Empirical Likelihood

Art B. Owen

Department of Statistics
Stanford University

These lectures

I) Basics of empirical likelihood

II) Estimating equations✓

III) Research frontier

Today: Estimating equations

1) Smooth functions of means
2) Defns for estimating equations
3) Side information and MELEs
4) Regression modeling
5) Time series
6) Finite populations
7) Computation

EL for other than the mean

Some simple statistics are available as smooth functions of a vector mean. Taylor expansion, as in the delta method, then extends empirical likelihood inferences to many such cases.

Much greater generality can be attained via estimating equations. These define a quantity θ implicitly via $\mathbb{E}(m(\boldsymbol{X}, \theta)) = 0$.

Smooth functions of means

$$\sigma = \sqrt{\mathbb{E}(X^2) - \mathbb{E}(X)^2}$$
$$\rho = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{\mathbb{E}(X^2) - \mathbb{E}(X)^2}\sqrt{\mathbb{E}(Y^2) - \mathbb{E}(Y)^2}}$$
$$\theta = h(\mathbb{E}(U, V, \dots, Z))$$

Generally

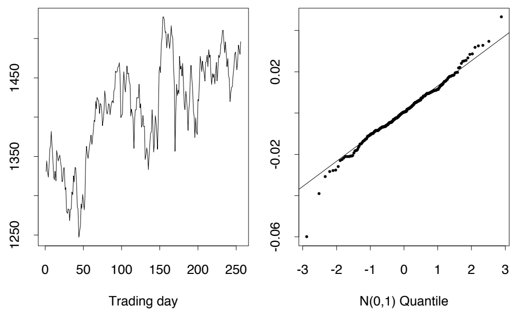
$$\boldsymbol{X} = (U, V, \dots, Z)$$
$$\theta = \mathbb{E}(h(\boldsymbol{X}))$$
$$\hat{\theta} = h(\bar{\boldsymbol{x}}) \doteq h(\mathbb{E}(\boldsymbol{X})) + (\bar{\boldsymbol{x}} - \mathbb{E}(\boldsymbol{X}))^{\mathsf{T}} \frac{\partial}{\partial \boldsymbol{x}} h(\mathbb{E}(\boldsymbol{X}))$$

h nearly linear near $\mathbb{E}(\boldsymbol{X}) \implies \theta$ nearly a mean

EL for smooth functions

$$\mathcal{R}(\theta) = \max\Big\{\prod_{i=1}^n nw_i \mid w_i \ge 0, \sum_{i=1}^n w_i = 1, h\Big(\sum_{i=1}^n w_i \boldsymbol{x}_i\Big) = \theta\Big\}$$

S&P 500 returns

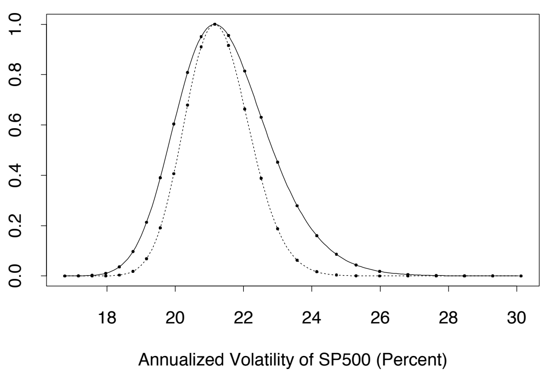


Return $= \log(x_{i+1}/x_i)$

Nearly $\mathcal{N}(0, \sigma^2)$ but heavy tails

Volatility σ is standard deviation of returns

S&P 500 returns



Solid = Empirical likelihood

Dashed = Normal likelihood

Estimating equations

More powerful and general than smooth functions

Define θ via $\mathbb{E}(m(\boldsymbol{X},\theta)) = 0$

Define $\hat\theta$ via $\frac{1}{n}\sum_{i=1}^n m(\boldsymbol{x}_i,\hat\theta) = 0$

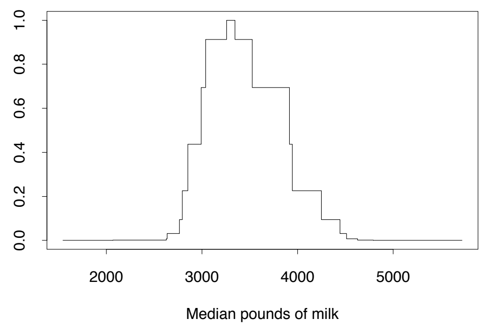
Usually $\dim(m) = \dim(\theta)$

Basic examples:

$m(\boldsymbol{X},\theta)$	Statistic
$\boldsymbol{X} - \theta$	Mean
$1_{\boldsymbol{X}\in A} - \theta$	Probability of set A
$1_{X\le\theta} - \frac{1}{2}$	Median
$\frac{\partial}{\partial\theta}\log(f(\boldsymbol{X};\theta))$	MLE under f

$$-2\log\mathcal{R}(\theta_0) \to \chi^2_{\text{Rank}(\text{Var}(m(\boldsymbol{X},\theta_0)))}$$

Empirical likelihood for a median



LR is constant between observations

$\mathbb{E}(1_{X\le m} - 1/2) = 0$

α-quantile: $\mathbb{E}(1_{X\le\theta} - \alpha) = 0$

Nuisance parameters

Sometimes it is not easy to write $\mathbb{E}(m(\boldsymbol{X},\theta)) = 0$ directly, but it may become much easier by introducing a few extra (nuisance) parameters not of direct interest.

$$\mathbb{E}(m(\boldsymbol{X},\theta,\nu)) = 0$$

where θ is of interest and ν is the nuisance. IE, we expand the parameter vector from θ to (θ,ν).

$$\mathcal{R}(\theta,\nu) = \max\Big\{\prod_{i=1}^n nw_i \mid w_i \ge 0, \sum_{i=1}^n w_i, \sum_{i=1}^n w_i m(\boldsymbol{x}_i,\theta,\nu)\Big\}$$
$$\mathcal{R}(\theta) = \max_\nu \mathcal{R}(\theta,\nu)$$

The first optimization is simple. The second may be difficult.

Typically $-2\log\mathcal{R}(\theta_0) \to \chi^2_{(\dim(\theta))}$

Example: correlation

Suppose we are interested in $\rho = \mathrm{Corr}(X,Y)$. Then,

$$0 = \mathbb{E}(X-\mu_x)$$
$$0 = \mathbb{E}(Y-\mu_y)$$
$$0 = \mathbb{E}((X-\mu_x)^2 - \sigma_x^2)$$
$$0 = \mathbb{E}((Y-\mu_y)^2 - \sigma_y^2)$$
$$0 = \mathbb{E}((X-\mu_x)(Y-\mu_y) - \rho\sigma_x\sigma_y)$$

Parameter and nuisance

$\theta = (\rho)$ and $\nu = (\mu_x,\mu_y,\sigma_x,\sigma_y)$

$\mathbb{E}(m(\boldsymbol{X},\theta,\nu)) = 0 = \frac{1}{n}\sum_{i=1}^n m(X_i,\hat\theta,\hat\nu)$

$m(\cdot)$ has the five components above

Huber's robust M-estimate

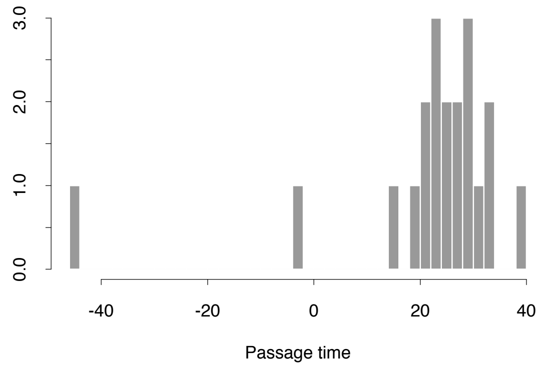
$$0 = \frac{1}{n}\sum_{i=1}^{n}\psi\Big(\frac{x_i-\mu}{\sigma}\Big) \qquad 0 = \frac{1}{n}\sum_{i=1}^{n}\Big[\psi\Big(\frac{x_i-\mu}{\sigma}\Big)^2 - 1\Big]$$

Like mean for small obs, median for outliers

$$\psi(z) = \begin{cases} z, & |z| \le 1.35 \\ 1.35\,\mathsf{sign}(z), & |z| \ge 1.35. \end{cases}$$

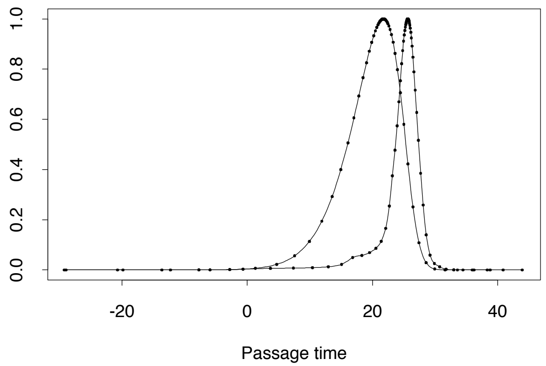
$$\mathcal{R}(\mu) = \max_{\sigma}\max\Big\{\prod_{i=1}^{n} nw_i \mid 0 \le w_i, \sum_i w_i = 1, \sum_i w_i\psi\Big(\frac{x_i-\mu}{\sigma}\Big) = 0,$$
$$\sum_i w_i\Big[\psi\Big(\frac{x_i-\mu}{\sigma}\Big)^2 - 1\Big] = 0\Big\}$$

Newcomb's passage times of light



From Stigler

EL for mean and Huber's location



Curve for the mean is much more skewed by the outlier.

Robust statistic slightly skewed.

Side information

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \in \mathbb{R}^{p+q} \quad \text{known } \mathbb{E}(\boldsymbol{X}) = \mu_{x0}$$

Use what we know

$$\mathcal{R}_{X,Y}(\mu_x,\mu_y) = \max\Big\{\prod_{i=1}^{n} nw_i \mid w_i \ge 0, \sum_i w_i\boldsymbol{x}_i = \mu_x, \sum_i w_i\boldsymbol{y}_i = \mu_y\Big\}$$

$$\mathcal{R}_{X}(\mu_x) = \max\Big\{\prod_{i=1}^{n} nw_i \mid w_i \ge 0, \sum_i w_i\boldsymbol{x}_i = \mu_x\Big\}$$

$$\mathcal{R}_{Y\mid X}(\mu_y \mid \mu_x) = \frac{\mathcal{R}_{X,Y}(\mu_x,\mu_y)}{\mathcal{R}_{X}(\mu_x)}$$

$$-2\log\mathcal{R}_{Y\mid X}(\mu_y \mid \mu_{x0}) \to \chi^2_{(p)}$$

Maximum E. L. estimates

$$\text{Var}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix} = \begin{pmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{pmatrix}$$

MELE $\widetilde{\mu}_y = \sum_{i=1}^n w_i \boldsymbol{y}_i \doteq \bar{\boldsymbol{Y}} - \Sigma_{yx}\Sigma_{xx}^{-1}(\bar{\boldsymbol{X}} - \mu_{x0})$

$$n\,\text{Var}(\widetilde{\mu}_y) \doteq \Sigma_{y|x} \equiv \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

Using known mean reduces variance when $\boldsymbol{Y}$ correlated with $\boldsymbol{X}$

General side information

Can be incorporated via estimating equations

Known parameter	Estimating equation
mean	$\boldsymbol{X} - \mu_x$
α quantile	$1_{X \le Q} - \alpha$
$\Pr(\boldsymbol{X} \in A \mid B)$	$(1_{\boldsymbol{X} \in A} - \rho)1_B$
$\mathbb{E}(\boldsymbol{X} \mid B)$	$(\boldsymbol{X} - \mu)1_B$

Qin has a nice example of Y vs X regression where $E(Y)$ is known

Maximum empirical likelihood estimates

Hartley & Rao	1968	means & finite population setting
O.	1991	means IID sampling
Qin & Lawless	1993	estimating eqns IID

Overdetermined equations

$$\mathbb{E}(m(\boldsymbol{X}, \theta)) = 0, \quad \dim(m) > \dim(\theta)$$

Popular in econometrics, e.g. Generalized Method of Moments Hansen

Approaches:

1) Drop $\dim(m) - \dim(\theta)$ equations
2) Replace $m(\boldsymbol{X}, \theta)$ by $m(\boldsymbol{X}, \theta)A(\theta)$ where
   A a $\dim(m) \times \dim(\theta)$ matrix (IE pick $\dim(\theta)$ linear comb. of m)
3) GMM: estimate the optimal A
4) MELE: $\widetilde{\theta} = \arg\max_\theta \max_{w_i} \prod_i nw_i$ st $\sum_{i=1}^n w_i m(\boldsymbol{x}_i, \theta) = 0$

MELE has same asymptotic variance as using optimal $A(\theta)$

Bias scales more favorably with dimensions for MELE than for $\hat{A}$ methods

Newey, Smith, Kitamura

Qin and Lawless result

$\dim(m) = p + q \geq p = \dim(\theta)$ MELE $\widetilde{\theta}$

$-2\log(\mathcal{R}(\theta_0)/\mathcal{R}(\widetilde{\theta})) \to \chi^2_{(p)}$ conf regions for θ_0

$-2\log \mathcal{R}(\widetilde{\theta}) \to \chi^2_{(q)}$ goodness of fit tests when $q > 0$

Uses only differentiability, moment, identifiability and non-degeneracy conditions, no parametric assumptions.

Regression

$$\mathbb{E}(Y \mid X = x) \doteq \beta_0 + \beta_1 x$$

Models (Freedman)

Correlation $(X_i, Y_i) \sim F_{XY}$ IID

Regression x_i fixed, $Y_i \sim F_{Y \mid X=(1,x_i)}$ indep

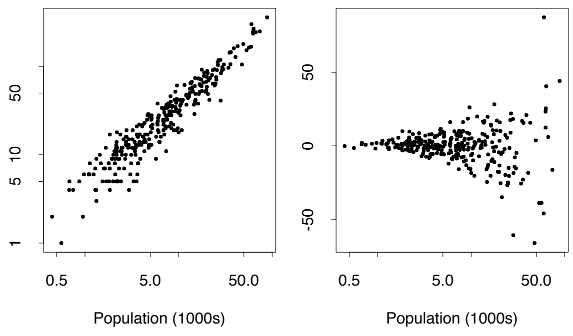
Correlation model

$$\beta = \mathbb{E}(X^{\mathsf{T}}X)^{-1}\mathbb{E}(X^{\mathsf{T}}Y)$$

$$\hat{\beta} = \Bigl(\frac{1}{n}\sum_{i=1}^n X_i^{\mathsf{T}}X_i\Bigr)^{-1}\frac{1}{n}\sum_{i=1}^n X_i^{\mathsf{T}}Y_i$$

β and $\hat{\beta}$ well defined even for lack of fit

Cancer deaths vs population, by county



Nearly linear regression nonconstant residual variance

Royall via Rice

Estimating equations for regression

$$\mathbb{E}(\boldsymbol{X}^{\mathsf{T}}(Y - \boldsymbol{X}^{\mathsf{T}}\beta)) = 0, \qquad \frac{1}{n}\sum_{i=1}^n (Y_i - \boldsymbol{x}_i^{\mathsf{T}}\hat{\beta})\boldsymbol{x}_i = 0$$

$$\mathcal{R}(\beta) = \max\Bigl\{\prod_{i=1}^n nw_i \mid \sum_{i=1}^n w_i \boldsymbol{Z}_i(\beta) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1\Bigr\}$$

$$\boldsymbol{Z}_i(\beta) = (Y_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)\boldsymbol{x}_i$$

need $\mathbb{E}(\|\boldsymbol{Z}\|^2) \leq \mathbb{E}\Bigl(\|\boldsymbol{X}\|^2(Y - \boldsymbol{X}^{\mathsf{T}}\beta)^2\Bigr) < \infty$

Don't need:

normality, constant variance, exact linearity

For cancer data

$P_i =$ population of i'th county in 1000s

$C_i =$ cancer deaths of i'th county in 20 years

$C_i \doteq \beta_0 + \beta_1 P_i$

$\hat{\beta}_1 = 3.58 \quad \Longrightarrow \quad 3.58/20 = 0.18$ deaths per thousand per year

$\hat{\beta}_0 = -0.53$ near zero, as we'd expect

Regression through the origin

$$C_i \doteq \beta_1 P_i$$

Residuals should have mean zero and be orthogonal to P_i

We want two equations in one unknown β_1

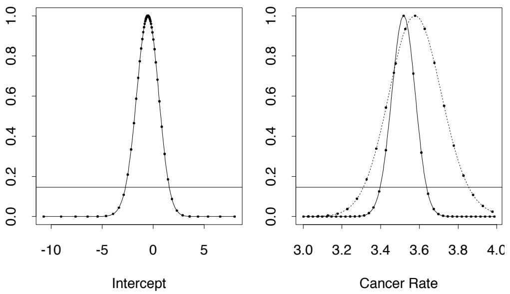
Equivalently, side information $\beta_0 = 0$

Least squares regression through origin does not solve both equations

MELE $\widetilde{\beta}_1 = \arg\max_{\beta_1} \mathcal{R}(\beta_1)$

$$\mathcal{R}(\beta_1) = \max\Big\{\prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i (C_i - P_i\beta_1) = 0, \sum_{i=1}^n w_i P_i (C_i - P_i\beta_1) = 0, \sum_{i=1}^n w_i = 1, w_i \ge 0\Big\}$$

Regression parameters



Intercept nearly 0, MELE smaller than MLE

CI based on conditional empirical likelihood

Constraint narrows CI for slope by over half

Fixed predictor regression model

$\mathbb{E}(Y_i) = \mu_i \doteq \beta_0 + \beta_1 x_i$ fixed, and $\mathrm{Var}(Y_i) = \sigma_i^2$

With lack of fit $\mu_i \ne \beta_0 + \beta_1 x_i$

No good definition of 'true' β given L.O.F.

$\boldsymbol{Z}_i = (Y_i - \boldsymbol{x}_i^\mathsf{T}\beta)\boldsymbol{x}_i$ have

1) $\mathbb{E}(\boldsymbol{Z}_i) = (\mu_i - \boldsymbol{x}_i^\mathsf{T}\beta)\boldsymbol{x}_i$ $\quad 0$ may be the common value

2) $\mathrm{Var}(\boldsymbol{Z}_i) = \boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}\sigma_i^2$ $\quad$ non-constant, even if σ_i^2 constant

Triangular array ELT

$$
\begin{array}{llllll}
Z_{11} & & & & & \\
Z_{12} & Z_{22} & & & & \\
Z_{13} & Z_{23} & Z_{33} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
Z_{1n} & Z_{2n} & Z_{3n} & \cdots & Z_{nn} & \\
\vdots & \vdots & \vdots & & & \ddots
\end{array}
$$

Row n has indep $Z_{1n}, \dots, Z_{nn}$, common mean 0 not ident distributed

Different rows have different distns

Still get $-\log \mathcal{R}(\text{Common mean} = 0) \to \chi^2_{\dim(Z)}$ under mild conditions

Applies for fixed $\boldsymbol{x}$ regression: $Z_{in} = (Y_i - \boldsymbol{x}_i^\mathsf{T}\beta)\boldsymbol{x}_i$

Variance modelling

Working model $Y \sim \mathcal{N}(\boldsymbol{x}^\mathsf{T}\beta, e^{2\boldsymbol{z}^\mathsf{T}\gamma})$

$$0 = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\mathsf{T}\beta)\, e^{-2\boldsymbol{z}_i^\mathsf{T}\gamma} \qquad (\text{weight} \propto 1/\text{var})$$

$$0 = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\Big(1 - \exp(-2\boldsymbol{z}_i^\mathsf{T}\gamma)(y_i - \boldsymbol{x}_i^\mathsf{T}\beta)^2\Big)$$

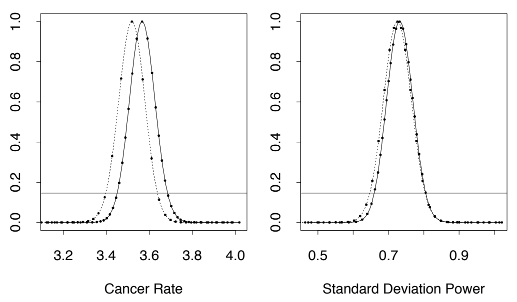
For cancer data

$$\boldsymbol{x}_i = (1, P_i)^\mathsf{T} \quad \boldsymbol{z}_i = (1, \log(P_i))^\mathsf{T}$$

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 P_i \quad \sqrt{\text{Var}(Y_i)} = \exp(\gamma_0 + \gamma_1 \log(P_i)) = e^{\gamma_0} P_i^{\gamma_1}$$

and $\beta_0 = 0$

Heteroscedastic model

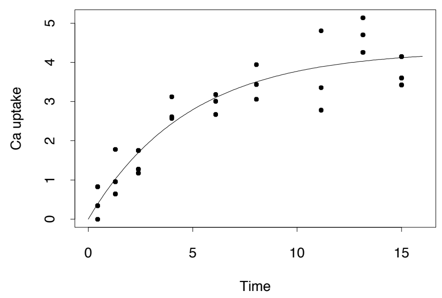


Left: solid curve accounts for nonconstant variance

Right: solid curve forces $\beta_0 = 0$, and,

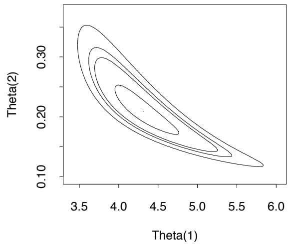
rules out $\gamma_1 = 1/2$ (Poisson) and $\gamma_1 = 1$ (Gamma)

Nonlinear regression



$$y \doteq f(x, \theta) \equiv \theta_1(1 - \exp(-\theta_2 x))$$

Nonlinear regression regions



$$0 = \sum_{i=1}^{n} w_i (Y_i - f(x_i, \theta)) \frac{\partial}{\partial \theta} f(x_i, \theta)$$

Don't need: normality or constant variance

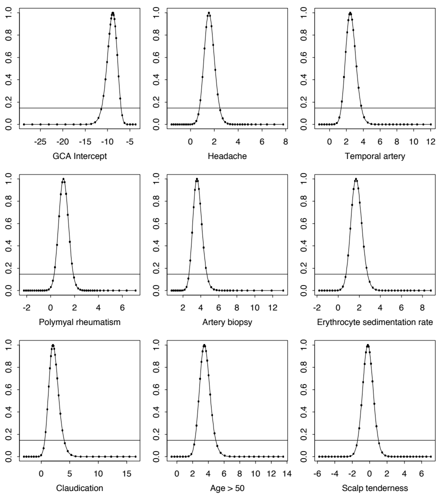
Logistic regression

- Giant cell arteritis is a type of vasculitis (inflamation of blood or lymph vessels)
- Not all vasculitis is GCA
- Try to predict GCA from 8 binary predictors

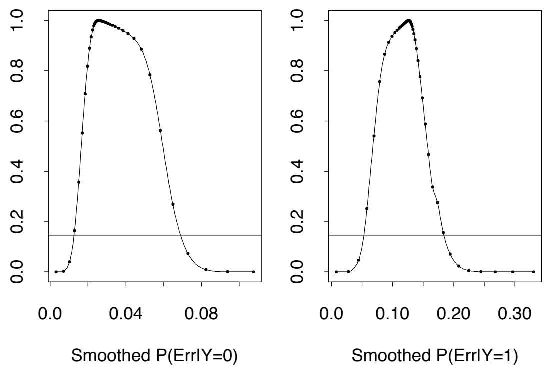
$$\Pr(\text{GCA}) \doteq \tau(X^\mathsf{T}\beta) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_8 X_8)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_8 X_8)}$$

Likelihood estimating equations reduce to: $\boldsymbol{Z}_i(\beta) = \boldsymbol{X}_i(Y_i - \tau(\boldsymbol{X}_i^\mathsf{T}\beta))$

Logistic regression coefficients



Prediction accuracy



Multiple biased samples

Population k sampled from F with bias $u_k(\cdot)$, $k = 1, \dots, s$

$$\boldsymbol{X}_{ik} \sim G_k, \qquad i = 1, \dots, n_k, \quad k = 1, \dots, s$$

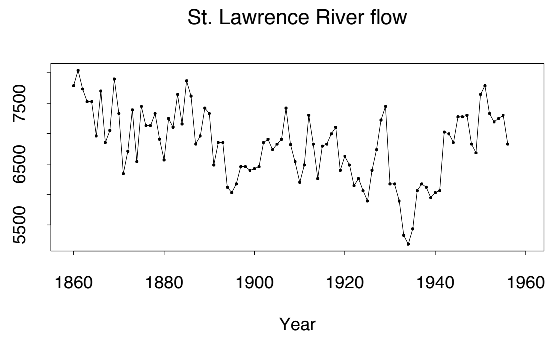
$$G_k(A) = \frac{\int_A u_k(\boldsymbol{y})\, dF(\boldsymbol{y})}{\int u_k(\boldsymbol{y})\, dF(\boldsymbol{y})}, \quad k = 1, \dots, s$$

Examples

1) clinical trials with varying enrolment criteria
2) mix of length biased and unbiased samples
3) telescopes with varying detection limits
4) sampling from different frames

NPMLEs Vardi (also Wellner) and χ^2 limits Qin by multiplying likelihoods

Time series



Reduce to independence

$$Y_i - \mu = \beta_1(Y_{i-1} - \mu) + \cdots + \beta_k(Y_{i-k} - \mu) + \epsilon_i$$
$$\mathbb{E}(\epsilon_i) = 0$$
$$\mathbb{E}(\epsilon_i^2) = \exp(2\tau)$$
$$\mathbb{E}(\epsilon_i(Y_{i-j} - \mu)) = 0$$

j	$\hat{\beta}_j$	$-2\log \mathcal{R}(\beta_j = 0)$
1	0.627	30.16
2	-0.093	0.48
3	0.214	4.05

Blocking of time series

Block i of observations, out of $n = \lfloor (T - M)/L + 1 \rfloor$ blocks

$$B_i = \big(Y_{(i-1)L+1}, \dots, Y_{(i-1)L+M}\big)$$

$M =$ length of blocks

$L =$ spacing of start points

Large $M = L \implies$ block dependence small

Large M $\implies$ block dependence predictable given L

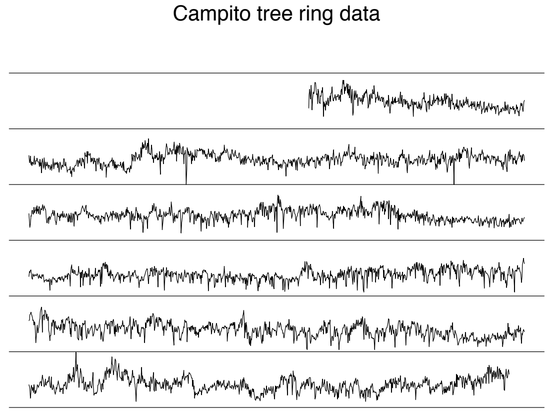
Blocked estimating equation, replace m by b

$$b(B_i, \theta) = \frac{1}{M} \sum_{j=1}^{M} m(X_{(i-1)L+j}, \theta)$$

$$-2\Big(\frac{T}{nM}\Big) \log \mathcal{R}(\theta_0) \to \chi^2 \qquad \text{as } M \to \infty, MT^{-1/2} \to 0 \quad \text{Kitamura}$$

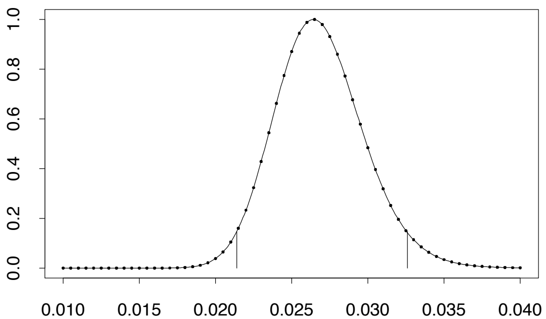
Bristlecone pine

5405 years of Bristlecone pine tree ring widths



0 to 100 in 0.01 mm   Fritts et al.

Probability of sharp decrease



Sharp $\equiv$ drop of over 0.2 mm from average of previous 10 years.

MELEs for finite population sampling

1) use side information
   (a) population means, totals, sizes
   (b) stratum means, totals, sizes
2) take unequal sampling probabilities
3) use non-negative observation weights

Hartley & Rao, Chen & Qin, Chen & Sitter

More finite population results

χ^2 limits	$-2\left(1-\frac{n}{N}\right)\mathcal{R}(\mu) \to \chi^2$	Zhong & Rao
EL variance ests	via pairwise inclusion probabilities	Sitter & Wu
Multiple samples	varying distortions	Zhong, Chen, & Rao

Curve estimation problems

$$\widehat{f}_h(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\Big(\frac{x_i - x}{h}\Big) \qquad \text{density}$$

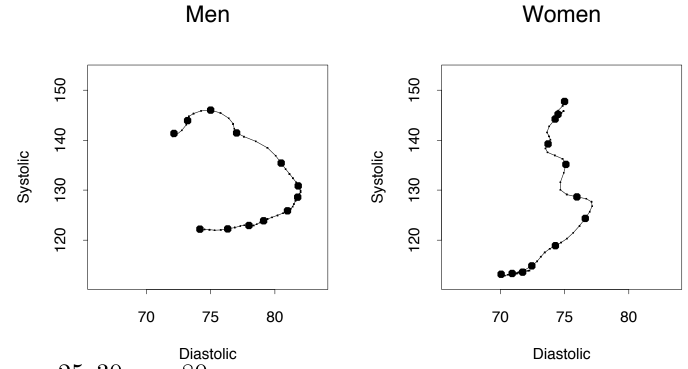
$$\widehat{\mu}_h(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\Big(\frac{x_i - x}{h}\Big)Y_i \qquad \text{regression}$$

Triangular array ELT applies   Bias adjustment issues

Dimensions and geometry

Dim(x)	Dim(y)	Estimate	Region
1	≥ 2	space curve	confidence tube
≥ 2	1	(hyper)-surface	confidence sandwich

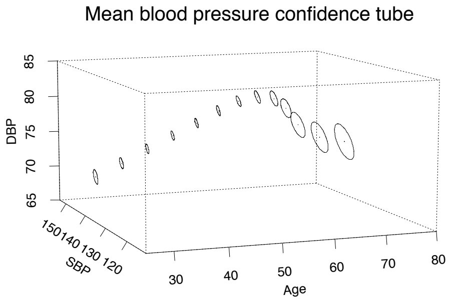
Trajectories of mean blood pressure



dots at ages $25, 30, \ldots, 80$

data from Jackson et al., courtesy of Yee

Confidence tube for men's mean SBP, DBP



Empirical likelihood vs bootstrap

1) EL gives shape of regions for $d > 1$
2) EL Bartlett correctable, bootstrap not
3) EL can be faster, but,
4) EL optimization can be hard

Computation

$$\log \mathcal{R}(\theta) = \max_{\nu} \log \mathcal{R}(\theta, \nu)$$
$$= \max_{\nu} \min_{\lambda} \mathbb{L}(\theta, \nu, \lambda), \quad \text{where,}$$
$$\mathbb{L}(\theta, \nu, \lambda) = -\sum_{i=1}^{n} \log\bigl(1 + \lambda^{\mathsf{T}} m(x_i, \theta, \nu)\bigr)$$

Inner and outer optimizations $\ll n$ dimensional

Used NPSOL, expensive and not public domain (but it works)

Algorithmic strategies

Newton's method to solve for a saddlepoint:

$$0 = \frac{\partial}{\partial \nu} \mathbb{L}(\theta, \nu, \lambda)$$
$$0 = \frac{\partial}{\partial \lambda} \mathbb{L}(\theta, \nu, \lambda)$$

Progress towards a saddle-point is more difficult to define than progress towards a mode.

Newton's method to solve

$$\max_{\nu} \mathcal{R}(\theta, \nu)$$

deriving gradient and Hessian from $\mathbb{L}(\theta, \nu, \lambda)$

These methods usually work well around the MLE.

As $n \to \infty$ the region where they work grows.

Next: research directions

Two main challenges for empirical likelihood are

1) escaping the convex hull

2) profiling out nuisance parameters

Problem 1 is important when the parameter is high dimensional. Less important when we only want a confidence statement on on or two of the components.

Problem 2 is also difficult for parametric likelihoods; usually we just make a second order Taylor approximation to the log likelihood around the MLE.

There has been great progress on problem 1.