Isotonic Regression
Statistics 710
September 26, 2006

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

\[ y_i = \theta_i + \epsilon_i, \ i = 1, \ldots, n, \]  

(1)

where

\[-\infty < \theta_1 \leq \theta_2 \leq \cdots \leq \theta_n < \infty \]  

(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 measurements of \( \theta_i \) 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 \( \mu \) is a non-decreasing function. The least squares estimates minimize

\[
SS = \sum_{i=1}^{n} w_i [y_i - \hat{\theta}_i]^2 := \|y - \hat{\theta}\|_w^2,
\]  

(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 \( \Omega \) of \( \theta \in \mathbb{R}^n \) for which (2) is a convex subset of \( \mathbb{R}^n \). Thus, the minimization problem has a unique solution \( \hat{\theta} \), the projection of \( y \) onto \( \Omega \) 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 \( \langle y - \hat{\theta}, \xi \rangle_w \leq 0 \) for all \( \xi \in \Omega \). The latter two conditions may be written

\[
\sum_{i=1}^{n} w_i \hat{\theta}_i(y_i - \hat{\theta}_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} w_i (y_i - \hat{\theta}_i)\xi_i \leq 0
\]  

(4)
Figure 1: Global Temperature Anomalies

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

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 $\hat{\Theta}'(t) = \hat{\theta}_k$ 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 $1_k = [1, \cdots, 1, 0, \cdots, 0]'$. Then $\hat{\theta} \pm \alpha 1_k \in \Omega$ for all sufficiently small $0 < \alpha < \hat{\theta}_{k+1} - \hat{\theta}_k$, so that $\langle y - \hat{\theta}, \hat{\theta} \pm \alpha 1_k \rangle_w \leq 0$; and this implies $\pm \langle y - \hat{\theta}, \alpha 1_k \rangle_w \leq 0$, 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 $0 < t < W_n$;
and let $j$ and $k$ be as above. Then

$$G(t) \leq \frac{(t - W_j)G(W_k) + (W_k - t)G(W_j)}{W_k - W_j}$$

$$\leq \frac{(t - W_j)Y_k + (W_k - t)Y_j}{W_k - W_j}$$

$$\leq \frac{(t - W_j)\hat{\Theta}(W_k) + (W_k - t)\hat{\Theta}(W_j)}{W_k - W_j} = \hat{\Theta}(t).$$

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 $\hat{\theta}_j = c$, or not. Then $\hat{\theta} \pm \alpha e \in \Omega$ for small $\alpha$, 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 $h : V \rightarrow \mathbb{R}$ is any function, then

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

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 $\hat{\theta}_{j-1}^k \leq \hat{\theta}_j^k$, let $\hat{\theta} = \hat{\theta}_j^k$ and stop.

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

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

and $\hat{\theta}_i^{k+1} \leq \hat{\theta}_i^k$ 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

$$av(i, j) = \frac{w_i y_i + \cdots + w_j y_j}{w_i + \cdots + w_j} = \frac{Y(W_j) - Y(W_{i-1})}{W_j - W_{i-1}}$$

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

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

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

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

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 $\psi \rightarrow \mathcal{R}$ be a convex function, and let

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

4
for \( w, z \in I \). If \( y_1, \ldots, y_n \in I \), 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

\[
\Psi(y_i, \theta_i) - [\Psi(y_i, \hat{\theta}_i) + \Psi(\hat{\theta}_i, \theta_i)] = \psi(y_i) - \psi(\theta_i) - \psi'(\theta_i)(y_i - \theta_i) \\
= [\psi(y_i) - \psi(\hat{\theta}_i) - \psi'(\hat{\theta}_i)(y_i - \hat{\theta}_i) \\
+ \psi(\hat{\theta}_i) - \psi(\theta_i) - \psi'(\theta_i)(\hat{\theta}_i - \theta_i) \\
= [\psi'(\hat{\theta}_i) - \psi'(\theta_i)](y_i - \hat{\theta}_i).
\]

So,

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

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 \( \psi' \) 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 \( \text{LHS}(7) - \text{RHS}(7) \geq 0 \), 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 \( \theta \). 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 \( \psi \) 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 \( \theta_i \) 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