The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

**1863-16**

**Advanced School and Conference on Statistics and Applied Probability in Life Sciences**

***24 September - 12 October, 2007***

# Estimating optimal step-function approximations in semiparametric models

Ian McKeague

*Mailman School of Public Health*
*Columbia University*
*New York NY 10032, USA*

# Estimating optimal step-function approximations in semiparametric models

Ian McKeague
Columbia University

October 1, 2007

Advanced School and Conference on Statistics and Applied Probability in Life Sciences, Trieste

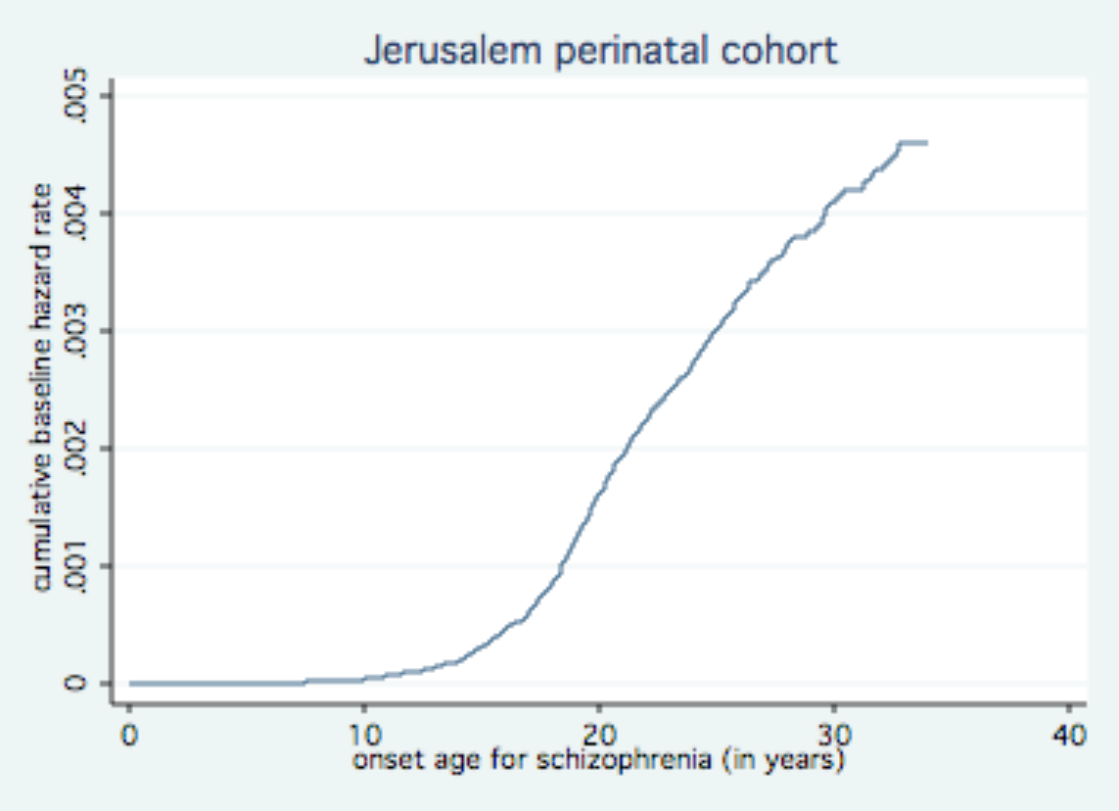
# Outline

1. Motivation: condensing functional information
2. Binary decision trees in
   - nonparametric regression
   - Cox regression
   - partially linear models
3. Confidence intervals for split points (thresholds)
4. Example: phosphorus threshold for the Florida Everglades
5. Example: age threshold for onset of schizophrenia
6. Conclusion

# Motivation

- Key information about functional parameters (nonparametric regression functions, hazard functions, . . . ) is not easily reported *non-graphically*.
- We are interested in condensing functional information into a form that can be communicated easily (using a few parameters).
- We investigate how this can be done in terms of the best-fitting binary decision tree approximation, defined by the location of an abrupt change and mean levels on either side.
- Binary decision trees are "weak learners" in terms of prediction (unless improved via bagging say), but are readily interpretable.

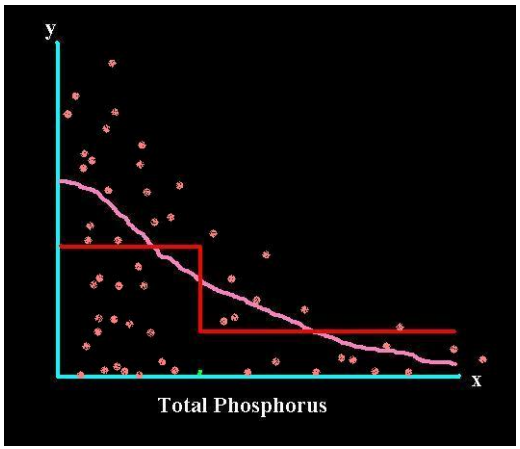
# Example: cumulative baseline hazard




What is the best way to interpret this plot?

# Binary decision trees in nonparametric regression (CART)

$Y = f(X) + \epsilon$
$E(\epsilon|X) = 0$
$f$ unknown
$X$ has unknown density $p_X(\cdot)$
$\epsilon$ has unknown conditional variance $\sigma^2(x)$
$n$ i.i.d. observations of $(X, Y)$




$$(\beta_l^0, \beta_u^0, d^0) = \text{argmin}_{\beta_l, \beta_u, d} \, E\,[Y - \beta_l \, 1(X < d) - \beta_u \, 1(X \geq d)]^2$$

# Least squares estimates

$$(\hat{\beta}_l, \hat{\beta}_u, \hat{d}_n) = \text{argmin}_{\beta_l, \beta_u, d} \sum_{i=1}^{n} [Y_i - \beta_l 1(X_i < d) - \beta_u 1(X_i \geq d)]^2$$

Normal equations:

$$\beta_l^0 = E(Y \mid X < d^0), \quad \beta_u^0 = E(Y \mid X \geq d^0), \quad f(d^0) = \frac{\beta_l^0 + \beta_u^0}{2}$$

# Bühlmann and Yu (2002)

$$n^{1/3}(\hat{d}_n - d^0) \rightarrow_d \text{argmax}_t\, Q(t)$$

where

$$Q(t) = a\, W(t) - b_0\, t^2, \qquad t \in \mathbb{R}$$

$W(t)$ is two-sided Brownian motion started from 0,

$$a^2 = p_X(d^0)\sigma^2(d^0), \qquad b_0 = \frac{1}{2} \mid p_X(d^0) f'(d^0) \mid > 0\,.$$

Error in proof: assumes that $\hat{\beta}_l$ and $\hat{\beta}_u$ converge at $n^{1/2}$-rate and make no contribution to the limiting distribution of $\hat{d}_n$.

# Conditions

(A1) There is a unique minimizer $(\beta_l^0, \beta_u^0, d^0)$ of the expectation in the least squares criterion with $\beta_l^0 \neq \beta_u^0$.

(A2) $f(x)$ is continuous and is continuously differentiable in an open neighborhood $N$ of $d^0$. Also, $f'(d^0) \neq 0$.

(A3) $p_X(x)$ does not vanish and is continuously differentiable on $N$.

(A4) $\sigma^2(x)$ is continuous on $N$.

(A5) $\sup_{x \in N} E[\epsilon^2 1\{|\epsilon| > \eta\} | X = x] \to 0$ as $\eta \to \infty$.

# Theorem (Banerjee and McKeague, 2006)

$$n^{1/3}\left(\hat{\beta}_l - \beta_l^0, \hat{\beta}_u - \beta_u^0, \hat{d}_n - d^0\right) \to_d (c_1, c_2, 1)\,\mathrm{argmax}_t\, Q(t),$$

where

$$Q(t) = a\,W(t) - b\,t^2,$$

$$b = b_0 - \frac{1}{8}|\beta_l^0 - \beta_u^0|\,p_X(d^0)^2\left(\frac{1}{F_X(d^0)} + \frac{1}{1 - F_X(d^0)}\right) > 0,$$

$$b_0 = p_X(d^0)|f'(d^0)|/2,$$

$$c_1 = \frac{p_X(d^0)(\beta_u^0 - \beta_l^0)}{2F_X(d^0)}, \quad c_2 = \frac{p_X(d^0)(\beta_u^0 - \beta_l^0)}{2(1 - F_X(d^0))}.$$

# Example: binary response

$Y|X \sim \mathrm{Ber}(f(X))$, where $f(x) = P(Y = 1|X = x)$.

The ratio of $\beta_u^0$ to $\beta_l^0$ is a relative risk:

$$\beta_u^0/\beta_l^0 = P(Y = 1|X > d^0)/P(Y = 1|X \leq d^0)$$

useful for comparing the risks before and after the split point.

# Binary decision trees in Cox regression

Conditional hazard function for the failure time $T$ of an individual with a $p$-vector of covariates $Z$

$$\lambda(t|Z) = \lambda(t)\exp\{\beta^T Z\},$$

$\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p$-vector of unknown regression coefficients and $\lambda(t)$ is an unspecified baseline hazard function.

$X = \min\{T, C\}$, where $T$ and the censoring time $C$ are assumed to be conditionally independent given $Z$.

$\delta = 1\{T \leq C\}$, indicator that failure is observed.

# Binary tree approximation to $\lambda(t)$

$$\bar{\lambda}(t; \lambda_l, \lambda_u, d) = \lambda_l \, 1(t \leq d) + \lambda_u \, 1(t > d)$$

where $d$ is the threshold (or jump point), $\lambda_l$ is the value to the left of the jump, and $\lambda_u$ is the value to the right of the jump.

$$(\lambda_l^0, \lambda_u^0, d^0) = \text{argmin}_{\lambda_l, \lambda_u, d} \int_0^\tau \left[\lambda(t) - \bar{\lambda}(t; \lambda_l, \lambda_u, d)\right]^2 dt,$$

where $\tau > 0$ is a given terminal time.

- The threshold $d^0$ is the main parameter of interest; it most accurately splits the time interval into two subintervals with the risk changing abruptly at the boundary.

Express the $L^2$-distance in terms of the cumulative baseline hazard function $\Lambda(t) = \int_0^t \lambda(u)\, du$:

$$(\lambda_l^0, \lambda_u^0, d^0) \quad = \quad \text{argmin}_{\lambda_l, \lambda_u, d}\, \mathbb{M}(\lambda_l, \lambda_u, d),$$

where $\mathbb{M}$ is the criterion function

$$\mathbb{M}(\lambda_l, \lambda_u, d) \equiv (\lambda_l^2 - \lambda_u^2)\, d + \lambda_u^2\, \tau + 2\,(\lambda_u - \lambda_l)\, \Lambda(d) - 2\,\lambda_u\, \Lambda(\tau)\,.$$

**Normal equations:**

$$\lambda_l^0 = \frac{\Lambda(d^0)}{d^0}, \qquad \lambda_u^0 = \frac{\Lambda(\tau) - \Lambda(d^0)}{\tau - d^0}, \qquad \lambda(d^0) = \frac{\lambda_l^0 + \lambda_u^0}{2}\,,$$

# Estimators

$n$ i.i.d. observations $(X_i, \delta_i, Z_i)$ of $(X, \delta, Z)$.

$$(\hat{\lambda}_l^0, \hat{\lambda}_u^0, \hat{d}_n) \quad = \quad \text{argmin}_{\lambda_l, \lambda_u, d} \, \mathbb{M}_n(\lambda_l, \lambda_u, d),$$

where

$$\mathbb{M}_n(\lambda_l, \lambda_u, d) \equiv (\lambda_l^2 - \lambda_u^2)\, d + \lambda_u^2\, \tau + 2\,(\lambda_u - \lambda_l)\, \hat{\Lambda}_n(d) - 2\,\lambda_u\, \hat{\Lambda}_n(\tau)$$

and $\hat{\Lambda}_n$ is Breslow's estimator

$$\hat{\Lambda}_n(t) = \mathbb{P}_n \left[ \frac{\delta 1\{X \leq t\}}{S^{(0)}(\hat{\beta}, X)} \right]$$

$\mathbb{P}_n$ = empirical distribution

$S^{(0)}(\beta, t) = \mathbb{P}_n[Y(t) e^{\beta^T Z}]$
$Y(t) = 1\{X \geq t\}$ at-risk indicator.

# Conditions

Usual regularity conditions for Cox model (Andersen and Gill, 1982). In particular, assume

$$s^{(0)}(\beta, t) = E[Y(t)e^{\beta^T Z}]$$

is bounded away from zero. Also assume bounded covariates.

(A1) There is a unique vector $(\lambda_l^0, \lambda_u^0, d^0)$ with $\lambda_l^0 \neq \lambda_u^0$ and $0 < d^0 < \tau$ that minimizes $\mathbb{M}$.

(A2) $\lambda$ is continuously differentiable in a neighborhood of $d^0$, and $\lambda'(d^0) \neq 0$.

# Theorem

$$n^{1/3}\left(\hat{\lambda}_l - \lambda_l^0, \hat{\lambda}_u - \lambda_u^0, \hat{d}_n - d^0\right) \to_d (c_1, c_2, 1)\,\mathrm{argmax}_t\, Q(t),$$

where

$$Q(t) = aW(t) - bt^2,$$

$W$ is two-sided Brownian motion, $a^2 = \lambda(d^0)/s^{(0)}(\beta_0, d^0)$,

$$b = b_0 - \frac{1}{8}|\lambda_l^0 - \lambda_u^0|\left(\frac{1}{d^0} + \frac{1}{\tau - d^0}\right) > 0,$$

$b_0 = |\lambda'(d^0)|/2$, and

$$c_1 = \frac{\lambda_u^0 - \lambda_l^0}{2d^0}, \quad c_2 = \frac{\lambda_u^0 - \lambda_l^0}{2(\tau - d^0)}.$$

# Proof

Uses a strategy applicable to general M-estimators $\hat{\theta} = \text{argmin}_\theta \mathbb{M}_n(\theta)$ in which we establish 1) the rate of convergence, 2) the weak convergence of a suitably localized version of the criterion function, and 3) apply the argmax (or argmin) continuous mapping theorem.

The rate of convergence is derived in terms of the expected continuity modulus of $\sqrt{n}(\mathbb{M}_n - \mathbb{M})$ at $\theta_0$:

$$\sqrt{n}E\left[\sup_{d(\theta,\theta_0)<\epsilon} |(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)|\right] = O(\sqrt{\epsilon})$$

for $\epsilon > 0$. This implies a $n^{1/3}$-rate of convergence.

# Binary decision trees for partially linear models

$Y = X^T\beta + g(Z) + \epsilon$
$\beta$ a $p$-vector of regression parameters
$\epsilon$ independent of $(X, Z)$

Binary tree approximation for $g$:

$$(\lambda_l^0, \lambda_u^0, d^0) = \text{argmin}_{\lambda_l,\lambda_u,d}\, E\left[g(Z) - \bar{g}(Z; \lambda_l, \lambda_u, d)\right]^2,$$

where

$$\bar{g}(z; \lambda_l, \lambda_u, d) = \lambda_l\, 1(z \leq d) + \lambda_u\, 1(z > d).$$

# Adapting our strategy from Cox regression

$$(\lambda_l^0, \lambda_u^0, d^0) = \text{argmin}_{\lambda_l, \lambda_u, d} \, \mathbb{M}(\lambda_l, \lambda_u, d),$$

where

$$\mathbb{M}(\lambda_l, \lambda_u, d) \equiv (\lambda_l^2 - \lambda_u^2) F(d) + \lambda_u^2 + 2\,(\lambda_u - \lambda_l)\,\Lambda(d) - 2\,\lambda_u\,\Lambda(1),$$

$F =$ cdf of $Z$, and $\Lambda(t) = E[g(Z)1\{Z \leq t\}]$.

Plug-in estimates:

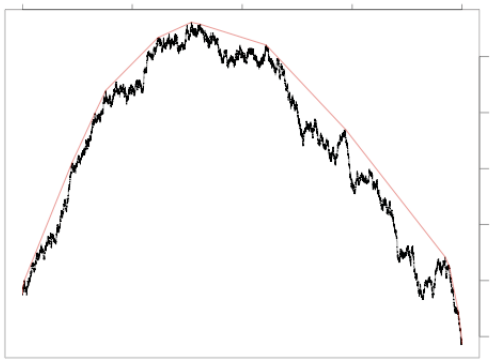
$$\hat{\Lambda}_n(t) = \mathbb{P}_n \left[ (Y - X^T \hat{\beta}) 1\{Z \leq t\} \right], \quad \hat{F}_n(t) = \mathbb{P}_n 1\{Z \leq t\}$$

where $\hat{\beta}$ is a $\sqrt{n}$-consistent estimator of $\beta$.

# Brownian scaling

$Q_{a,b}(t) = a\,W(t) - b\,t^2$ is a scaled, time-changed version of $Q_{1,1}$:

$$Q_{a,b}(t) \stackrel{\mathcal{D}}{=} a\,(a/b)^{1/3}\,Q_{1,1}((b/a)^{2/3}\,t)$$



# Chernoff's distribution

$$Z = \text{argmax}_t Q_{1,1}(t).$$

has a density that can be expressed analytically in terms of zeros of the Airy function (Groeneboom, 1985); quantiles calculated exactly via an algorithm of Groeneboom and Wellner (2001). 0.975 quantile is 0.998181.

# Confidence intervals for the split point

From Brownian scaling,

$$n^{1/3}(\hat{d}_n - d^0) \to_d kZ,$$

where $k = (a/b)^{2/3}$, so we have the Wald-type CI

$$\hat{d}_n \pm n^{-1/3}\hat{k}p_{\alpha/2},$$

where $p_\alpha$ is the upper $\alpha$-quantile of Chernoff's distribution.

**Problem:** smoothing needed to estimate $k$.

# Bootstrap of M-estimators under cube-root asymptotics

- subsampling (resampling without replacement), Politis and Romano (1994)
- Delgado, Rodriquez-Poo and Wolf (2001) gave simulation evidence for inconsistency of the empirical bootstrap
- Abrevaya and Huang (2005) proposed a corrected empirical bootstrap
- $m$ out of $n$ bootstrap, Lee and Pun (2006)

# Abrevaya and Huang (2005, Econometrica)

*Econometrica*, Vol. 73, No. 4 (July, 2005), 1175–1204

## ON THE BOOTSTRAP OF THE MAXIMUM SCORE ESTIMATOR

BY JASON ABREVAYA AND JIAN HUANG[1]

This paper shows that the bootstrap does not consistently estimate the asymptotic distribution of the maximum score estimator. The theory developed also applies to other estimators within a cube-root convergence class. For some single-parameter estimators in this class, the results suggest a simple method for inference based upon the bootstrap.

KEYWORDS: Maximum score estimation, bootstrap, cube-root asymptotics.

### 1. INTRODUCTION

THE MAXIMUM SCORE ESTIMATOR of Manski (1975) was the first semiparametric estimator proposed for the (latent-variable) binary response model. The model considered by Manski (1975) replaced parametric assumptions on the error disturbance with a conditional median restriction,

$$y = \{x'\beta_0 - \epsilon \geq 0\}, \quad \text{Median}(\epsilon|x) = 0, \tag{1}$$

where $\{\cdot\}$ denotes an indicator function. If $(y_1, x_1), \ldots, (y_n, x_n)$ are the observed data that satisfy (1), the maximum score estimator $\beta_n$ is defined as the maximizer of the objective function[2]

$$n^{-1}\sum_{i=1}^{n}(2y_i - 1)\{x_i'\beta \geq 0\} \tag{2}$$

# Correcting the bootstrap

Abrevaya and Huang claim to have proved that, for an M-estimator converging at cube-root rate, conditionally on the data

$$n^{1/3}(\hat{d}_n^* - \hat{d}_n) \rightarrow_d kZ^*$$

almost surely, where $k$ is the constant to be estimated,

$$Z^* = \text{argmax}_t(W(t) - t^2) - \text{argmax}_t(W(t) + W^*(t) - t^2)$$

and $W$ and $W^*$ are independent two-sided Brownian motions.

Unfortunately, their proof has a serious error!

# Comparison of densities of $Z$ and $Z^*$

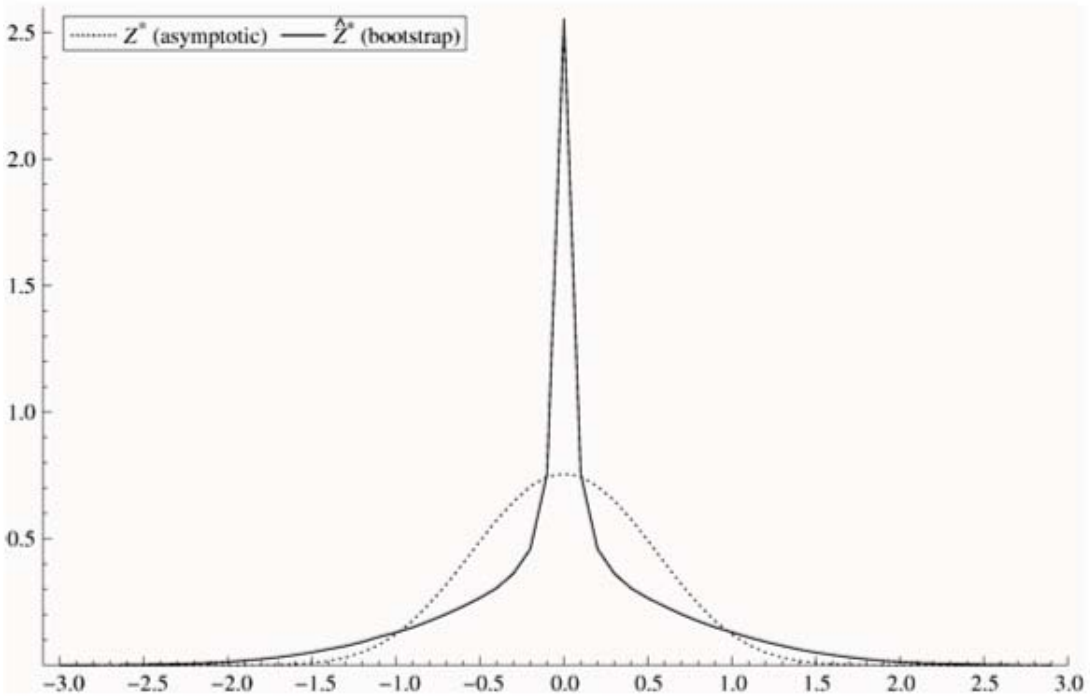



FIGURE 1.—Densities of $Z^*$ and $\hat{Z}^*$.

# Lee and Pun (2006, JASA)

Show that $m$ out of $n$ bootstrap works for general M-estimators.

Disadvantage: calibration (choice of $m$) still needed, as with subsampling.



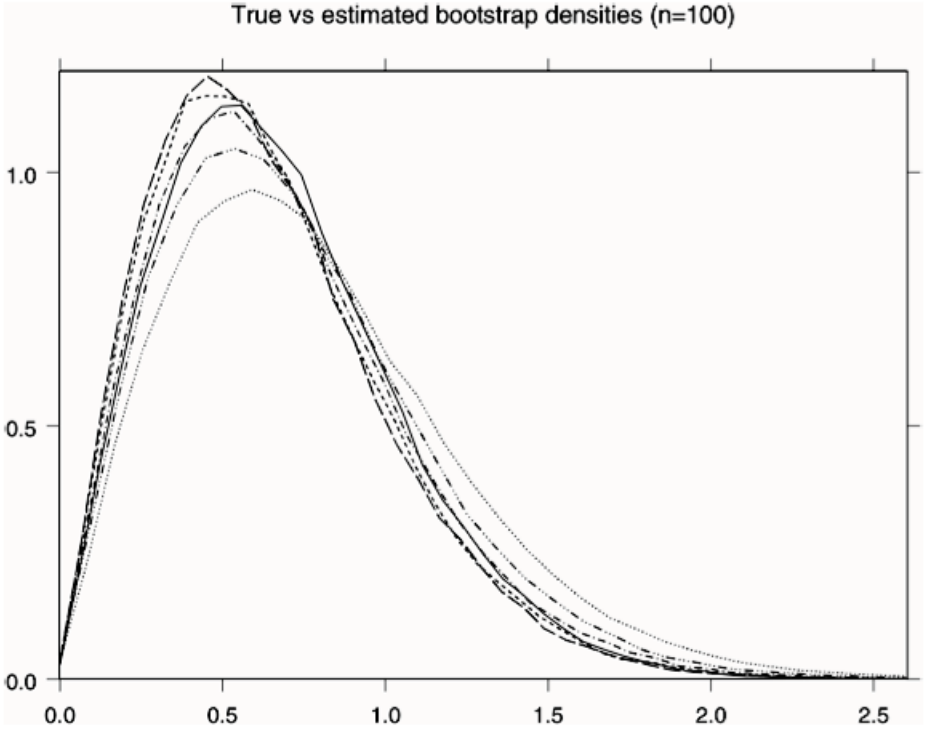



*Figure 3. Case 2, Studentized Location M-Estimation: Kernel-Smoothed Mean Density Functions of $m^{1/3}|\hat{\xi}_n - \xi_n|$ Compared With Density Function of $n^{1/3}|\xi_n - \xi_0|$ (———), for $m = 30$ (·····), 50 (–·–·–), 70 (–··–··–), 90 (– – –), 100 (— — —) and $n = 100$.*

# Confidence sets based on deviance

Use a deviance function as an asymptotic pivot:

$$\mathbb{D}_n(d) = \mathbb{M}_n(\hat{\lambda}_l^d, \hat{\lambda}_u^d, d) - \mathbb{M}_n(\hat{\lambda}_l, \hat{\lambda}_u, \hat{d}_n),$$

where

$$\hat{\lambda}_l^d = \frac{\hat{\Lambda}_n(d)}{d}, \qquad \hat{\lambda}_u^d = \frac{\hat{\Lambda}_n(\tau) - \hat{\Lambda}_n(d)}{\tau - d}.$$

It can be shown that

$$n^{2/3}\mathbb{D}_n(d^0) \to_d k \max_t(W(t) - t^2),$$

where $k$ is a constant that can be estimated.

Invert $\mathbb{D}_n(d)$ to get a CI for $d^0$.

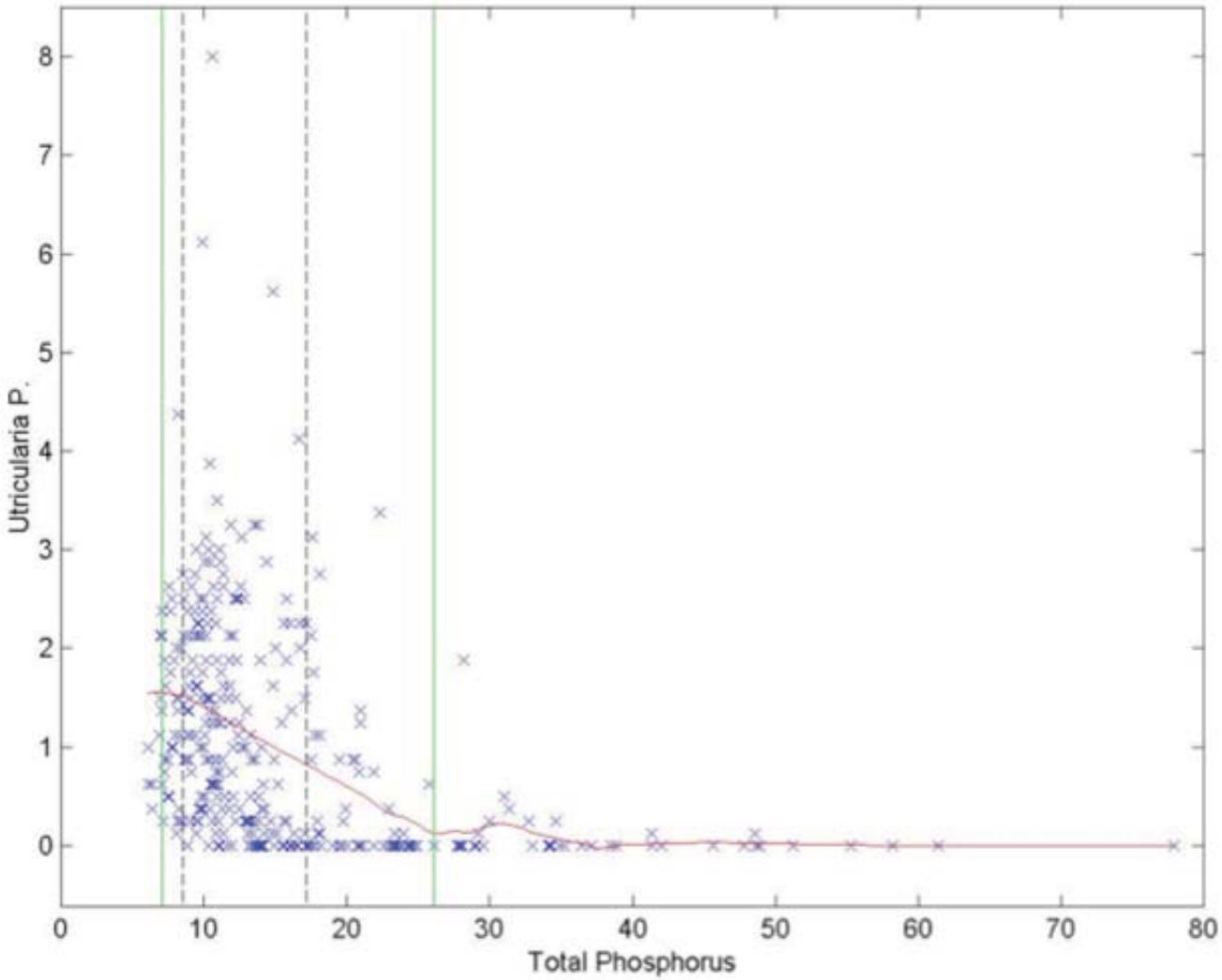
# Example: DUWC Dosing Study

Experiment to find threshold level of total phosphorus at which biological imbalance occurs in the Everglades; data comes from 1992–1998 at two unimpacted sites. $n = 340$.

In 1994, the Florida legislature passed the Everglades Forever Act which called for a threshold level of total phosphorus that would prevent an "imbalance in natural populations of aquatic flora or fauna."

This threshold may eventually be set at around 10 or 15 parts per billion (ppb), but it remains undecided despite extensive scientific study and much political and legal debate.

# DUWC Dosing Study (cont'd)



# 95% CIs for the phosphorus threshold

Bayesian changepoint: 13.7–15.4 ppb

Wald-type: 0.7–24.9 ppb

Subsampling: 8.5–17.1 ppb

Deviance-type: 7.1–26.1 ppb

# Simulation example (hazard function setting)

Baseline hazard: $\lambda(t) = t$, terminal time $\tau = 1.5$.

Note: there is no abrupt change in $\lambda(t)$! Yet the threshold is well-defined: $d^0 = .75$.

Covariates: $Z \sim \text{Unif}[0,1]^p$, for $p = 1$ and 5

$\beta_0 = 1/p$

censoring time $C$ exponential with mean 3

1000 replicated samples

Table: Coverage and average CI length, $p = 1$

| | Wald | | Deviance type | |
|---|---|---|---|---|
| $n$ | Coverage | Length | Coverage | Length |
| 50 | 98.4 | 1.77 | 93.6 | 1.10 |
| 100 | 99.0 | 1.42 | 93.1 | 1.02 |
| 150 | 97.6 | 1.16 | 93.9 | 0.90 |
| 200 | 98.0 | 1.05 | 95.1 | 0.85 |
| 250 | 98.0 | 0.95 | 94.6 | 0.79 |
| 300 | 97.5 | 0.89 | 94.5 | 0.74 |
| 350 | 95.6 | 0.81 | 94.5 | 0.69 |
| 400 | 94.8 | 0.76 | 96.4 | 0.66 |
| 450 | 94.0 | 0.73 | 94.4 | 0.62 |
| 500 | 93.7 | 0.70 | 93.8 | 0.59 |

Table: Coverage and average CI length, $p = 5$

| | Wald | | Deviance type | |
|---|---|---|---|---|
| $n$ | Coverage | Length | Coverage | Length |
| 50 | 92.4 | 2.44 | 83.4 | 0.98 |
| 100 | 95.4 | 1.47 | 89.8 | 0.96 |
| 150 | 96.0 | 1.23 | 92.3 | 0.89 |
| 200 | 96.8 | 1.09 | 92.0 | 0.82 |
| 250 | 94.5 | 0.99 | 93.5 | 0.78 |
| 300 | 94.6 | 0.89 | 94.9 | 0.73 |
| 350 | 95.3 | 0.82 | 93.5 | 0.68 |
| 400 | 93.9 | 0.79 | 92.4 | 0.65 |
| 450 | 91.9 | 0.73 | 94.4 | 0.61 |
| 500 | 92.4 | 0.70 | 92.0 | 0.59 |

# Jerusalem Perinatal Cohort Schizophrenia Study

Follow-up data on 92,000 individuals born between 1964 and 1976 to Israeli women living in Jerusalem and the adjoining rural areas.

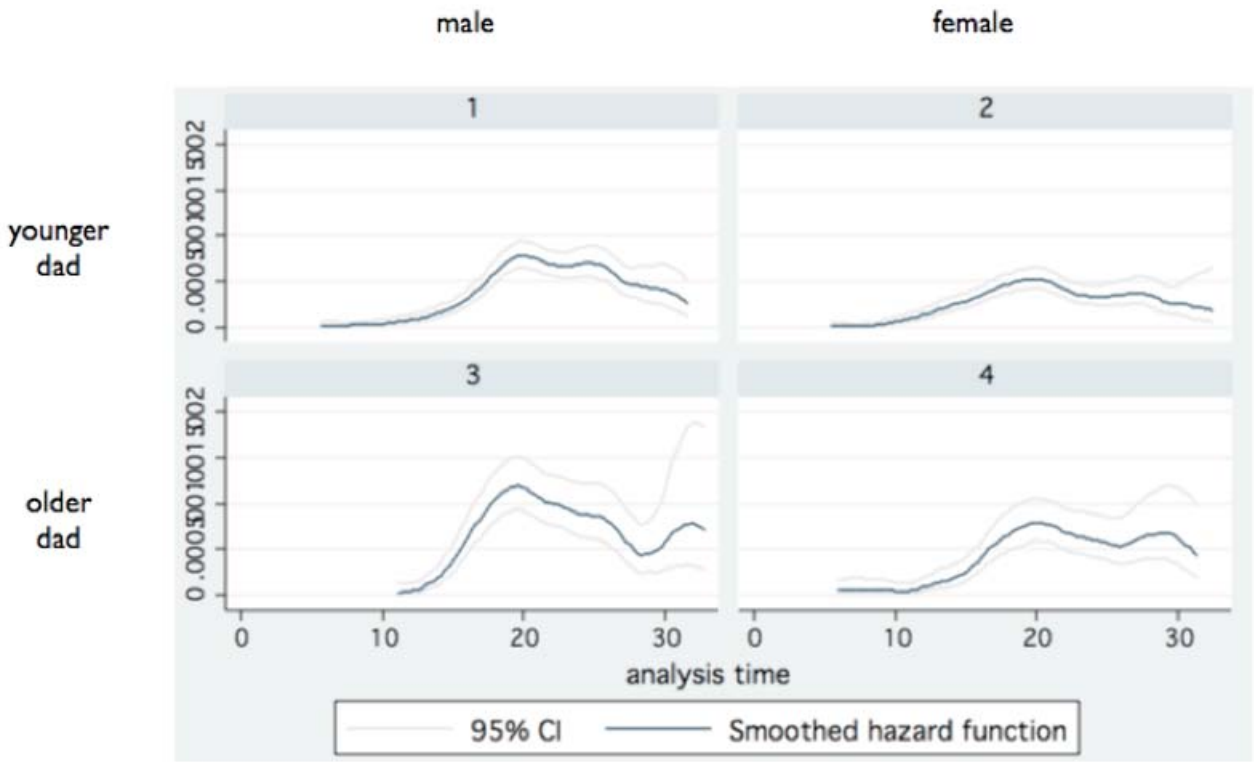
We analyze the data on 87,642 of these individuals for which complete covariate information is available.

Right censored survival data: indicator of a diagnosis of schizophrenia, and age (in years) at the time of diagnosis.

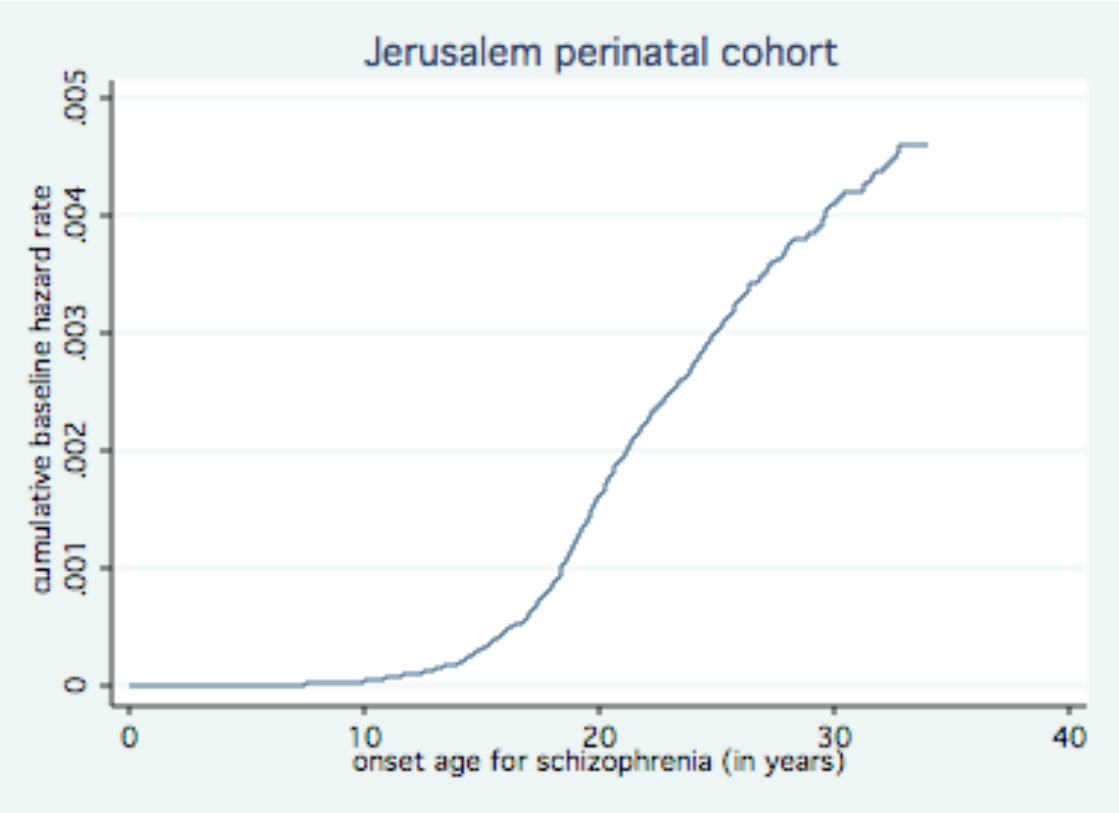
Covariates of interest : indicator male, indicator of low social class, paternal age at the time of the individual's birth

- Malaspina et. al. (2001) demonstrate a steady increase in schizophrenia risk with advanced paternal age.
- The rate of genetic mutation in paternal germ cells is known to increase significantly with age.
- Such increased mutation frequency has a strong clinical association with strong paternal age effects for multiple diseases and disorders, possibly because of accumulating replication errors in spermatogonial cell lines.

# Subgroup analyses



# Breslow's estimate $\hat{\Lambda}_n(t)$



# Results

$\tau = 30$ years

bandwidth for estimating $\lambda'(d^0)$ is 1 year

$\hat{d}_n = 16.69$ years

95% Wald CI for $d^0$: 15.79–17.59

95% Deviance CI for $d^0$: 16.29–17.06

$\hat{\lambda}_l = 2.39 \times 10^{-5}$

$\hat{\lambda}_u = 20.16 \times 10^{-5}$

# Other potential applications

- Deciding the time (or age) at which vaccination or diagnostic testing is advisable from a public health point of view.
- Our results can be used to determine a confidence interval not only for the threshold, but also for the relative risk $\lambda_u^0/\lambda_l^0$ across the threshold.
- Large values of this relative risk would indicate a greater necessity for medical intervention.
- Covariate thresholds are also of interest, e.g., in paternal age related effects in schizophrenia risk. Pons (2003) studies a covariate change-point Cox model, but not in the misspecified setting.

# Discussion

- Our approach is complimentary to change-point analysis in which the aim is to estimate the locations of existing jump discontinuities in an otherwise smooth curve.
- In change-point analysis, no distinction is made between the working model that has the jump point and the model that is assumed to generate the data.
- We use a model-robust approach that applies under arbitrary misspecification of the discontinuous working model.

# Summary

- We have studied the estimation of optimal binary decision tree approximations for functional parameters.
- The convergence rate $n^{1/3}$ contrasts with the rate of $n$ under (correctly specified) change-point models.
- The estimators of the mean levels on either side of the threshold are also $n^{1/3}$-consistent, in contrast to the corresponding change-point estimators which are $\sqrt{n}$-consistent with normal limits.

# References

M. Banerjee and I. W. McKeague. Confidence Sets for Split Points in Decision Trees. *Annals of Statistics* **35** 543–574 (2007).

M. Banerjee and I. W. McKeague. Estimating Optimal Step-function Approximations to Instantaneous Hazard Rates. *Bernoulli* **13** 279–299 (2007).