THE MINIMUM ABSOLUTE DEVIATION TREND LINE

CHARLES F. COOK

> "Since the desired curve is to be used for estimating, or predicting purposes, it is reasonable to require that the curve be such that it makes the errors of estimation small. . . . However, sums of absolute values are not convenient to work with mathematically; consequently [it is required] that the sum of the squares of the errors be a minimum."
>
> —Paul G. Hoel

Two problems arise out of the use of the method of least squares for determining an average claim cost trend line. First, a single odd point in the data has an excessive influence on the fitted line, and second, the oldest and newest points are given excessive weight relative to intermediate points, which may result in an inordinately large change in slope when a new point is added to the data and the oldest is deleted. These problems are not unique to our trend lines, but apply to all lines fitted by the method of least squares. They are the direct result of squaring the deviations between the data and the fittted line, which is simply not as reasonable a criterion of "best fit" as the absolute value of the deviation.

Why then do we use the method of least squares? Simply because absolute values are alleged to be mathematically inconvenient. This is not true; a trend line minimizing the sum of the absolute values of the deviations can be calculated, by the method shown in this paper, more easily than a least squares trend line. I do not mean to claim that all authors of books on mathematical statistics are wrong; but what is mathematically inconvenient to them is not necessarily inconvenient to an actuary. A minimum absolute deviation method of fitting a line is mathematically inconvenient for the following reasons:

1. It will not fit a *curved* line.
2. It requires equal intervals between measurements.
3. The form of calculation is an algorithm of the operations analysis type, rather than a concise mathematical formula.
4. It does not always produce a unique result; rather the minimum may be achieved for any slope a such that $m \leq a \leq n$.
5. It does not necessarily pass through the mean, so that the average deviation may not be zero, as it is for a line fitted by the method of least squares.

Inconveniences (1), (2), and (3) do not apply to our specific problem. Number (4) appears to be only theoretical; in practice, it seems adequate to use $a = (m + n)/2$, the mean of all slopes which produce the minimum.

We have chosen to resolve inconvenience (5) by *defining* our line as the one which minimizes the sum of absolute deviations, *subject to the condition that the average deviation be zero.* Incidentally, this condition not only yields an intuitively more reasonable result, but reduces the computational labor by about one-third.

THE MINIMUM ABSOLUTE DEVIATION ALGORITHM

Given n observations y_i associated with equally-spaced points x_i, the problem is to determine the values a, b such that $\sum_{i=1}^{n} | ax_i + b - y_i |$ is a minimum, subject to the condition $\sum_{i=1}^{n} (ax_i + b - y_i) = 0$.

1. If n is odd, set $x_i = -\left(\frac{n+1}{2}\right) + i$. If n is even, set $x_i = -(n + 1) + 2i$.
2. Calculate $\sum_{i=1}^{n} yi/n = \bar{y}$ and $\sum_{i=1}^{n} | x_i |/2 = MX$.
3. Calculate $a_i = \frac{y_i - \bar{y}}{x_i}$ for all i.
4. Order the a_i from least to greatest, such that $a_{i1} \leq a_{i2} \leq \ldots \leq a_{in}$.
5. Order the x_i the same way as their associated a_i.
6. Accumulate the $| x_{i_j} |$ to form $Z_k = \sum_{j=1}^{k} | x_{i_j} |$.
7. Find k^*, the least k for which $Z_k \geqslant MX$.
8. If $Z_{k^*} = MX$, then the desired line is $y' = \frac{a_{i_{k^*}} + a_{i_{k^*+1}}}{2} x + \bar{y}$.

   If $Z_{k^*} > MX$, then the desired line is $y' = a_{i_{k^*}} x + \bar{y}$.

Example

While at first reading the algorithm may seem complex, it is very simple to perform. All arithmetic except possibly the division $\sum_{i=1}^{n} y_i/n = \bar{y}$ in step (2) may be done mentally. All other divisors are small integers; there is no multiplication or squaring at all. The simplicity of the procedure is illustrated by the following example, in which all work is shown. It may be enlightening for the reader to try fitting a least squares line to the same data without benefit of calculator, slide rule, or scratch paper.

x_i	y_i	$(y_i - \bar{y})$	a_i	Rank	Z_k
−6	110	−3.4	.567	5	18
−5	109	−4.4	.880	10	
−4	112	−1.4	.350	4	12
−3	111	−2.4	.800	8	
−2	115	+1.6	−.800	1	2
−1	112	−1.4	1.400	12	
0	113	−0.4	——	—	
1	114	+0.6	.600	6	19
2	112	−1.4	−.700	2	4
3	116	+2.6	.867	9	
4	114	+0.6	.150	3	8
5	117	+3.6	.720*	7	24*
6	119	+5.6	.933	11	

$\sum |x_i| = 42$ $\quad$ $\sum y_i = 1{,}474$

$\frac{\sum |x_i|}{2} = 21$ $\quad$ $\bar{y} = 113.4$ $\quad$ $y' = .72\,x + 113.4$

Proof of the Algorithm

Lemma 1: If $E(a) = \sum_{j=1}^{n} | ax_i + \bar{y} - y_j |$, $0 < \Delta \leq (a_{i_{k+1}} - a_{i_k})$, and

$$\epsilon_k = E(a_{i_k} + \Delta) - E(a_{i_k}), \text{ then for any } a_{i_k}$$

$$\epsilon_k = \Delta \left(\sum_{j=1}^{k} | x_{i_j} | - \sum_{j=k+1}^{n} | x_{i_j} | \right)$$

Proof: By the definition of ϵ_k we have

$$\epsilon_k = \sum_{j=1}^{n} | (a_{i_k} + \Delta)\, x_j + \bar{y} - y_j | - \sum_{j=1}^{n} | a_{i_k} x_j + \bar{y} - y_j |$$

By substituting with the equation $y_j = a_j x_j + \bar{y}$, we get

$$\epsilon_k = \sum_{j=1}^{n} |(a_{i_k} + \Delta - a_j)\, x_j| - \sum_{j=1}^{n} |(a_{i_k} - a_j)\, x_j|$$

$$= \sum_{j=1}^{n} |x_j| \cdot \Big(|a_{i_k} + \Delta - a_j| - |a_{i_k} - a_j| \Big)$$

It is clear that if $a_{i_k} \geqq a_j$, then $(a_{i_k} + \Delta) > a_j$, so that

$$|a_{i_k} + \Delta - a_j| - |a_{i_k} - a_j| = \Delta$$

Likewise if $a_{i_k} < a_j$, then $(a_{i_k} + \Delta) \leqq a_{i_{k+1}} \leqq a_j$, so that

$$|a_{i_k} + \Delta - a_j| - |a_{i_k} - a_j| = -\Delta$$

But by construction $a_{i_k} \geqslant a_j$ for $j = i_1, i_2, \ldots i_k$, and similarly

$a_{i_k} < a_j$ for $j = i_{k+1}, i_{k+2} \ldots i_n$. Therefore

$$\epsilon_k = \sum_{j=1}^{k} |x_{i_j}|\, \Delta + \sum_{j=k+1}^{n} |x_{i_j}| (-\Delta)$$

$$= \Delta \left(\sum_{j=1}^{k} |x_{i_j}| - \sum_{j=k+1}^{n} |x_{i_j}| \right) \qquad \text{Q.E.D.}$$

Lemma 2: If $a_{i_{k+1}} = a_{i_k}$, then $\epsilon_k = 0$. *If* $a_{i_{k+1}} > a_{i_k}$, then ϵ_k is $> 0, = 0,$ or < 0 according to whether $\left(\sum_{j=1}^{k} |x_{i_j}| - MX \right)$ is $> 0, = 0,$ or < 0, respectively.

Proof: If $a_{i_{k+1}} = a_{i_k}$, then $\epsilon_k = E(a_{i_{k+1}}) - E(a_{i_k}) = 0$.

If $a_{i_{k+1}} > a_{i_k}$, we have by the definition of MX that

$$\sum_{j=1}^{k} |x_{i_j}| + \sum_{j=k+1}^{n} |x_{i_j}| = 2MX$$

$$\sum_{j=1}^{k} |x_{i_j}| - MX = MX - \sum_{j=k+1}^{n} |x_{i_j}| = 1/2 \left(\sum_{j=1}^{k} |x_{i_j}| - \sum_{j=k+1}^{n} |x_{i_j}| \right)$$

$$= \frac{\epsilon_k}{2\,\Delta} \quad \text{(by Lemma 1)}$$

This completes the proof because $\Delta > 0$ and thus cannot affect the sign.

Theorem 1: If $Z_{k^*} > MX$, then for all $a \neq a_{i_{k^*}}$, $E(a_{i_{k^*}}) < E(a)$.

Proof: For $a_{i_{k^*}} < a \leqq a_{i_{k^*+1}}$, let a $= a_{i_{k^*}} + \Delta$. Then $\epsilon_{k^*} > 0$, because

$$\sum_{j=1}^{k^*} | x_{i_j} | = Z_{k^*} > MX$$

For $a > a_{i_{k^*+1}}$, the argument holds *a fortiori*, because for all $k \geqslant k^*$

$$\sum_{j=1}^{k} | x_{i_j} | \geqslant Z_{k^*}$$

so that each $\epsilon_k \geqslant 0$. At least one such $\epsilon_k > 0$ because $a \neq a_{i_{k^*}}$. Therefore

$$E(a_{i_{k^*}}) \leqslant E(a_{i_{k^*+1}}) \leqslant \ldots < E(a)$$

For $a < a_{i_{k^*}}$ the same arguments hold in reverse; by the algorithm, for all $k < k^*$

$$\sum_{j=1}^{k} | x_{i_j} | < MX$$

so that each $\epsilon_k \leqslant 0$. At least one such $\epsilon_k < 0$ because $a \neq a_{i_{k^*}}$. Therefore

$$E(a) > \ldots \geqslant E(a_{i_{k^*-1}}) \geqslant E(a_{i_{k^*}})$$

Theorem 2: If $Z_{k^*} = MX$, then for all a, b such that $a_{i_{k^*}} \leqq b \leqq a_{i_{k^*+1}}$ and $a < a_{i_{k^*}}$ or $a > a_{i_{k^*+1}}$, $E(b) = E(a_{i_{k^*}}) < E(a)$.

Proof: If $a_{i_{k^*}} = a_{i_{k^*+1}}$, we have $Z_{k^*+1} > MX$ and $b = a_{i_{k^*}} = a_{i_{k^*+1}}$, so that Theorem 1 can be applied. Therefore we need only consider the case in which $a_{i_{k^*+1}} > a_{i_{k^*}}$.

Let $b = a_{i_{k^*}} + \Delta$. Then $\epsilon_{k^*} = 0$ by Lemma 2, since

$$\sum_{j=1}^{k^*} | x_{i_j} | - MX = Z_{k^*} - MX = 0$$

If $a_{i_{k^*+1}} < a < a_{i_{k^*+2}}$, let $a = a_{i_{k^*+1}} + \Delta$. Then $\epsilon > 0$ by Lemma 2, since

$$\sum_{j=1}^{k^*+1} | x_{i_j} | = Z_{k^*} + | x_{i_{k^*+1}} | > MX$$

For $a > a_{i_{k^*+2}}$ and $a < a_{i_{k^*}}$, the proof of theorem 1 is applicable.