Isotonic Regression
Statistics 710
September 26, 2006

The Problem. The isotonic regression problem may be stated as follows: suppose that

$$y_i = \theta_i + \epsilon_i, \;\; i = 1, \cdots, n, \tag{1}$$

where

$$-\infty < \theta_1 \le \theta_2 \le \cdots \le \theta_n < \infty \tag{2}$$

and $\epsilon_1, \cdots, \epsilon_n$ are uncorrelated random errors with means 0 and variances of the form $E(\epsilon_i^2) = \sigma_i^2/w_i$, where $w_1, \cdots, w_n > 0$ are known and $0 < \sigma^2 < \infty$ may be known or unknown. For example, if y_i is the average of n_i independent measurments of θ_i and all random errors have the same variance, then $w_i = n_i$. For another example, suppose that $\theta_i = f(t_i)$, where $-\infty < t_1 < t_2 < \cdots < t_n < \infty$ and μ is a non-decreasing function. The least squares estimates minimize

$$SS = \sum_{i=1}^{n} w_i[y_i - \theta_i]^2 := \|y - \theta\|_w^2, \tag{3}$$

with respect of $\theta = [\theta_1 \cdots, \theta_n]'$, subject to (2), where $y = [y_1, \cdots, y_n]'$ and $w = [w_1, \cdots, w_n]'$.

Example 1 *The following data (part of a larger data set) give temperature anomalies from 1856 to 1880 (1900 = base):* $-.381, -.461, -.415, -.225, \cdots, -.289, -.295$. *A plot of the data is given in Figure 1.*

The Solution. The set Ω of $\theta \in I\!R^n$ for which (2) is a convex subset of $I\!R^n$. Thus, the minimization problem has a unique solution $\hat\theta$, the projection of y onto Ω with respect to $\langle \cdot, \cdot \rangle_w$; and $\hat\theta \in \Omega$ is characterized by the conditions $\hat\theta \in \Omega$, $\langle y - \hat\theta, \hat\theta \rangle_w = 0$, and $\langle y - \hat\theta, \xi \rangle_w \le 0$ for all $\xi \in \Omega$. The latter two conditions may be written

$$\sum_{i=1}^{n} w_i \hat\theta_i (y_i - \hat\theta_i) = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} w_i (y_i - \hat\theta_i)\xi_i \le 0 \tag{4}$$

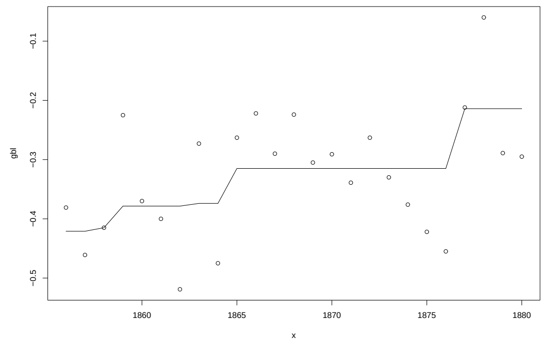

Figure 1: Global Temperature Anomalies

for all $\xi \in \Omega$. As a corollary $\langle y - \hat{\theta}, \mathbf{1} \rangle_w = 0$, since $\pm \mathbf{1} \in \Omega$.

The Cumulative Sum Diagram. Let

$$\hat{\Theta}_k = \hat{\theta}_1 + \cdots + \hat{\theta}_k$$
$$Y_k = y_1 + \cdots + y_k,$$

and

$$W_k = w_1 + \cdots + w_k$$

for $k = 0, \cdots, n$; and let $\hat{\Theta}$ and Y be piecewise linear functions with knots at $W_0, \cdots, W_n$ for which $\hat{\Theta}(W_k) = \hat{\Theta}_k$ and $Y(W_k) = Y_k$. Then $\hat{\Theta}$ is a convex function, since $\hat{\Theta}'_\ell(t) = \hat{\theta}_k$ for $W_{k-1} < t \leq W_k,\ k = 1, \cdots, n$, and this is a non-decreasing function. Moreover $\hat{\Theta}(t) \leq Y(t)$ for $0 \leq t \leq W_n$. By the piecewise linearity, it suffices to show this when $t = W_k$. Let $\xi = [-1, \cdots, -1, 0, \cdots]'$ ($k - 1$'s). Then $-\sum_{i=1}^k w_k[y_k - \hat{\theta}_k] \leq 0$ and, therefore $\hat{\Theta}(W_k) = \hat{\Theta}_k \leq Y_k = Y(W_k)$.

It will be shown that $\hat{\Theta}$ *is the largest convex function that is less than or equal to* Y, but two preliminary results are needed first. If $\hat{\theta}_k < \hat{\theta}_{k+1}$, then $\hat{\Theta}_k = Y_k$. To see this, let $\mathbf{1}_k = [1, \cdots, 1, 0, \cdots, 0]'$. Then $\hat{\theta} \pm \alpha \mathbf{1}_k \in \Omega$ for all sufficiently small $0 < \alpha < \theta_{k+1} - \theta_k$, so that $\langle y - \hat{\theta}, \hat{\theta} \pm \alpha \mathbf{1}_k \rangle_w \leq 0$; and this implies $\pm \langle y - \hat{\theta}, \alpha \mathbf{1}_k \rangle_w \leq 0$, or equivalently, $Y_k = \hat{\Theta}_k$. Next, if $0 < t < W_n$, let $j \geq 0$ be the largest index for which $W_j < t$ and $\hat{\Theta}_j = Y_j$, and let $k \leq n$ be the smallest index for which $\hat{\Theta}_k = Y_k$. Then $\hat{\Theta}$ is linear on $[W_j, W_k]$. For otherwise, there would be an i for which $j < i < k$ and $\hat{\theta}_i < \hat{\theta}_{i+1}$; but then $Y_i = \hat{\Theta}(W_i)$, contradicting the definition of j or k.

Now, let G be any convex function for which $G(t) \leq Y(t)$ for $0 \leq t \leq W_n$; let $0 < t < W_n$;

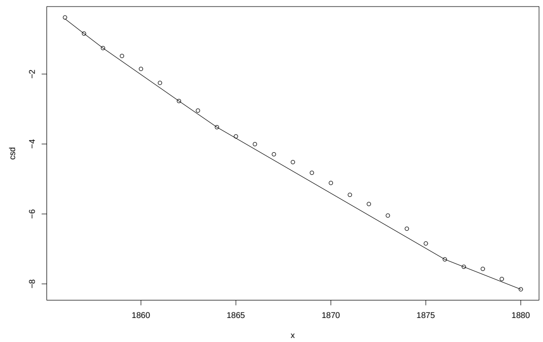

Figure 2: The Cumulative Sum Diagram and its Greatest Convex Minorant

and let j and k be as above. Then

$$\begin{aligned} G(t) &\leq \frac{(t-W_j)G(W_k)+(W_k-t)G(W_j)}{W_k-W_j} \\ &\leq \frac{(t-W_j)Y_k+(W_k-t)Y_j}{W_k-W_j} \\ &\leq \frac{(t-W_j)\hat{\Theta}(W_k)+(W_k-t)\hat{\Theta}(W_j)}{W_k-W_j} = \hat{\Theta}(t). \end{aligned}$$

The cumulative sum diagram and its greatest convex minorant are displayed in Figure 2. *Thus, $\hat{\theta}_k$ is the left hand derivative of the greatest convex minorant $\hat{\Theta}$ to the cumulative sum diagram Y.*

The following is implicit in the derivation. If $c \in \{\hat{\theta}_1, \cdots, \hat{\theta}_n\} = V$, say, then

$$\sum_{j:\hat{\theta}_j = c} (y_j - c)c_j = 0.$$

For the set of j for which $\hat{\theta}_j = c$ is an interval $\{i, \cdots, k\}$. Let $e_j = 1$ or 0 depending on whether $\theta_j = c$, or not. Then $\hat{\theta} \pm \alpha e \in \Omega$ for small α, so that $\langle y - \hat{\theta}, e\rangle_w = 0$ by (4) and

$$\sum_{j:\hat{\theta}_j = c} (y_j - c)w_j = 0 = \langle y - \hat{\theta}, e\rangle_w = 0$$

As a consequence, if $h : V \to I\!R$ is any function, then

$$\sum_{i=1}^{n} (y_i - \hat{\theta}_i)h(\hat{\theta}_i)w_i = 0. \tag{5}$$

This may be seen by summing over the distinct values of h.

The Pool Adjacent Violators Algorithm. The characterization of $\hat{\Theta}$ is the basis for the following algorithm: starting with $\hat{\theta}^0 = y$:

a) If $\hat{\theta}^k_{j-1} \leq \hat{\theta}^k_j$, let $\hat{\theta} = \hat{\theta}^k$ and stop.

b) Otherwise, let j be the smallest index for which $\hat{\theta}^k_{j-1} > \hat{\theta}^k_j$; let

$$\hat{\theta}^{k+1}_{j-1} = \hat{\theta}^{k+1}_j = \frac{w_{j-1}\hat{\theta}^k_{j-1} + w_j\hat{\theta}^k_j}{w_{j-1} + w_j}$$

and $\hat{\theta}^{k+1}_i \leq \hat{\theta}^k_I$ for $j-1 \neq i \neq j$; then go back to a). The algorithm terminates after a finite number of steps. The proof that it delivers $\hat{\theta}$ is left as an exercise.

Problem 1 . Global temperature anomalies from 1856 to 2005 may be found at the website

http://cdiac.ornl.gov/ftp/trends/temp/jonescru/global.dat or by entering "temperature anomalies" in Google. Find the isotonic regression with equal weights for this data set, and superimpose the estimated regression function on a scatter plot, as in Figure 1.

The Min-Max Formula. Let

$$\mathrm{av}(i,j) = \frac{w_iy_i + \cdots + w_jy_j}{w_i + \cdots + w_j} = \frac{Y(W_j) - Y_(W_{i-1})}{W_j - W_{i-1}}$$

for $1 \leq i \leq j \leq n$. Then

$$\hat{\theta}_k = \max_{i \leq k} \min_{j \geq k} \mathrm{av}(i,j) \tag{6}$$

for $1 \leq k \leq n$. To see this (geometrically), let $S = \{\ell : \hat{\theta}_\ell < \hat{\theta}_\ell\} \cup \{0, n\}$ and observe that

$$\hat{\theta}_k = \max_{i \leq k} \min_{j \geq k} \frac{\hat{\Theta}(W_j) - \hat{\Theta}(W_{i-1})}{W_j - W_{i-1}} = \max_{i \in S, i \leq k} \min_{j \in S, j \geq k} \frac{\hat{\Theta}(W_j) - \hat{\Theta}(W_{i-1})}{W_j - W_{i-1}},$$

by convexity. Next, recalling that $\hat{\Theta}_\ell = Y_\ell$ for $\ell \in S$,

$$\hat{\theta}_k = \max_{i \in S, i \leq k} \min_{j \in S, j \geq k} \frac{Y(W_j) - Y(W_{i-1})}{W_j - W_{i-1}}.$$

Clearly,

$$\min_{j \geq k} \frac{Y(W_j) - Y(W_{i-1})}{W_j - W_{i-1}} \leq \min_{j \in S, j \geq k} \frac{Y(W_j) - Y(W_{i-1})}{W_j - W_{i-1}}$$

In fact, there is equality. For if $j \geq k$ and $i \in S$, then

$$\frac{Y(W_j) - Y(W_{i-1})}{W_j - W_{i-1}} \geq \frac{\hat{\Theta}(W_j) - \hat{\Theta}(W_{i-1})}{W_j - W_{i-1}} \geq \min_{j' \in S, j' \geq k} \frac{\hat{\Theta}(W_{j'}) - \hat{\Theta}(W_{i-1})}{W_{j'} - W_{i-1}}.$$

Relation (6) follows from this and a dual argument for i.

Generalized Isotonic Regression. Now let I be an interval; let $\psi \to I\!R$ be a convex function, and let

$$\Psi(w,z) = \psi(w) - \psi(z) - \psi'(z)(w-z)$$

for $w, z \in I$. *If* $y_1, \cdots, y_n \in I$, *then*

$$\sum_{i=1}^{n} \Psi(y_i, \theta_i) w_i \geq \sum_{i=1}^{n} \Psi(y_i, \hat{\theta}_i) w_i + \sum_{i=1}^{n} \Psi(\hat{\theta}_i, \theta_i) w_i \tag{7}$$

for all $\theta \in \Omega$. *Consequently, the left side of (7) is minimized when* $\theta = \hat{\theta}$. To see this, observe first that

$$\begin{aligned}
\Psi(y_i, \theta_i) - [\Psi(y_i, \hat{\theta}_i) + \Psi(\hat{\theta}_i, \theta_i)] &= \psi(y_i) - \psi(\theta_i) - \psi'(\theta_i)(y_i - \theta_i) \\
&\quad - [\psi(y_i) - \psi(\hat{\theta}_i) - \psi'(\hat{\theta}_i)(y_i - \hat{\theta}_i) \\
&\qquad + \psi(\hat{\theta}_i) - \psi(\theta_i) - \psi'(\theta_i)(\hat{\theta}_i - \theta_i) \\
&= [\psi'(\hat{\theta}_i) - \psi'(\theta_i)](y_i - \hat{\theta}_i).
\end{aligned}$$

So,

$$\mathrm{LHS}(7) - \mathrm{RHS}(7) = \sum_{i=1}^{n} w_i [\psi'(\hat{\theta}_i) - \psi'(\theta_i)](y_i - \hat{\theta}_i).$$

Here

$$\sum_{i=1}^{n} w_i \psi'(\hat{\theta}_i)(y_i - \hat{\theta}_i) = 0,$$

by (5). Next $\xi = [\psi(\theta_1), \cdots, \psi(\theta_n)]' \in \Omega$, since ψ' is non-decreasing, and

$$\sum_{i=1}^{n} w_i \psi'(\theta_i)(y_i - \hat{\theta}_i) = \langle \xi, y - \hat{\theta} \rangle_w \leq 0,$$

by (4). It follows that LHS(7)-RHS(7) ≥ 0, completing the proof of (7).

Example 2 . If $Y_i \sim \text{Poisson}(w_i \theta_i),\ i = 1, \cdots, n$ are independent, then the log-likelihood function is

$$\ell(\theta | y) = \sum_{i=1}^{n} w_i [y_i \log(\theta_i) - \theta_i] + C,$$

where C does not depend on θ. Here $y = [y_1, \cdots, y_n]'$ and $\theta = [\theta_1, \cdots, \theta_n]'$ denote the vectors. Let $\psi(z) = z \log(z) - z$. Then $\psi'(z) = \log(z)$, so that ψ is convex. Next

$$\Psi(y_i, \theta_i) = [y_i \log(y_i) - y_i] - [\theta_i \log(\theta_i) - \theta_i] - (y_i - \theta_i) \log(\theta_i) = \theta_i - y_i \theta_i + \psi(y_i),$$

and

$$-\ell(\theta | y) = \sum_{i=1}^{n} w_i \Psi(y_i, \theta_i) + C'.$$

Suppose now that the θ_i are non-decreasing, so that $\theta \in \Omega$. Then the MLE is isotonic regression $\hat{\theta}$ of y with weights w.

Remarks. *This material is taken from [1].*

References

[1] Robertson, T., F. Wright, and R. Dykstra (1988). *Order Restricted Inference*. Wiley