Estimating conditional extremes

Keith Knight
University of Toronto

e-mail: `keith@utstat.toronto.edu`

homepage: `www.utstat.toronto.edu/keith/home.html`

Research supported by NSERC

Outline of talk

I. Introduction

II. Estimation

- M-estimation
- invariance in location case

III. Asymptotics

- point process convergence
- epi-convergence in distribution
- asymptotics for M-estimators

IV. Other things

- Barrier regularization
- “Soft” extremes

I. Introduction

- Consider a linear regression model with positive errors:

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + W_i \quad (i = 1, \cdots, n)$$

  where the W_i's are independent with

$$\begin{aligned} \text{ess inf } W_i &= 0 \\ P(W_i \leq w | \boldsymbol{x}_i) &= \lambda(\boldsymbol{x}_i) w^{\alpha} L(w) \quad (\alpha > 0). \end{aligned}$$

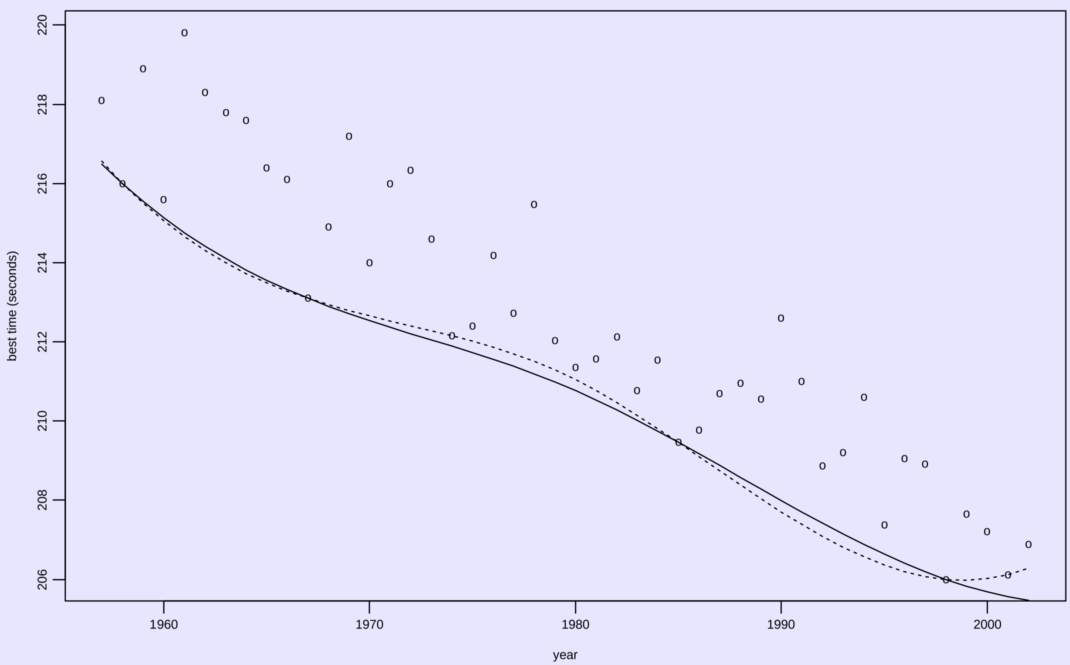
  ($L(w)$ slowly varying at 0.)

- We can view $\boldsymbol{x}_i^T \boldsymbol{\beta}$ as conditional minimum of Y_i.
- This type of model is also appropriate for "record" data.

Example: Yearly best times in men's (outdoor) 1500m races from 1957 to 2002:

$$\text{Time}(\text{year}) = g(\text{year}) + W(\text{year})$$

where g can be interpreted as the absolutely best possible time.



Spline estimates (4 knots) using constrained least squares and L_1 estimation

- Intuitively, we should be able to estimate $\boldsymbol{\beta}$ most efficiently when the boundary is well-defined by the observations $\Rightarrow W_i$'s have significant probability mass around 0.
- Similar issues arise in
  - production frontier estimation (Aigner & Chu, 1968; Simar & Wilson, 2000; Florens & Simar, 2002)
  - estimation of point process boundaries (e.g. Girard & Menneteau, 2003; Bouchard *et al.*, 2003) .

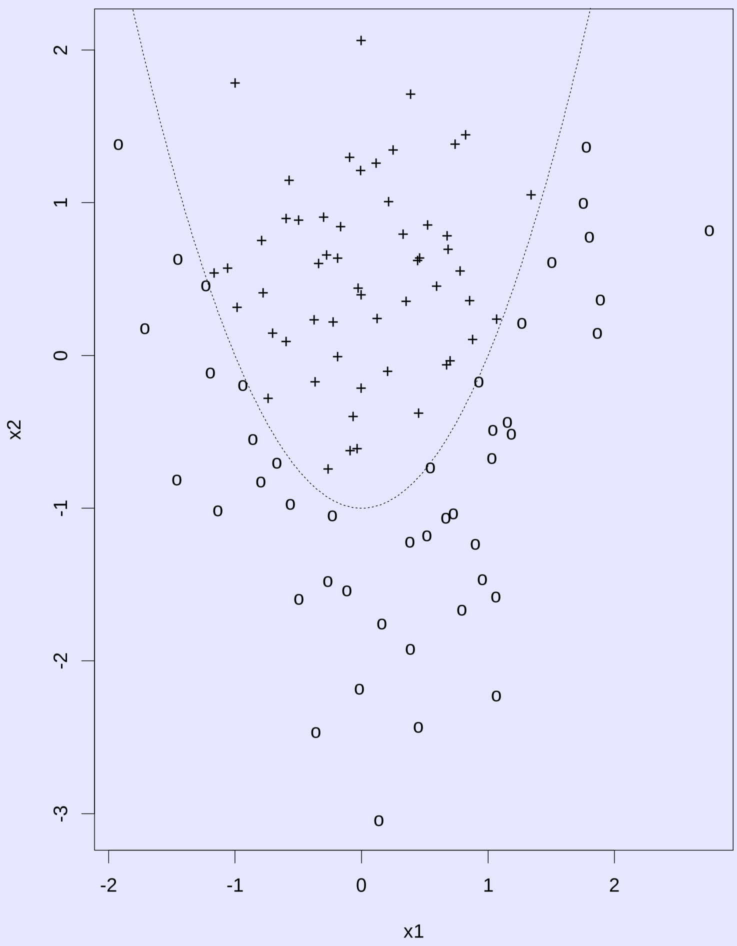
$\Rightarrow$ Different models but similar issues in estimation and asymptotics.

- We are assuming that $\{W_i\}$ are in the domain of attraction of a Type III extreme value (Weibull) distribution.
- In this case, the conditional minimum is well-defined.
- We can also consider properties of estimators for $\{W_i\}$ in other extreme value domains of attraction.

- Similar problems arise also in classification, particularly when we can assume “separability”.
- Data consist of “feature” $\{\boldsymbol{x}_i\}$ and classes labelled by $\{Y_i\}$ — assume simple case $Y_i = \pm 1$.
- Classification rule: $\widehat{Y} = \mathrm{sgn}(\widehat{g}(\boldsymbol{x}))$, for example, $\widehat{g}(\boldsymbol{x}) = \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}$.
- Maximum margin estimator: Maximize $h \geq 0$ subject to

$$Y_i \boldsymbol{x}_i^T \boldsymbol{\beta} \geq h \quad \text{for } i = 1, \cdots, n$$

and $\|\boldsymbol{\beta}\|_1 = 1$.



Maximum margin estimate of a quadratic boundary.

II. Estimation

1. M-estimation

- Minimal requirement for $\widehat{\boldsymbol{\beta}}$:

$$Y_i \geq \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} \quad \text{for all } i$$

(since $Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\beta}$ for all i).

- Pseudo-ML consideration: Assume the W_i's have a density

$$f(w) = \exp(-\rho(w)) \quad (w > 0)$$

$\rho(w) \to \infty$ as $w \to \infty$. Then the MLE $\widehat{\boldsymbol{\beta}}_n$ minimizes

$$\sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}) \quad \text{subject to} \quad Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\phi}$$

for $i = 1, \cdots, n$.

- Aigner & Chu (1968) consider estimation with $\rho(w) = w$ and $\rho(w) = w^2$ for estimating production frontier functions.
- For $\rho(w) = w$, $\widehat{\boldsymbol{\beta}}_n$ is the solution of a linear programming problem and can also be viewed as a regression quantile estimator (Koenker & Bassett, 1978) of order $\alpha = 0$; that is, as $\alpha \to 0$, $\widehat{\boldsymbol{\beta}}$ is the limit of

$$\operatorname{argmin} \sum_{i=1}^{n} \rho_\alpha(Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})$$

  where $\rho_\alpha(x) = x[\alpha - I(x < 0)]$.
- Asymptotics for this estimator are given by Smith (1994), Portnoy & Jureckova (1999), and Knight (2001) under various regularity conditions.

- Assume smoothness for ρ:

$$\rho(w) = \int_0^w \psi(t)\, dt$$

  where ψ is Hölder continuous.

- We will also assume that the *right* tail of $\{W_i\}$ is not too heavy relative to ψ.

Problem: What are the asymptotics for general ρ?

- How does the asymptotic behaviour depend on ρ?
- What determines the asymptotics of $\widehat{\boldsymbol{\beta}}_n$ in general?

2. Location case

- In the location case (i.e. $Y_i = \theta + W_i$), the situation is straightforward: If $\widehat{\theta}_n$ minimizes

$$\sum_{i=1}^{n} \rho(Y_i - \phi) \quad \text{subject to} \quad Y_i \geq \phi \quad \text{for all } i$$

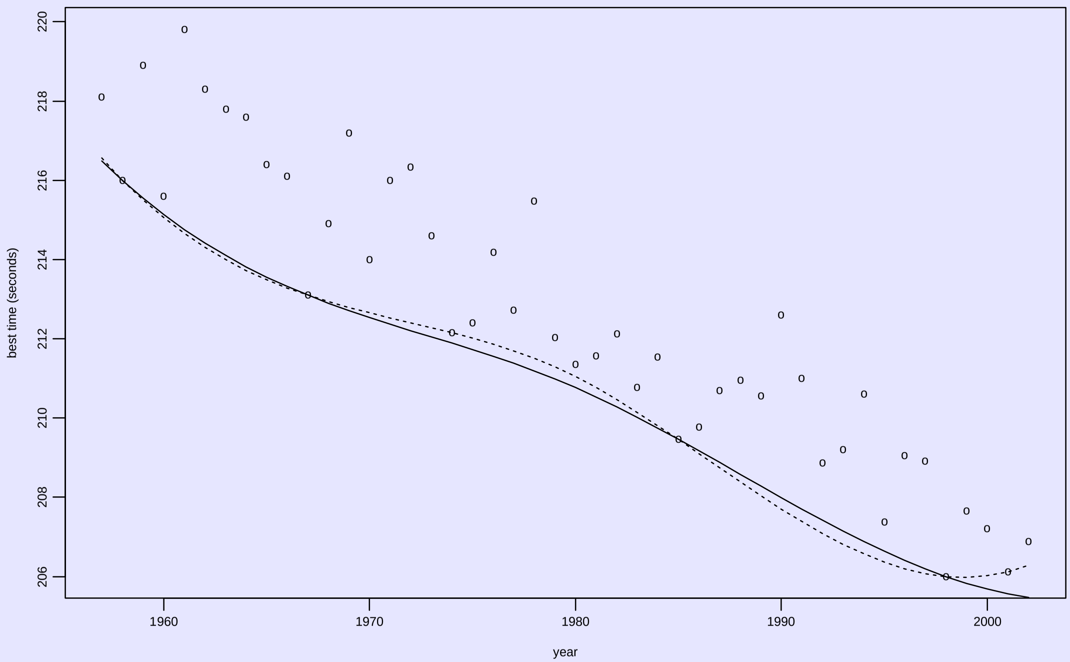
  then $\widehat{\theta}_n = \min_{i \leq n} Y_i$ (at least for sufficiently large n) over a wide class of ρ with $E[\psi(W_i)] > 0$.

- $\widehat{\theta}_n$ inherits the asymptotic properties of $\min_{i \leq n} Y_i$.

Question: How does this "invariance" extend to the regression setting?

Example: 1500 metre data (1957-2002)

- Look (again) at estimates for spline basis with with 4 knots using $\rho(w) = w$ (dotted) and $\rho(w) = w^2$ (solid).



- Estimates are close but not equal; what determines the dependence on ρ?

III. Asymptotics

1. Convergence of point processes and epi-convergence in distribution

- There are two issues to confront in determining asymptotics for boundary estimators:
  - (i) estimators are essentially determined by observations close to the boundary (i.e. influence of distant observations is negligible);
  - (ii) "classical" asymptotic techniques are difficult to apply due to the constraints.
- We will deal with (i) using point process asymptotics and with (ii) using epi-convergence in distribution.

Point process convergence

- Characterize point processes as random integer-valued measures:

$$N(A) = \#\text{ of points lying in } A$$

- Convergence of a sequence of point processes $\{N_n\}$ characterized by weak convergence of integrals:

$$N_n \xrightarrow{d} N_0 \quad \text{iff} \quad \int g(t)\, N_n(dt) \xrightarrow{d} \int g(t)\, N_0(dt)$$

  for all bounded continuous functions g with compact support.

- If N_0 is a Poisson process (i.e. $N_0(A) \sim \text{Pois}(\lambda(A))$ for each A) then the $\xrightarrow{d}$ condition can be simplified.

Epi-convergence in distribution

- Suppose that $\boldsymbol{U}_n$ minimizes an objective function ξ_n over some (closed) set C_n.
- This is equivalent to minimizing

$$Z_n(\boldsymbol{u}) = \begin{cases} \xi_n(\boldsymbol{u}) & \text{if } \boldsymbol{u} \in C_n \\ +\infty & \text{otherwise} \end{cases}$$

Question: What's the weakest form of weak convergence of $\{Z_n\}$ to Z that guarantees

$$\boldsymbol{U}_n = \text{argmin}(Z_n) \xrightarrow{d} \text{argmin}(Z)$$

when $\text{argmin}(Z_n) = O_p(1)$ and $\text{argmin}(Z)$ is unique?

Answer: Epi-convergence in distribution. (see Pflug, 1994; Geyer, 1996)

- Epi-convergence is actually convergence (with respect to the appropriate topology) of the epi-graphs of the objective functions (which are assumed to be lower-semicontinuous).
- For convex objective functions, finite dimensional weak convergence is sufficient for epi-convergence in distribution provided that the limit is finite on an open set.

2. Asymptotics for boundary M-estimators

- $\widehat{\boldsymbol{\beta}}_n$ minimizes

$$\sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}) \quad \text{subject to} \quad Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\phi}$$

  for $i = 1, \cdots, n$ where ρ is convex and reasonably smooth.

- Look at case where W_i's are i.i.d. first; assume that
  - $F(w) = P(W_i \leq w) = w^\alpha L(w)$,
  - for some probability measure μ,

$$\frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{x}_i \in A) \to \mu(A).$$

  Define $\{a_n\}$ such that $n\, F(t/a_n) = t^\alpha \Rightarrow a_n = n^{1/\alpha} L^*(n)$.

Key point: The asymptotics are determined by $O(1)$ points within $O(a_n^{-1})$ of the boundary $\Rightarrow$ point process asymptotics

- We start by defining the objective function

$$Z_n(\boldsymbol{u}) = \frac{a_n}{n} \sum_{i=1}^{n} \left[\rho(W_i - \boldsymbol{x}_i^T \boldsymbol{u}/a_n) - \rho(W_i)\right]$$

  if $a_n W_i \geq \boldsymbol{x}_i^T \boldsymbol{u}$ for all i with $Z_n(\boldsymbol{u}) = +\infty$ otherwise.

- Note that $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \text{argmin}(Z_n)$.
- We need to determine the epi-limit of $\{Z_n\}$.
- Assume that $E[\psi^2(W_1)] < \infty$ and some additional regularity conditions.

- Using point process techniques, we can show that $Z_n \stackrel{e-d}{\longrightarrow} Z$ where

$$\begin{aligned} Z(\boldsymbol{u}) &= -E[\psi(W_1)] \int \boldsymbol{u}^T \boldsymbol{x}\, \mu(d\boldsymbol{x}) \\ &= -E[\psi(W_1)] \boldsymbol{u}^T \boldsymbol{\gamma} \\ &\quad \text{if } \Gamma_k \geq \boldsymbol{X}_k^T \boldsymbol{u} \text{ for } k = 1, 2, \cdots \end{aligned}$$

  and $Z(\boldsymbol{u}) = +\infty$ otherwise.

- $\{(\Gamma_k, \boldsymbol{X}_k) : k \geq 1\}$ are the points of a Poisson process N_0 with

$$E[N_0(ds \times d\boldsymbol{x})] = \alpha s^{\alpha-1}\, ds\, \mu(d\boldsymbol{x}).$$

- $\{\Gamma_k\}$ and $\{\boldsymbol{X}_k\}$ are independent sequences.

- Then $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \text{argmin}(Z)$, which is the solution of a linear program where the (random) constraints are determined by the Poisson process.
- Note that the limiting distribution does not depend on ρ, at least when $E[\psi^2(W_1)] < \infty \Rightarrow$ *asymptotic invariance*.

- However, the invariance fails in the non-i.i.d. case where the distribution of W_i depends on $\boldsymbol{x}_i$.
- Here we have $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \text{argmin}(Z)$ where

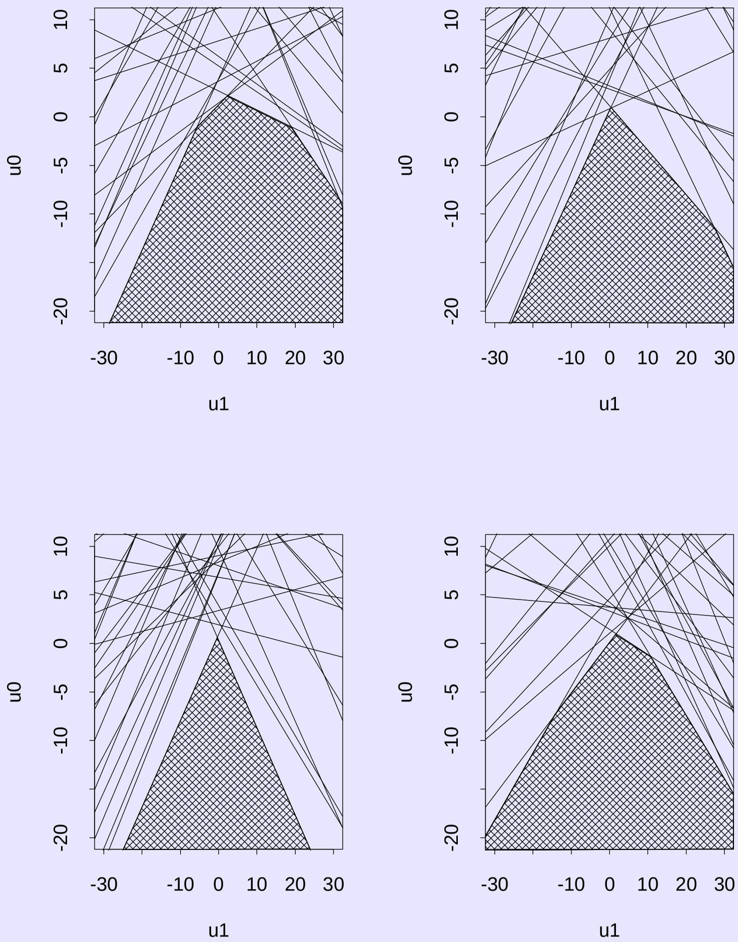
$$Z(\boldsymbol{u}) = -\int E[\psi(W|\boldsymbol{x})]\boldsymbol{x}^T\boldsymbol{u}\,\mu(d\boldsymbol{x}) = -\boldsymbol{u}^T\boldsymbol{\gamma}(\rho)$$

if $\Gamma_k^* \geq \boldsymbol{X}_k^T\boldsymbol{u}$ for $k = 1, 2, \cdots$ and $Z(\boldsymbol{u}) = \infty$ otherwise.

- As before, $\{\Gamma_k^*, \boldsymbol{X}_k\}$ are points of a (possibly different) Poisson process that does *not* depend on ρ.

- Only *finite* part of the limiting objective function depends on ρ.
  - The constraints do *not* depend on ρ.
  - If $\boldsymbol{\gamma}(\rho_1)$ is close to $\boldsymbol{\gamma}(\rho_2)$ then the respective minimizers will be exactly equal with high probability.
- Thus we have "near" invariance.

Example: Look at feasible regions and constraint lines for $\alpha = 1$ and $\boldsymbol{X}_k = (1, U_k)$ where $\{U_k\}$ are i.i.d. uniform r.v.'s on $[-1, 1]$. In this case, $\boldsymbol{\gamma}(\rho) \propto (1, c_\rho)^T$ where $-1 < c_\rho < 1$.



Feasible regions (shaded) and constraint lines

Other extreme value domains of attraction

- It's possible to extend the results to other extreme value domains of attraction:
  - Type I: $P(W < -x) \to 0$ exponentially as $x \to \infty$.
  - Type II: $P(W < -x) = x^{-\alpha} L(x)$ for $\alpha > 0$ and L slowly varying.
- To derive limiting distributions, we need to be careful to define $\rho(w)$ appropriately for $w < 0$.

IV. Other things

1. Barrier regularization

- $\boldsymbol{x}^T \widehat{\boldsymbol{\beta}}_n$ tends to be biased upwards.
- One possible way of removing bias is to add a barrier function to push estimated conditional minimum downwards.
- For a positive tuning parameter ϵ define $\widehat{\boldsymbol{\beta}}_n(\epsilon)$ to minimize

$$\sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}) + \epsilon \sum_{i=1}^{n} \tau(Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi})$$

  subject to $Y_i \geq \boldsymbol{x}_i^T \boldsymbol{\phi}$ for all i.

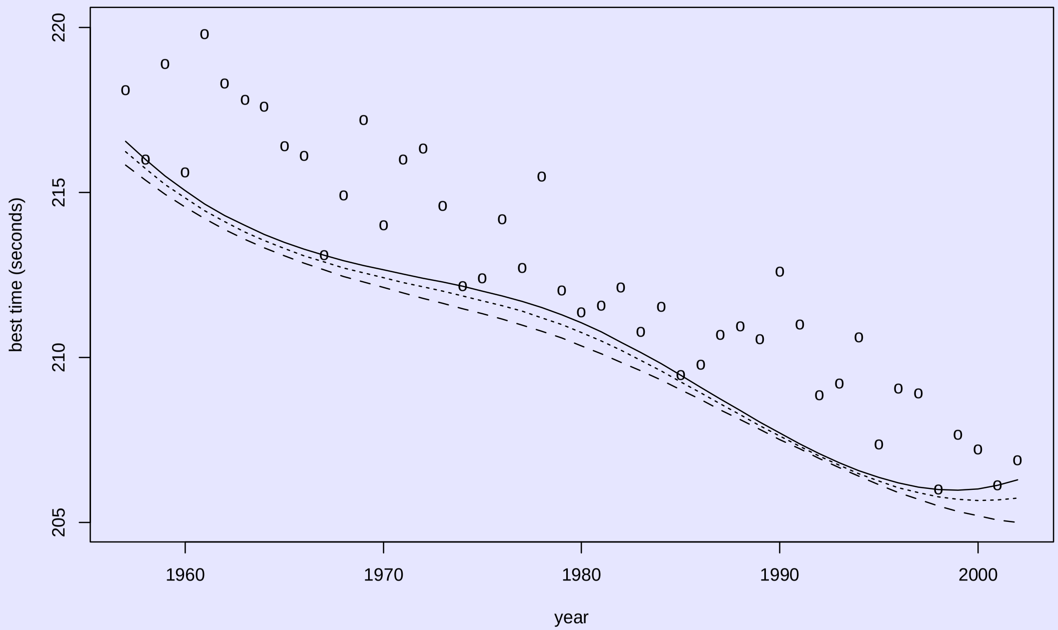
- $\tau(w)$ (barrier function) is a convex function satisfying

$$\lim_{w \downarrow 0} \tau(w) = +\infty.$$

- We can take $\tau(w) = w^{-r}$ for $r > 0$ or $\tau(w) = -\ln(w)$.
- For a given $\epsilon > 0$, $\widehat{\boldsymbol{\beta}}_n$ lies in the interior of the constraint set; that is,

$$Y_i > \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_n(\epsilon) \quad \text{for all } i$$

- Computational advantages:
  - $\widehat{\boldsymbol{\beta}}_n(\epsilon)$ can be computed using Newton or quasi-Newton methods;
  - $\widehat{\boldsymbol{\beta}}_n$ can be obtaining from $\{\widehat{\boldsymbol{\beta}}_n(\epsilon)\}$ by taking $\epsilon \downarrow 0$ — interior point algorithms (Fiacco & McCormick, 1990; Koenker & Portnoy, 1997).



Barrier regularized estimates using $\rho(w) = w$ and $\tau(w) = w^{-2}$.

Solid line is the extreme regression quantile line.

2. “Soft” conditional extremes

- **Idea:** Allow a small number of the constraints to be violated.
- **Rationale:** Robustness
  - Estimates of conditional extremes are naturally very sensitive to extreme observations.
  - It’s often desirable to downweight or ignore such observations in the interest of model fidelity.
- But we don’t want to specify *a priori* the number of constraints to be violated.

- Note that the M-estimator $\widehat{\boldsymbol{\beta}}_n$ minimizes

$$\sum_{i=1}^{n} \varrho(Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi})$$

where

$$\varrho(w) = \begin{cases} \rho(w) & \text{for } w \geq 0 \\ +\infty & \text{for } w < 0. \end{cases}$$

- Replace ϱ by the "softened" version

$$\bar{\varrho}(w) = \begin{cases} \rho(w) & \text{for } w \geq 0 \\ \epsilon^{-1}\psi(w) & \text{for } w < 0 \end{cases}$$

where $\epsilon > 0$ and $\psi(w) \to +\infty$ as $w \to -\infty$.

- ψ should be a concave function to get the desired result, for example, $\psi(w) = (-w)^r$ for $0 < r < 1$.
  - Taking $\psi(w)$ to be convex, we get essentially (for small ϵ) regression quantiles.
  - Concavity of ψ allows some adaptability and allows for $\widehat{\boldsymbol{\beta}}_n(\epsilon) = \widehat{\boldsymbol{\beta}}_n$.
- More work needs to be done:
  - Computational algorithm for $\widehat{\boldsymbol{\beta}}_n(\epsilon)$.
  - If we let $\epsilon \downarrow 0$, we get an exterior point algorithm for computing $\widehat{\boldsymbol{\beta}}_n$ — see Fiacco & McCormick (1990).
  - Asymptotics.