : \theta(F) = \theta_0\}}{\sup\{L(F)\}}$$

(with $\sup \emptyset \equiv 0$)

EL hypothesis tests:

Accept $\theta(F) = \theta_0$ when $\mathrm{EL}_n(\theta_0) \geq r_0$ for some threshold r_0.

EL confidence regions:

$$\{\theta : \mathrm{EL}_n(\theta) \geq r_0\}$$

with r_0 chosen via an EL analogue of Wilks's theorem.

EL for means

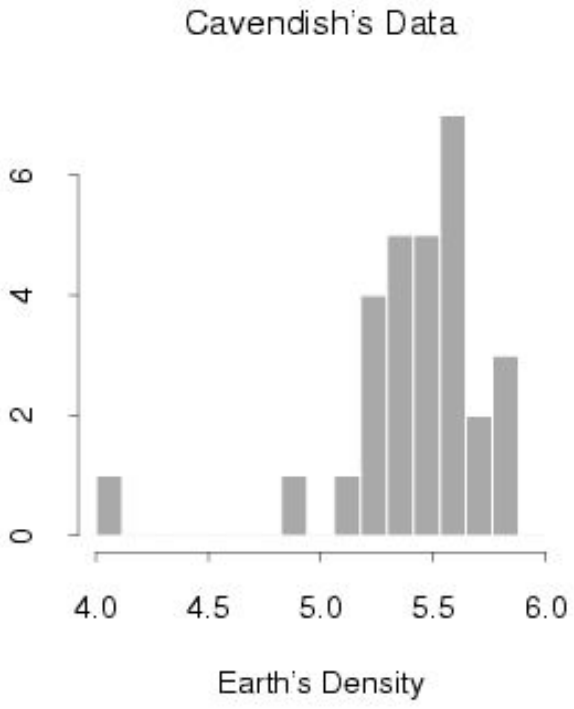
$\mu = E(X) \in \mathbb{R}^d$

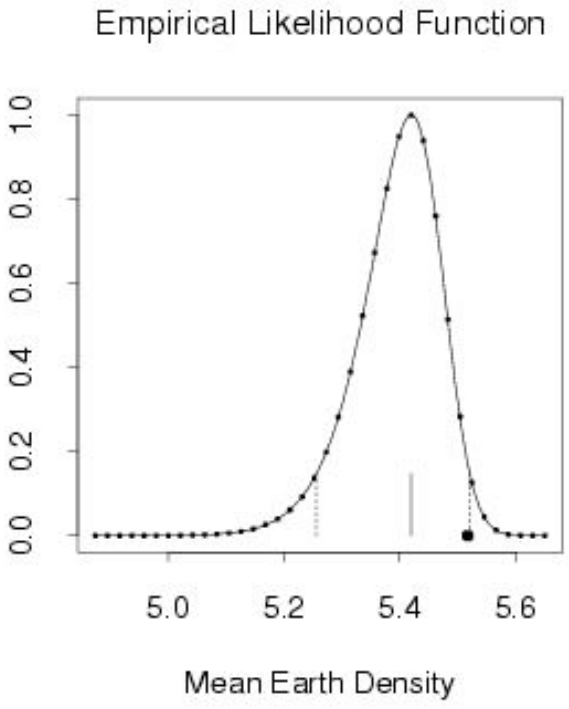
$$\mathrm{EL}_n(\mu) = \max\left\{ \prod_{i=1}^n np_i : \ \sum_{i=1}^n p_i X_i = \mu, p_i \geq 0, \sum_{i=1}^n p_i = 1 \right\}$$

Confidence region:

$$\{\mu : \mathrm{EL}_n(\mu) \geq r_0\} = \left\{ \sum_{i=1}^n p_i X_i : \ \prod_{i=1}^n np_i \geq r_0, p_i \geq 0, \sum_{i=1}^n p_i = 1 \right\}$$

Example



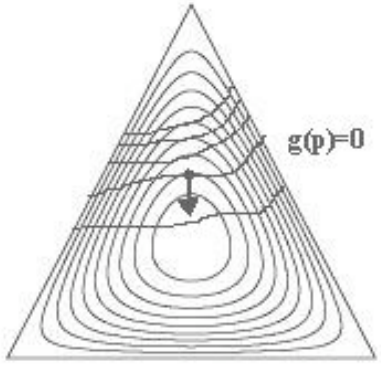


$\mathrm{EL}_n(\mu)$ (solid curve); 95% confidence limits (dotted bars); from Owen (2001).

Method of Lagrange multipliers

Maximize $f(x)$ subject to the (multivariate) constraint $g(x) = 0$.

Find $x^* = x^*(\lambda)$ maximizing $f(x) - \lambda^T g(x)$ such that $g(x^*) = 0$.



Then x^* solves the constrained problem.

Geometric intuition: At the maximum, ∇f and ∇g (when g is univariate) must be parallel:

$$\nabla f = \lambda \nabla g$$

for some constant λ (Lagrange multiplier).

Maximize $\log R(F) = \sum_{i=1}^{n} \log(np_i)$ under the constraints:

$$n \sum_{i=1}^{n} p_i(X_i - \mu) = 0, \quad 1 - \sum_{i=1}^{n} p_i = 0.$$

Let

$$G = \sum_{i=1}^{n} \log(np_i) - n\lambda \sum_{i=1}^{n} p_i(X_i - \mu) - \gamma \left(1 - \sum_{i=1}^{n} p_i\right)$$

where λ and γ are Lagrange multipliers.

$$\frac{\partial G}{\partial p_i} = \frac{1}{p_i} - n\lambda(X_i - \mu) + \gamma = 0$$

so $0 = \sum_{i=1}^{n} p_i \frac{\partial G}{\partial p_i} = n + \gamma$ giving $\gamma = -n$. Thus $p_i = \frac{1}{n} \frac{1}{1+\lambda(X_i-\mu)}$ and

$$g(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - \mu}{1 + \lambda(X_i - \mu)} = 0$$

This equation has a unique solution for $\lambda = \lambda(\mu)$.

Basic EL Theorem (Owen 1990)

$X_1, \ldots, X_n$ iid with finite mean μ_0, finite covariance matrix of rank $q > 0$. Then

$$-2 \log \mathrm{EL}_n(\mu_0) \xrightarrow{d} \chi^2_q.$$

Proof: Case $d = 1$. The Lagrange multiplier λ is the solution to

$$g(\lambda) = n^{-1} \sum_{i=1}^{n} \frac{X_i - \mu_0}{1 + \lambda(X_i - \mu_0)} = 0$$

and note that $g(0) = \bar{X} - \mu_0$. Denote $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \mu_0)^2$. Taylor expanding g gives

$$\begin{aligned} 0 &= g(\lambda) = g(0) + \lambda g'(0) + o_P(n^{-1/2}) \\ &= \bar{X} - \mu_0 - \lambda \hat{\sigma}^2 + o_P(n^{-1/2}) \end{aligned}$$

Thus $\lambda = (\bar{X} - \mu_0)/\hat{\sigma}^2 + o_P(n^{-1/2}) = O_P(n^{-1/2})$. Recall

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda(X_i - \mu_0)}$$

so, using the Taylor expansion $\log(1 + x) = x - x^2/2 + O(x^3)$,

$$\begin{aligned}
-2 \log \mathrm{EL}_n(\mu_0) &= -2 \sum_{i=1}^{n} \log(np_i) = 2 \sum_{i=1}^{n} \log(1 + \lambda(X_i - \mu_0)) \\
&= 2n\lambda(\bar{X} - \mu_0) - n\lambda^2 \hat{\sigma}^2 + o_P(1) \\
&= 2n(\bar{X} - \mu_0)^2/\hat{\sigma}^2 - n(\bar{X} - \mu_0)^2/\hat{\sigma}^2 + o_P(1) \\
&= n(\bar{X} - \mu_0)^2/\hat{\sigma}^2 + o_P(1) \\
&\xrightarrow{d} \chi^2_1
\end{aligned}$$

□

Calibration

This suggests the χ^2-**calibration** with threshold

$$r_0 = \exp(-\chi^2_{q,\alpha}/2)$$

for a $100(1-\alpha)\%$ confidence region; actual coverage $1-\alpha+O(n^{-1})$.

Bartlett correction

$$\left(1+\frac{a}{n}\right)\chi^2_{q,\alpha}$$

a involves higher-order moments of X, and needs to be estimated. Coverage improves to $1-\alpha+O(n^{-2})$.

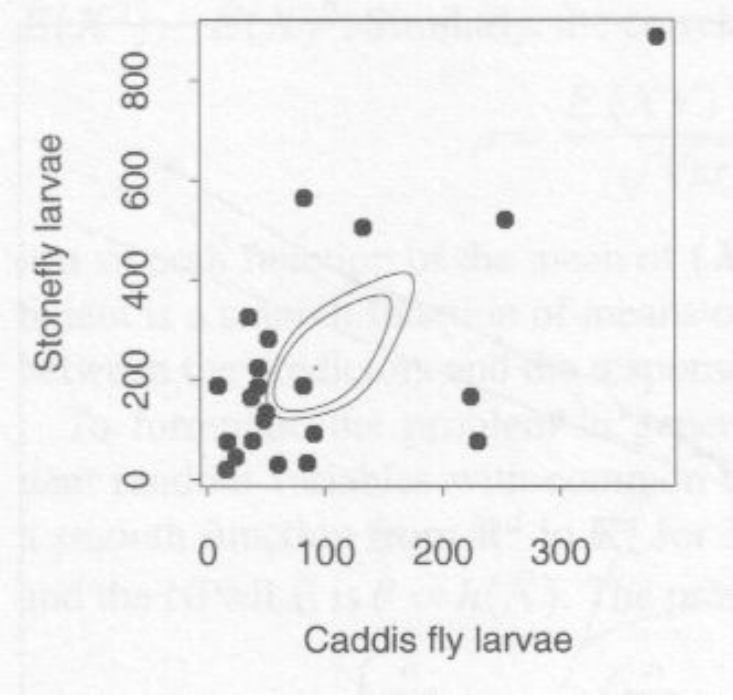
Bootstrap calibration

$X_1^*, \ldots, X_n^*$ iid from F_n. Simulation used to find the upper α-quantile of $-2 \log \mathrm{EL}_n^*(\bar{X})$, where

$$\mathrm{EL}_n^*(\bar{X}) = \max \left\{ \prod_{i=1}^n n p_i : \sum_{i=1}^n p_i X_i^* = \bar{X}, p_i \geq 0, \sum_{i=1}^n p_i = 1 \right\}$$

Example

Counts of two types of aquatic larvae at 22 locations in Wales.



Bivariate 95% confidence regions calibrated by χ^2 and by the bootstrap (larger region); from Owen (2001).

Expanding the scope of EL

- Linear functionals of F: $\theta = E(h(X)) = \int h(x)\, dF(x)$.
- Implicitly defined parameters: $E(m(X,\theta)) = 0$ where $m(X,\theta)$ is the estimating function; e.g., median, $m(X,\theta) = 1\{X \le \theta\} - .5$.

$$\mathrm{EL}_n(\theta) = \max\left\{\prod_{i=1}^n np_i : \sum_{i=1}^n p_i m(X_i,\theta) = 0, p_i \ge 0, \sum_{i=1}^n p_i = 1\right\}$$

- Functional parameters (e.g. F itself)
- Smooth functions of means: $\theta = h(\mu)$. EL-delta method?
- Nuisance parameters
- Non-iid data
- High-dimensional data
- Conditional estimating equations

Example: simultaneous band for F

Local EL function at $\theta_0 = F_0(t)$:

$$\begin{aligned}
\mathrm{EL}_n(t) &= \frac{\sup\{L(F) : F(t) = F_0(t)\}}{\sup\{L(F)\}} \\
&= \frac{\left(\frac{F_0(t)}{nF_n(t)}\right)^{nF_n(t)} \left(\frac{1-F_0(t)}{n(1-F_n(t))}\right)^{n(1-F_n(t))}}{\left(\frac{1}{n}\right)^n} \\
&= \left(\frac{F_0(t)}{F_n(t)}\right)^{nF_n(t)} \left(\frac{1-F_0(t)}{1-F_n(t)}\right)^{n(1-F_n(t))}.
\end{aligned}$$

Hence

$$\begin{aligned}
-2\log \mathrm{EL}_n(t) &= -2nF_n(t)\log\frac{F_0(t)}{F_n(t)} \\
&\quad -2n\left(1-F_n(t)\right)\log\frac{1-F_0(t)}{1-F_n(t)}.
\end{aligned}$$

Taylor expanding $\log(1+x) = x - x^2/2 + O(x^3)$ we have

$$-2\log \mathrm{EL}_n(t) \quad = \quad \left(\frac{\sqrt{n}(F_n(t) - F_0(t))}{\sqrt{F_0(t)(1-F_0(t))}} \right)^2 + o_P(1)$$

As a *process* in $t \in [a, b]$:

$$\begin{aligned} -2\log \mathrm{EL}_n(t) \quad &\xrightarrow{d} \quad \left(\frac{W^o(F_0(t))}{\sqrt{F_0(t)(1-F_0(t))}} \right)^2 \\ &\stackrel{d}{=} \quad \left(\frac{W(\sigma^2(t))}{\sigma(t)} \right)^2, \end{aligned}$$

W^o standard tied-down Wiener process (Brownian bridge)
W standard Brownian motion

$$\sigma^2(t) = \frac{F_0(t)}{1 - F_0(t)}.$$

Simultaneous confidence band over an interval $[a, b]$:

$$\{F_0 : -2 \log \mathrm{EL}_n(t) \leq C_\alpha, t \in [a, b]\}$$

C_α the upper α-quantile of

$$\sup_{t \in [\hat{\sigma}^2(a), \hat{\sigma}^2(b)]} \frac{W^2(t)}{t}.$$

Equal precision EL band.

Hollander, McKeague, Yang (1997)

Example: testing $F = F_0$

Under $H_0 : F = F_0$,

$$\begin{aligned} T_n &= -2 \int_{-\infty}^{\infty} \log \mathrm{EL}_n(t)\, dF_n(t) \\ &\xrightarrow{d} \int_0^1 \left(\frac{W^o(t)}{\sqrt{t(1-t)}} \right)^2 dt. \end{aligned}$$

T_n is asymptotically equivalent to the Anderson–Darling statistic

$$n \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))}\, dF_0(x).$$

Einmahl and McKeague (2003)

Example: testing for symmetry

$$H_0 : F(-x) = 1 - F(x-), \quad \text{for all } x > 0.$$

Local EL function:

$$\mathrm{EL}_n(x) = \frac{\sup\{L(F) : F(-x) = 1 - F(x-)\}}{\sup\{L(F)\}}, \quad x > 0.$$

Treat F as a function of $0 \le p \le 1$, where F puts mass

- $p/2$ on $(-\infty, -x]$, and on $[x, \infty)$
- $1 - p$ on $(-x, x)$

Point masses on observations in the respective intervals:

$$\frac{p/2}{n\hat{p}_1}, \frac{p/2}{n\hat{p}_2}, \frac{1-p}{n(1-\hat{p})},$$

$\hat{p} = \hat{p}_1 + \hat{p}_2$, $\hat{p}_1 = F_n(-x)$, $\hat{p}_2 = 1 - F_n(x-)$.

Maximum of

$$\left(\frac{p/2}{n\hat{p}_1}\right)^{n\hat{p}_1}\left(\frac{p/2}{n\hat{p}_2}\right)^{n\hat{p}_2}\left(\frac{1-p}{n(1-\hat{p})}\right)^{n(1-\hat{p})},$$

attained at $p = \hat{p}$.

$$\begin{aligned}\log \mathrm{EL}_n(x) &= n\hat{p}_1 \log\frac{\hat{p}}{2\hat{p}_1} + n\hat{p}_2 \log\frac{\hat{p}}{2\hat{p}_2} \\ &= nF_n(-x)\log\frac{F_n(-x)+1-F_n(x-)}{2F_n(-x)} \\ &\quad + n(1-F_n(x-))\log\frac{F_n(-x)+1-F_n(x-)}{2(1-F_n(x-))}\end{aligned}$$

Test statistic:

$$T_n = -2\int_0^\infty \log \mathrm{EL}_n(x)\, dG_n(x),$$

G_n is the empirical cdf of the $|X_i|$. If F is continuous, then under H_0

$$T_n \xrightarrow{d} \int_0^1 \frac{W^2(t)}{t}\, dt.$$

Einmahl and McKeague (2003): EL tests for symmetry, exponentiality, independence and changes in distribution.

Nuisance parameters

Now include a nuisance parameter h:

$$\mathrm{EL}_n(\theta, h) = \max\Big\{\prod_{i=1}^{n}(nw_i)\colon w_i > 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i m(X_i, \theta, h) = 0\Big\}.$$

Profile EL Theorem (Qin and Lawless, 1994)

For the iid case with $Em(X, \theta_0, h_0) = 0$ and m "sufficiently smooth" in $(\theta, h) \in \mathbb{R}^q \times \mathbb{R}^\ell$,

$$-2 \log \frac{\sup_h \mathrm{EL}_n(\theta_0, h)}{\sup_{\theta,h} \mathrm{EL}_n(\theta, h)} \to_d \chi^2_q$$

What can be done with *infinite* dimensional nuisance parameters?

EL with plug-in

Use a plug-in estimator $\widehat{h}$ in place of h (which can be arbitrary).

Generalized ELT

$$-2a_n^{-1} \log \mathrm{EL}_n(\theta_0, \widehat{h}) \to_d U^T V_2^{-1} U$$

provided

(A0) $P(\mathrm{EL}_n(\theta_0, \widehat{h}) = 0) \to 0$.

(A1) $U_n = \sum_{i=1}^n X_{ni} \to_d U$, where, for example, $U \sim \mathrm{N}_p(0, V_1)$.

(A2) $V_n = a_n \sum_{i=1}^n X_{ni} X_{ni}^T \to_P V_2$.

(A3) $a_n \max_{1 \le i \le n} \|X_{ni}\| \to_P 0$.

Here $X_{ni} = m_n(X_i, \theta_0, \widehat{h})$ has dimension p, a_n is bounded away from zero, and V_2 is a $p \times p$ positive definite covariance matrix.

Sketch of proof in the case $a_n = 1$

By the Lagrange multiplier argument,

$$\mathrm{EL}_n(\theta_0, \widehat{h}) = \mathrm{EL}_n = \prod_{i=1}^{n} (1 + \widehat{\lambda}' X_{ni})^{-1}$$

where $\sum_{i=1}^{n} X_{ni}(1 + \widehat{\lambda}^T X_{ni})^{-1} = 0$. In terms of the dual optimization problem:

$$-2 \log \mathrm{EL}_n = 2 \sum_{i=1}^{n} \log(1 + \widehat{\lambda}^T X_{ni}) = \max_{\lambda} G_n(\lambda),$$

where

$$\begin{aligned} G_n(\lambda) &= 2 \sum_{i=1}^{n} \log(1 + \lambda^T X_{ni}) \\ &= 2\lambda^T U_n - \lambda^T V_n \lambda + o_P(1) \\ \rightarrow_d & \quad G(\lambda) = 2\lambda^T U - \lambda^T V_2 \lambda \end{aligned}$$

uniformly over compacta.

It can be shown that

$$\widehat{\lambda} = \operatorname{argmax}_{\lambda} G_n(\lambda) = O_P(1)$$

and, via the proof of the argmax continuous mapping theorem,

$$\max_{\lambda} G_n(\lambda) \to_d \max_{\lambda} G(\lambda).$$

We conclude that

$$-2 \log \mathrm{EL}_n \to_d U^T V_2^{-1} U.$$

Hjort, McKeague and Van Keilegom (2007)

Remarks

- (A0) is the basic existence condition for EL to be useful: the zero vector has to be in the interior of the convex hull of

$$\{m_n(X_i, \theta_0, \widehat{h}),\ i = 1, \ldots, n\}.$$

- Owen's EL theorem follows using $m_n = m/\sqrt{n}$ and $a_n = 1$. For i.i.d. observations, (A0) holds by the Glivenko–Cantelli theorem over half-spaces (using the separating hyperplane theorem), (A1) by the CLT, (A2) by the WLLN, and (A3) by Borel–Cantelli.
- When $U \sim \mathrm{N}_p(0, V_1)$ with V_1 positive definite, the limit is

$$r_1\chi^2_{1,1} + \cdots + r_p\chi^2_{1,p}$$

where the $\chi^2_{1,j}$s are independent χ^2_1 and $r_1, \ldots, r_p$ are the eigenvalues of $V_2^{-1}V_1$.

Example: weakly dependent observations

- Suppose $\{X_i\}$ is a stationary sequence with a rapidly decaying mixing coefficient, and we have an unbiased estimating function $m(X, \theta)$ without a nuisance parameter. Setting $m_n = m/\sqrt{n}$ and $a_n = 1$, (A1) can be checked using a CLT for stationary sequences; (A2) follows from the ergodic theorem.

- $-2 \log \mathrm{EL}_n(\theta_0) \to_d r\chi_1^2$ where

$$r = \sum_{i=1}^{\infty} \mathrm{Corr}\{m(X_1, \theta_0), m(X_i, \theta_0)\}.$$

- Kitamura (1997) showed that blockwise EL has greater efficiency than the "naive" EL in this setting.

Example: long range dependence

- Estimate the mean θ_0 of the stationary ergodic process $X_i = G(Z_i)$, where G is a Borel function and $\{Z_i\}$ is a mean-zero, unit-variance, stationary Gaussian process such that

$$\mathrm{Cov}(Z_i, Z_{i+n}) = n^{-\alpha} L(n)$$

  for some $0 < \alpha < 1$ and slowly varying $L(\cdot)$.

- Estimating function: $m_n(X_i, \theta) = b_n(X_i - \theta)$, where b_n depends on the (slower than $\sqrt{n}$) rate of convergence of the sample mean.

Nordman, D., Sibbertsen, P. and Lahiri, S.N. (2007). *Journal of Time Series Analysis* 28, 576–599.

- Condition (A1) is checked using a result of Taqqu (1975):

$$b_n \sum_{i=1}^{n} (X_i - \theta_0) \to_d U$$

where $b_n = n^{\alpha/2-1} L(n)^{-1/2}$ and U is a certain multiple Wiener integral.

- Condition (A2) is checked by setting

$$a_n = n^{-1} b_n^{-2} = n^{1-\alpha} L(n)$$

and using the ergodic theorem:

$$a_n \sum_{i=1}^{n} m_n(X_i, \theta_0)^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta_0)^2 \to_{\text{a.s.}} V_2.$$

- In this case the choice of a_n tends to infinity, and it is not possible to arrange $a_n = 1$.

Example: symmetric cdf F

Estimate $\theta_0 = F(x)$ using n i.i.d. observations from a cdf F that is symmetric about a, using plug-in of the sample median $\widehat{a}$.

$$m_n(X, \theta, a) = n^{-1/2} \begin{pmatrix} 1\{X \le x\} - \theta \\ 1\{X > 2a - x\} - \theta \end{pmatrix}.$$

Let $\eta_0 = \min(\theta_0, 1 - \theta_0)$ and suppose $0 < \theta_0 < 1$. Condition (A2) holds with

$$V_2 = \begin{pmatrix} \theta_0(1-\theta_0) & -\eta_0^2 \\ -\eta_0^2 & \theta_0(1-\theta_0) \end{pmatrix}.$$

when $\theta_0 \neq 1/2$. Note that V_2 is singular when $\theta_0 = 1/2$.

To check (A1), note that

$$
\begin{aligned}
& n^{1/2}[1 - \widehat{F}(2\widehat{a} - x) - \theta_0] \\
&= n^{1/2}[1 - F(2\widehat{a} - x) - \widehat{F}(2a - x) + F(2a - x) - \theta_0] + o_P(1) \\
&= n^{1/2}[1 - \widehat{F}(2a - x) - \theta_0] - 2f(2a - x)n^{1/2}(\widehat{a} - a) + o_P(1) \\
&= n^{1/2}[1 - \widehat{F}(2a - x) - \theta_0] - 2f(x)f(a)^{-1}n^{1/2}[\widehat{F}(a) - 1/2] + o_P(1)
\end{aligned}
$$

provided $f(a) > 0$. Use the CLT to conclude that $U \sim \mathrm{N}_2(0, V_1)$ with

$$
V_1 = \begin{pmatrix} \theta_0(1-\theta_0) & -\eta_0^2 - f(x)f(a)^{-1}\eta_0 \\ -\eta_0^2 - f(x)f(a)^{-1}\eta_0 & \theta_0(1-\theta_0) + f(x)^2[f(a)]^{-2} + 2f(x)f(a)^{-1}\eta_0 \end{pmatrix}.
$$

Example: integral of a squared density

Of interest for various problems related to nonparametric density estimation:

$$\theta_0 = \int f_0^2 \, dx.$$

Estimating function

$$m(X, \theta, f) = f(X) - \theta$$

with plug-in of $\widehat{f}$, a kernel density estimator having a symmetric kernel and bandwidth $b = b_n$.

Our result yields

$$-2 \log \mathrm{EL}_n(\theta_0, \widehat{f}) \to_d 4\chi_1^2$$

Example: density estimation

$X_1, \ldots, X_n$ i.i.d. density f_0, $\theta_0 = f_0(t)$ for a fixed t.

Hall and Owen (1993) constructed EL confidence bands for $f_0(t)$.

Chen (1996) showed that the pointwise EL confidence intervals are more accurate than those based on the bootstrap.

Estimating function

$$m_n(x, \theta) = n^{-1/2} b^{1/2} \{k_b(x - t) - \theta\}$$

where $k_b(u) = b^{-1} k(b^{-1} u)$, and k is a symmetric, bounded kernel function supported on $[-1, 1]$. Suppose the bandwidth $b = b_n$ satisfies $nb \to \infty$ and $nb^5 \to 0$.

(A2) can be checked under mild conditions on the density, as it follows from standard asymptotic theory for kernel density estimators that

$$\sum_{i=1}^{n} m_n(X_i, \theta_0) = (nb)^{1/2}\{\widehat{f}_n(t) - f_0(t)\} \to_d \mathrm{N}(0, V_1),$$

where

$$V_1 = f_0(t)R(k) \quad \text{and} \quad R(k) = \int k(u)^2 \, \mathrm{d}u.$$

For (A3),

$$\begin{aligned} \sum_{i=1}^{n} m_n^2(X_i, \theta_0) &= \frac{b}{n}\sum_{i=1}^{n}\{k_b(X_i - t) - \theta_0\}^2 \\ &= \frac{1}{nb}\sum_{i=1}^{n} k((X_i - t)/b)^2 + O_P(b) \to_P f_0(t)R(k) = V_1. \end{aligned}$$

Conclude that

$$-2 \log \mathrm{EL}_n(\theta_0) \to_d \chi_1^2$$

Survival analysis background

$T \sim F$ represents a lifetime

Survival function: $S = 1 - F$, $S(0) = 1$

Cumulative hazard function (chf):

$$A(t) = \int_{(0,t]} \frac{dF(s)}{1 - F(s-)}$$

There is a 1-1 correspondence between survival functions and cumulative hazards.

If F is continuous: $S = \exp(-A)$, $A = -\log(S)$.

Lemma

If F is a discrete cdf, the corresponding cumulative hazard function is

$$A(t) = \sum_{s \leq t} \frac{\Delta F(s)}{1 - F(s-)}$$

where $\Delta F(t) = F(t) - F(t-)$ is the jump in F at t. Conversely, if A is a discrete chf, the corresponding survival function is

$$S(t) = \prod_{s \leq t} (1 - \Delta A(t)).$$

Proof

Given a discrete chf A, write $S(t) = \prod_{s \le t}(1 - \Delta A(t))$. Then S has chf A, because $S(t-) = S(t)/(1 - \Delta A(t))$ and

$$\Delta A(t) = 1 - \frac{S(t)}{S(t-)} = \frac{\Delta F(s)}{1 - F(s-)}.$$

Conversely, given a discrete survival function S, then

$$\begin{aligned} S(t) &= \prod_{u \le t} \frac{S(u)}{S(u-)} = \prod_{u \le t} \left(1 + \frac{\Delta S(u)}{S(u-)}\right) \\ &= \prod_{u \le t} (1 - \Delta A(u)) \end{aligned}$$

where A is the chf. □

Hazard functions

If $T \sim F$ has density f, define the hazard function

$$\alpha(t) = A'(t) = f(t)/S(t) \approx P(T \in [t, t+dt) | T \geq t)/dt$$

Thus

$$P(T \in [t, t+dt) | T \geq t) \approx \alpha(t)\, dt$$

Cox proportional hazards model

$$\alpha(t|Z) = \lambda_0(t) e^{\beta^T Z}$$

adjusts for a (p-dimensional) covariate Z.

Independent right-censoring

$X = \min(T, C),\ \delta = 1\{T \leq C\}$
T and C independent

Counting process:

$$N(t) = 1\{X \leq t, \delta = 1\}$$

At risk indicator: $Y(t) = 1\{X \geq t\}$

$M(t) = N(t) - \int_0^t Y(s)\alpha(s)\,ds$ is a martingale:

$dN(t) \sim$ Bernoulli$(Y(t)\alpha(t)\,dt)$ given the past $\mathcal{F}_t$, so

$$E(dM(t)|\mathcal{F}_t) = E(dN(t) - Y(t)\alpha(t)\,dt|\mathcal{F}_t) = 0$$

Quadratic variation:

$$\langle M\rangle(t) = \int_0^t E(dM(s)^2|\mathcal{F}_s) = \int_0^t Y(s)\alpha(s)\,ds$$

EL for right-censored data (no covariates)

Nonparametric likelihood

$$L(F) = \prod_{i=1}^{n} (F(X_i) - F(X_i-))^{\delta_i} (1 - F(X_i))^{1-\delta_i}$$

Note: this is a *partial* likelihood—the full likelihood is the product of $L(F)$ and a similar expression involving the cdf G of C.

EL function

$$\mathrm{EL}_n(\theta_0) = \frac{\sup\{L(F) : \theta(F) = \theta_0\}}{\sup\{L(F)\}}$$

where the maximization is restricted to cdfs F supported by the *uncensored* lifetimes.

$L(F)$ in terms of the chf

Order the uncensored lifetimes: $0 < T_1 \leq \ldots \leq T_k$, $T_0 = 0$

$h_j = \Delta A(T_j) = 1 - S(T_j)/S(T_{j-1})$ jump in chf at T_j

$r_j = \sum_{i=1}^{n} 1\{X_i \geq T_j\}$ size of the risk set at T_j-, with $r_{k+1} = 0$.

$d_j \geq 1$ denotes the number of uncensored failures at T_j.

Lemma If F is supported by the uncensored lifetimes, then

$$L(F) = \prod_{j=1}^{k} h_j^{d_j} (1 - h_j)^{r_j - d_j}$$

Proof

Note that the number of censored lifetimes in $[T_j, T_{j+1})$ is $r_j - d_j - r_{j+1}$, so

$$
\begin{aligned}
L(F) &= \prod_{i=1}^{n} (S(X_i-) - S(X_i))^{\delta_i} (S(X_i))^{1-\delta_i} \\
&= \left\{ \prod_{j=1}^{k} (S(T_j-) - S(T_j))^{d_j} \right\} \left\{ \prod_{j=1}^{k} S(T_j)^{r_j - d_j - r_{j+1}} \right\} \\
&= \left\{ \prod_{j=1}^{k} h_j^{d_j} S(T_{j-1})^{d_j} \right\} \left\{ \prod_{j=1}^{k} \frac{S(T_j)^{r_j - d_j}}{S(T_{j-1})^{r_j}} \right\} \\
&= \prod_{j=1}^{k} h_j^{d_j} (1 - h_j)^{r_j - d_j}
\end{aligned}
$$

Nonparametric MLEs

$L(S)$ is maximized when $h_j = d_j/r_j$, giving the Nelson–Aalen estimator:

$$A_n(t) = \sum_{j:T_j \le t} \frac{d_j}{r_j}$$

Kaplan–Meier estimator:

$$S_n(t) = \prod_{j:T_j \le t} \left(1 - \frac{d_j}{r_j}\right)$$

and $F_n = 1 - S_n$.

Limit distributions

Assume now F is continuous. Then

$$\sqrt{n}(A_n(t) - A(t)) \xrightarrow{d} W(\sigma^2(t))$$

$$\sqrt{n}(S_n(t) - S(t)) \xrightarrow{d} S(t)W(\sigma^2(t))$$

where

$$\sigma^2(t) = \int_0^t \frac{dF(s)}{(1 - F(s))^2(1 - G(s-))}$$

Without censoring, simplifies to

$$\sigma^2(t) = \frac{F(t)}{1 - F(t)}.$$

Counting process approach

In counting process notation, the Nelson–Aalen estimator is

$$A_n(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\sum_{i=1}^{n} Y_i(s)}$$

where $dN_i(s) = Y_i(s)\alpha(s)\,ds + dM_i(s)$. Thus

$$\sqrt{n}(A_n(t) - A(t)) = U_n(t) + o_P(1)$$

where

$$U_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t \frac{dM_i(s)}{\mathbb{P}_n Y(s)}$$

is a martingale with quadratic variation

$$\langle U_n \rangle(t) = \mathbb{P}_n \int_0^t \frac{Y(s)\alpha(s)\,ds}{(\mathbb{P}_n Y(s))^2} \xrightarrow{p} \int_0^t \frac{\alpha(s)\,ds}{EY(s)} = \sigma^2(t).$$

EL for the survival function at a fixed point

$\theta_0 = S(t)$, with t fixed. The estimate $\hat{S}$ maximizing $L(S)$ subject to the constraint $S(t) = \theta$ is

$$\hat{S}(t) = \prod_{j:T_j \le t} \left(1 - \frac{d_j}{r_j + \lambda}\right)$$

where the Lagrange multiplier λ is the solution to

$$\prod_{j:T_j \le t} \left(1 - \frac{d_j}{r_j + \lambda}\right) \quad = \quad \theta.$$

Equivalently,

$$g(\lambda) = \sum_{j:T_j \le t} \log\left(1 - \frac{d_j}{r_j + \lambda}\right) = \log\theta = -A(t)$$

Theorem If F is continuous, $0 < \theta_0 = S(t) < 1$, $G(t) < 1$, then

$$-2\log \mathrm{EL}_n(\theta_0) \xrightarrow{d} \chi^2_1$$

Thomas and Grunkemeier (1975), Li (1995), Murphy (1995)

Proof: Taylor expansion of g leads to

$$\lambda = n(A(t) - A_n(t))/\hat{\sigma}^2 + O_P(1)$$

where $\hat{\sigma}^2$ is an estimate of $\sigma^2(t)$.

$$\begin{aligned}
-2\log \mathrm{EL}_n(\theta_0) &= -2(\log(L(\hat{S}) - \log(L(S_n)) \\
&= -2\sum_{i:T_j \le t} \left\{ (r_j - d_j)\log\left(1 + \frac{\lambda}{r_j - d_j}\right) \right. \\
&\qquad \left. -r_j \log\left(1 + \frac{\lambda}{r_j}\right) \right\} \\
&= \lambda^2 \hat{\sigma}^2/n + o_P(1) \\
&= n(A_n(t) - A(t))^2/\hat{\sigma}^2 + o_P(1) \\
&\xrightarrow{d} \chi^2_1
\end{aligned}$$

Simultaneous EL band for S

As a *process* in $t \in [a, b]$,

$$-2 \log \mathrm{EL}_n(S(t)) \quad \xrightarrow{d} \quad \left(\frac{W(\sigma^2(t))}{\sigma(t)} \right)^2,$$

Simultaneous confidence band for S over an interval $[a, b]$:

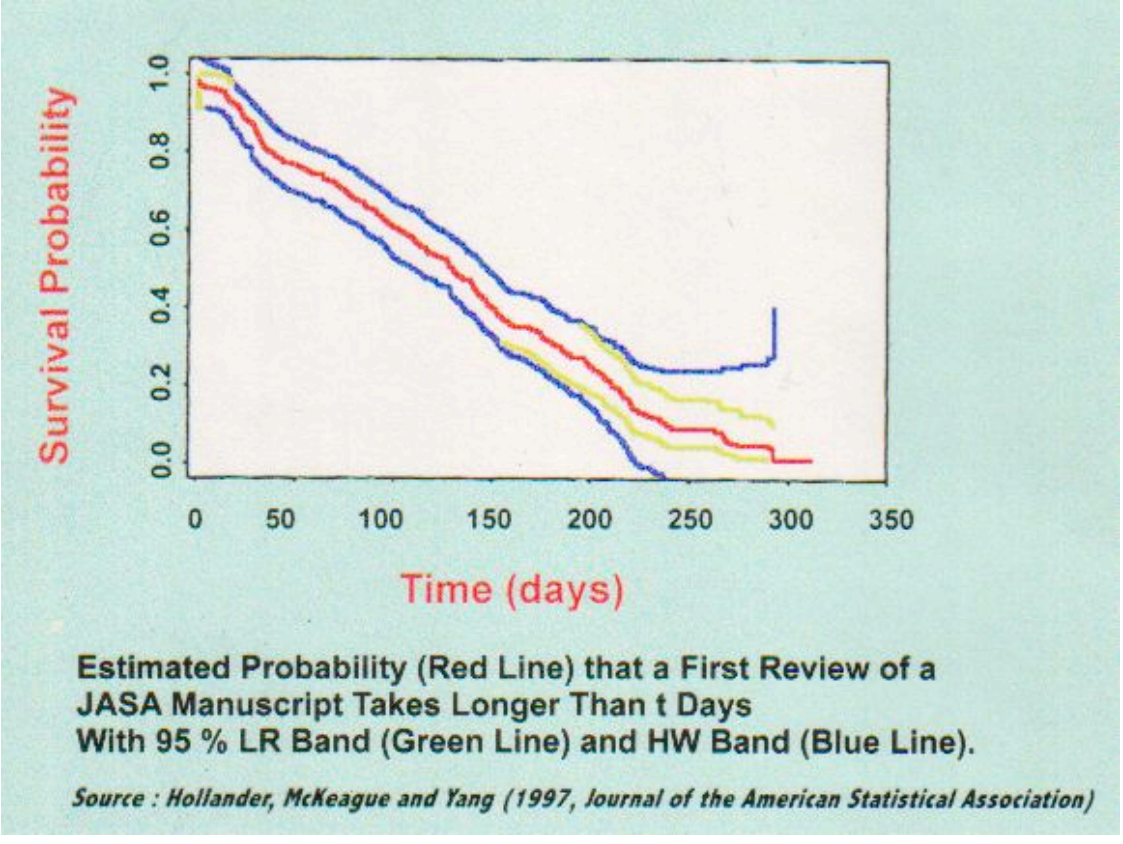
$$\{S(t) : -2 \log \mathrm{EL}_n(S(t)) \leq C_\alpha, t \in [a, b]\}$$

C_α the upper α-quantile of

$$\sup_{t \in [\hat{\sigma}^2(a), \hat{\sigma}^2(b)]} \frac{W^2(t)}{t}$$

Equal precision band. Hollander, McKeague, Yang (1997)

Example: Data on 432 manuscripts submitted to JASA during 1994. Time to first review censored by the end of the year.



Estimated Probability (Red Line) that a First Review of a JASA Manuscript Takes Longer Than t Days With 95 % LR Band (Green Line) and HW Band (Blue Line).

Source : Hollander, McKeague and Yang (1997, Journal of the American Statistical Association)

EL for Cox regression parameters

$$\alpha(t|Z) = \lambda_0(t)e^{\beta^T Z}$$

Estimating function:

$$m(\beta, s^{(0)}, s^{(1)}) = \int_0^\tau \left(Z - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right) dN(t)$$

is an unbiased estimating function with (functional) nuisance parameters $s^{(j)}(\beta, t) = E[Y(t)Z^j e^{\beta^T Z}]$ for $j = 0, 1$.

Reason: $dN(t) = dM(t) + Y(t)\lambda_0(t)e^{\beta_0^T Z}\, dt$, and $\int_0^\tau (\cdots)\, dM(t)$ is a martingale in τ.

Plug-in estimator $\widehat{h}$ is $S^{(j)}(\beta, t) = \mathbb{P}_n[Y(t)Z^j e^{\beta^T Z}], j = 0, 1$

Theorem Under conditions of Andersen and Gill (1982),

$$-2 \log \mathrm{EL}_n(\beta_0, \widehat{h}) \to_d \chi^2_p$$

Proof: Let $X_{ni} = m(\beta_0, \widehat{h})/\sqrt{n}$, and apply the

Generalized ELT

$$-2 \log \mathrm{EL}_n(\beta_0, \widehat{h}) \to_d U^T V_2^{-1} U$$

provided

(A0) $P(\mathrm{EL}_n(\beta_0, \widehat{h}) = 0) \to 0$.

(A1) $U_n = \sum_{i=1}^n X_{ni} \to_d U \sim \mathrm{N}_p(0, V_1)$.

(A2) $V_n = \sum_{i=1}^n X_{ni} X_{ni}^T \to_P V_2$.

(A3) $\max_{1 \le i \le n} \|X_{ni}\| \to_P 0$.

U_n is a martingale in τ:

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\tau \left(Z_i - \frac{S^{(1)}(\beta_0, t)}{S^{(0)}(\beta_0, t)} \right) \, dM_i(t) \to_d \mathrm{N}_p(0, \Sigma)$$

by the martingale CLT.

$$\begin{aligned} V_n &= \mathbb{P}_n \int_0^\tau \left(Z - \frac{S^{(1)}(\beta_0, t)}{S^{(0)}(\beta_0, t)} \right)^{\otimes 2} \, dN(t) \\ &= \mathbb{P}_n \int_0^\tau \left(Z - \frac{S^{(1)}(\beta_0, t)}{S^{(0)}(\beta_0, t)} \right)^{\otimes 2} Y(t) \lambda_0(t) e^{\beta_0^T Z} \, dt + o_P(1) \\ &= \langle U_n \rangle_\tau + o_P(1) \xrightarrow{p} \Sigma. \end{aligned}$$

Thus $V_1 = V_2 = \Sigma$, so the EL statistic has a χ^2_p limit. □

Cox model with time-dependent coefficients

$$\alpha(t|Z) = \lambda_0(t)e^{\beta(t)^T Z}$$

$\theta_0 = \beta(t)$ for some fixed t

As in the density estimation example, localize the estimating function in a neighborhood of t using a kernel k_b:

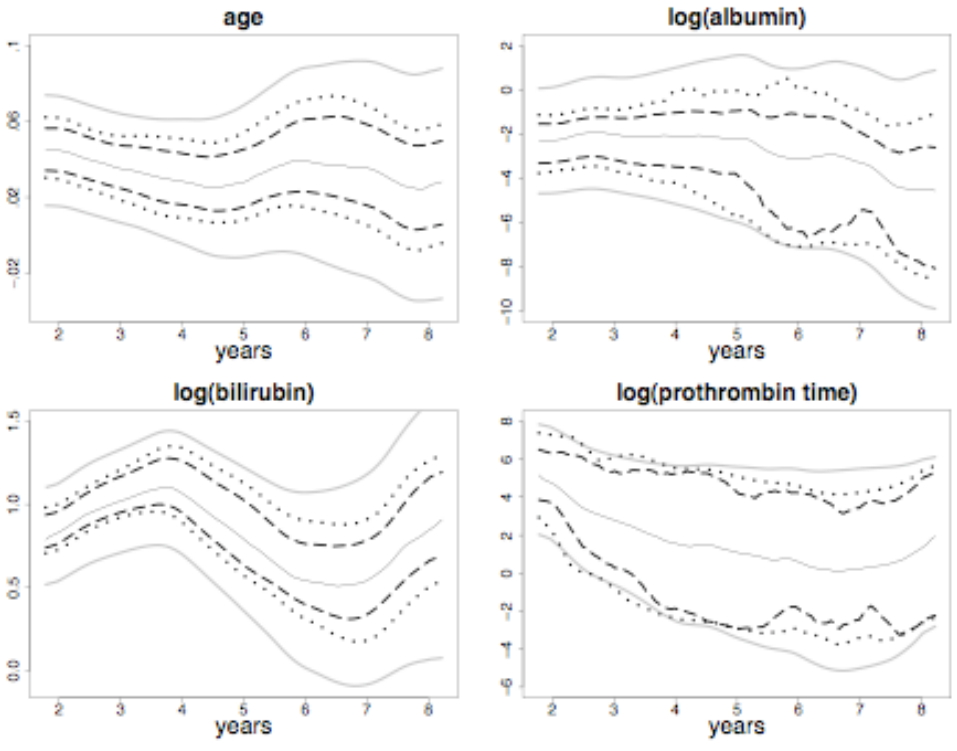
$$m_n(\theta_0, s^{(0)}, s^{(1)}) = \sqrt{\frac{b}{n}} \int_0^\tau k_b(u-t) \left(Z - \frac{s^{(1)}(\theta_0, u)}{s^{(0)}(\theta_0, u)} \right) dN(u)$$

Generalized ELT: plug-in EL statistic has a χ^2_p limit.

Sun, Sundaram, Zhao (2007): simultaneous EL band for $\beta(\cdot)$

Example: Mayo Clinic primary biliary cirrhosis study

Right-censored survival times of 416 patients.



Pointwise CI (dashed), simultaneous band (dotted), Cai and Sun (2003) (solid grey)

From Sun, Sundaram, Zhao (2007)

EL for current status data

$T \sim F$ failure time, $S = 1 - F$, pdf f, $\theta_0 = S(t)$
$C \sim G$ check-up time, assumed independent of T, pdf g

Only get to observe $X = (C, \Delta)$ where $\Delta = 1\{T \leq C\}$.

Nonparametric likelihood:

$$\begin{aligned} L(S) &= \prod_{i=1}^{n} (1 - S(C_i))^{\Delta_i} S(C_i)^{1-\Delta_i} \\ S_n(t) &= \arg\max_S L(S). \end{aligned}$$

Groeneboom (1987) showed

$$n^{1/3}(S_n(t) - S(t)) \to_d 2c \operatorname{argmin}_{s \in \mathbb{R}} (W(s) + s^2)$$

where $c = \{F(t)(1 - F(t)f(t)/(2g(t))\}^{1/3}$
W two-sided Brownian motion.

Banerjee and Wellner (2002) found a universal limit law for $-2 \log \mathcal{L}_n(\theta_0)$, where

$$\mathcal{L}_n(\theta) = \frac{\sup\{L(S) : S(t) = \theta\}}{\sup\{L(S)\}}.$$

Limit is an integral involving greatest convex minorants of $W(s) + s^2$.

The estimating function approach also works.

Idea is to apply the generalized ELT to an efficient estimating function.

van der Laan and Robins (1998) found an efficient influence curve for the functional

$$\theta_0 = \int_0^\infty k(u)S(u)\,du$$

(estimable at root-n rate):

$$\begin{aligned} m(X,\theta,F,g,k) &= \frac{k(C)(1-\Delta)}{g(C)} - \theta - \frac{k(C)(F(C)-1)}{g(C)} \\ &\quad + \int_0^\infty k(u)(1-F(u))\,du \end{aligned}$$

Provides an efficient (plug-in) estimating function $m(X,\theta,\widehat{F},\widehat{g},k)$ when $\widehat{F}$ or $\widehat{g}$ is consistent

van der Vaart and van der Laan (2006, IJB) found an asymptotically normal estimator for $\theta_0 = S(t)$:

$$n^{1/3}(\widehat{S}(t) - S(t)) \to_d N(0, \sigma^2),$$

where σ^2 depends on $F(t)$, $g(t)$ and the limits of $\widehat{g}(t)$, $\widehat{F}(t)$.

Estimating function

$$m_n(X, \theta, \widehat{F}, \widehat{g}) = n^{-2/3} m(X, \theta, \hat{F}, \hat{g}, k_n)$$

where $k_n(u) = k((u - t)/b)/b$ is a kernel function of bandwidth $b = b_n = b_1 n^{-1/3}$ centered at t.

Assume

- $\widehat{g}$, $\widehat{F}$ belong to classes of functions having uniform entropy of order $(1/\epsilon)^V$, $V < 2$, w.p. tending to 1
- $\widehat{g}$ or $\widehat{F}$ locally consistent at t.

Note: If $\widehat{g}' \to g'$ uniformly in probability, then $\hat{g}$ belongs to the class of Lipschitz functions, w.p. tending to 1.

Generalized ELT:

$$-2 \log \mathrm{EL}_n(S(t), \widehat{F}, \widehat{g}) \to_d \chi^2_1$$

Approach can be extended to adjust for covariates.

References

M. Hollander, I. W. McKeague and J. Yang. Likelihood Ratio Based Confidence Bands for Survival Functions. *Journal of the American Statistical Association*, **92** 215–226 (1997).

J. Einmahl and I. W. McKeague. Confidence Tubes for Multiple Quantile Plots via Empirical Likelihood. *The Annals of Statistics* **27** 1348–1367 (1999).

I. W. McKeague and Y. Zhao. Simultaneous Confidence Bands for Ratios of Survival Functions via Empirical Likelihood. *Statist. Probab. Letters* **60**, 405–415 (2002).

J. Einmahl and I. W. McKeague. Empirical Likelihood based Hypothesis Testing. *Bernoulli*, **9**, 267–290 (2003).

References

I. W. McKeague and Y. Zhao. Comparing Distribution Functions via Empirical Likelihood. *International Journal of Biostatistics* **1**, Issue 1, Article 5, (2005).

I. W. McKeague and Y. Zhao. Width-Scaled Confidence Bands for Survival Functions. *Statist. Probab. Letters* **76** 327–339 (2006).

A. El Gouch, I. Van Keilegom and I. W. McKeague. Empirical Likelihood Confidence Intervals for Dependent Duration Data. *Econometric Theory* **27** 178–198 (2010).

N. Hjort, I. W. McKeague and I. Van Keilegom. Extending the Scope of Empirical Likelihood. *The Annals of Statistics* **37** 1079–1111 (2009).

H. El Barmi and I. W. McKeague. Empirical Likelihood Based Tests for Stochastic Ordering. *Bernoulli* **19** 295–307 (2013).

H. El Barmi and I. W. McKeague. Testing for uniform stochastic ordering via empirical likelihood. *Ann. Inst. Math. Statist.* 68, 955–976 (2016).

H.-W. Chang and I. W. McKeague. Empirical Likelihood Based Tests for Stochastic Ordering under Right Censorship. *Electronic J. Statist.* 10, 2511–2536 (2016).