Staff Paper Series

Staff Paper P68-2 October 1968

USING ABSOLUTE DEVIATIONS TO COMPUTE LINES OF BEST FIT

By

J. P. Houck and R. D. Hunt

Department of Agricultural Economics

University of Minnesota
Institute of Agriculture
St. Paul, Minnesota 55108

USING ABSOLUTE DEVIATIONS TO COMPUTE LINES OF BEST FIT

J. P. Houck and R. D. Hunt

October 1968 P 68-2

USING ABSOLUTE DEVIATIONS TO COMPUTE LINES OF BEST FIT

J. P. Houck and R. D. Hunt

There is some renewed interest among agricultural economists and others in the old technique of using minimized absolute deviations in computing lines of best fit. H. B. Jones and J. C. Thompson, in a recent article in Agricultural Economics Research, discussed philosophical and practical questions raised when fitted lines using squared and unsquared deviations are compared [4]. However, agricultural economics literature, (including the Jones-Thompson paper) and elementary statistics and econometric literature seems to contain no concise description of the straightforward and simple procedure for calculating lines of best fit which minimize the sum of absolute deviations from the empirical observations to the fitted line. In some experimental work at the University of Minnesota, we find that lines fitted by minimizing the sum of absolute deviations can be useful in certain circumstances.

Our purpose is to present, without proving, a simple step-by-step method for calculating such lines of best fit for the two variable case. An illustrative example is appended. Those wishing to investigate the historical development and mathematical basis for these methods are directed to the References section, especially [1] [2] [5] [6] [7]. The article by Kaarst is particularly helpful.

It is our view that the minimized sum of absoluted deviations (MSAD) can be quite useful for linear trend fitting. Computing a trend line generally does not involve assumptions about the error term's statistical distribution. Therefore, unless some quadatric loss function is postulated, minimizing absolute deviations has as much theoretical appeal as minimizing squared deviations and rests on a simpler conceptual base. Selection of the MSAD procedure over least squares might hinge on the existence of one or more extreme or "unusual" data points which strongly influence the position of a least squares line. The influence of these extreme points is reduced with the MSAD approach. Moreover, an MSAD line is easy to compute with a desk calculator, perhaps easier than simple least squares.

Consider a set of n observations on Y and X to which a linear function is to be fitted. The fitted function is to have the form

$$Y = a + bX$$

where a and b are parameter estimates.

Suppose that, instead of an ordinary least squares line of best fit, a MSAD line is desired.

Two versions of this line can be computed [5]. The first is the MSAD line constrained to pass through a pre-selected point such as the means, the medians, or the origin. The second is the unconstrained MSAD line. Unlike the ordinary least squares line, the unconstrained MSAD line does not, in general, pass through the point of means.

The Constrained MSAD Line

1. Select the reference point (Y_o, X_o) through which the MSAD line is to pass. For each of the n observations, compute

$$y_i = (Y_i - Y_o) \qquad i = 1,2,3, \ldots\ldots n$$
$$x_i = (X_i - X_o)$$

2. Calculate the n ratios, y_i/x_i. Then rank these ratios in ascending algebraic order, beginning with largest negative or the smallest positive number.

3. Sum the absolute values of x_i; $\sum_n |x_i|$

4. To the value $-\frac{1}{2} \sum_n |x_i|$, add the absolute value of x_i from the ratio having the lowest rank. Then add the absolute value of x_i from the ratio with the next lowest rank. Continue until the algebraic sum changes sign from negative to positive.

5. Note the value of the ratio (y_k/x_k) at which the sign change occurs. This ratio is the slope, b, of the constrained MSAD line. This line will pass through both the pre-selected point and the point at which the ratio (x_k/y_k) is computed. Having calculated b:

$$a = Y_o - bX_o$$

The Unconstrained MSAD Line

Finding the unconstrained MSAD line of best fit involves a simple iterative procedure based on the method described above for the constrained line. The procedure is:

1. Compute a constrained MSAD line, but select one of the data points as the initial reference point (Fewer iterations are needed if the selected point is on or near a freehand line of best fit.)

2. This constrained line will pass through at least one other data point in the sample. Select this point as the next reference point, and compute another MSAD line.

3. This second line will pass through still another data point. Using this point as the reference, re-compute.

4. Continue until the fitted line reflects back through a data point already used. This line is then the unconstrained MSAD line of best fit. By judicious selection of the initial reference point on or near a freehand

trend line, only one or two iterations usually will be needed.

Goodness of Fit

Measuring the goodness of fit of an MSAD line is not as straightforward as with the least squares technique. This is because the total sum of absolute deviations of Y_i about a point, say the mean or median, cannot be partitioned unambiguously into that portion accounted for by the fitted line and that portion not accounted for by the fitted line, as can be done with squared deviations about the sample mean in least squares. However, we suggest the following coefficient as a measure of how well an MSAD line fits the sample observations on Y_i.

$$f = 1 - \sum_n \left| Y_i - \hat{Y}_i \right| \Big/ \sum_n \left| Y_i - \tilde{Y} \right|$$

Where $\hat{Y}_i$ are values of Y along the fitted MSAD line associated with the sample X_i and $\tilde{Y}$ is the median value of Y.

The median is selected as the reference point since, for any given set of numbers, the sum of the absolute deviations is smaller when measured from the median than from any other number [8]. The coefficient, f, can range between zero and +1.0. If there is no systematic association between X and Y in the data, then the minimizing of absolute deviations will yield $\hat{Y}_i = \tilde{Y}$ and f will be zero. On the other hand, if the fit is perfect and $\hat{Y}_i = Y_i$, then f will be +1.0. Intermediate values of f will then indicate how well the MSAD line fits the data relative to a scale of zero to +1.0. As a criterion for judging goodness of fit, the coefficient f probably should not be compared with the least squares r^2 for the same set of data. Its use should be confined to comparisons of MSAD lines of best fit. For example, the coefficient f can be used to compare constrained _vs_ unconstrained MSAD lines or constrained MSAD lines fitted through various preselected points.

Concluding Comments

Tables 1 and 2 contain an illustrative example of the calculation of both the constrained and the unconstrained MSAD line based on a single set of hypothetical data. The f coefficient is computed for both. Note that f is slightly larger with the unconstrained line. Figure 1 illustrates the data and the two computed lines.

Extending these methods beyond two variables is not easy, though it has been done mathematically [3] [7]. But perhaps the step-by-step sequences presented here will be useful to agricultural economists and other researchers who can visualize using MSAD lines of best fit in their work.

Table 1: Calculation of <u>constrained</u> MSAD line of best fit through the data means.

X	Y	x $(X-\bar{X})$	y $(Y-\bar{Y})$	y/x	Ranking of y/x
1	10	-5	5.18	-1.036	4
2	7	-4	2.18	- .545	7
3	11	-3	6.18	-2.060	2
4	4	-2	- .82	- .410	9
5	6	-1	1.18	-1.180	3
6	3	0	-1.82	---	11
7	1	1	-3.82	-3.820	1
8	5	2	.18	.090	10
9	2	3	-2.82	- .940	5
10	3	4	-1.82	- .455	8
11	1	5	-3.82	- .764	6

$\bar{X} = 6.00 \qquad \bar{Y} = 4.82 \qquad Y = 4.0$

$$-\tfrac{1}{2}\sum |x_i| = -15$$

(here the subscript on x_i denotes ranking)

$$|x_1| = +1 \quad \rightarrow \quad -14$$

$$|x_2| = +3 \quad \rightarrow \quad -11$$

$$|x_3| = +1 \quad \rightarrow \quad -10$$

$$|x_4| = +5 \quad \rightarrow \quad -5$$

$$|x_5| = +3 \quad \rightarrow \quad -2$$

$$|x_6| = +5 \quad \rightarrow \quad +3$$

Note sign change at rank 6

Then:

$$b = (y_6 / x_6) = -.764$$

$$a = \bar{Y} - b\,\bar{X} = 9.402$$

$$f = 1 - (17.238 / 29.000) = .406$$

Table 2: Calculation of <u>unconstrained</u> MSAD line of best fit selecting X = 5, Y = 6 as the first reference point.

(First iteration)

X	Y	x (X-5)	y (Y-6)	y/x	Ranking
1	10	-4	4	-1.000	6
2	7	-3	1	- .333	10
3	11	-2	5	-2.500	3
4	4	-1	-2	2.000	4
5	6	0	0	---	11
6	3	1	-3	-3.000	1
7	1	2	-5	-2.500	2
8	5	3	-1	- .333	9
9	2	4	-4	-1.000	5
10	3	5	-3	- .600	8
11	1	6	-5	- .833	7

$-\frac{1}{2}\sum|x_i| = -15$

$|x_1| = \underline{+1}$ (subscript denotes ranking)

-14

$|x_2| = \underline{+2}$

-12

$|x_3| = \underline{+2}$

-10

$|x_4| = \underline{+1}$

-9

$|x_5| = \underline{+4}$

-5

$|x_6| = \underline{+4}$

-1 ← Note sign change at rank 7

$|x_7| = \underline{+6}$

$+5$ ←

Then for the first iteration: ($b = y_7 / x_7 = -.833.$)
This line passes through X = 11, Y = 1. Use this point as reference for next iteration

(continued next page)

Table 2 (continued)

(Second iteration)

X	Y	x (X-11)	y (Y-1)	y/x	Ranking
1	10	-10	9	-.900	4
2	7	- 9	6	-.667	6
3	11	- 8	10	-1.250	3
4	4	- 7	3	-.429	8
5	6	- 6	5	-.833	5
6	3	- 5	2	-.400	9
7	1	- 4	0	0	10
8	5	- 3	4	-1.333	2
9	2	- 2	1	-.500	7
10	3	- 1	2	-2.000	1
11	1	0	0	---	11

$-\frac{1}{2}\sum|x_i| = -27.5$

$|x_1| = +1.0$ (Subscript denotes ranking)

-26.5

$|x_2| = +3.0$

-23.5

$|x_3| = +8.0$

-15.5

$|x_4| = +10.0$

-5.5

$|x_5| = +6.0$

$+0.5$

Note sign change at rank 5

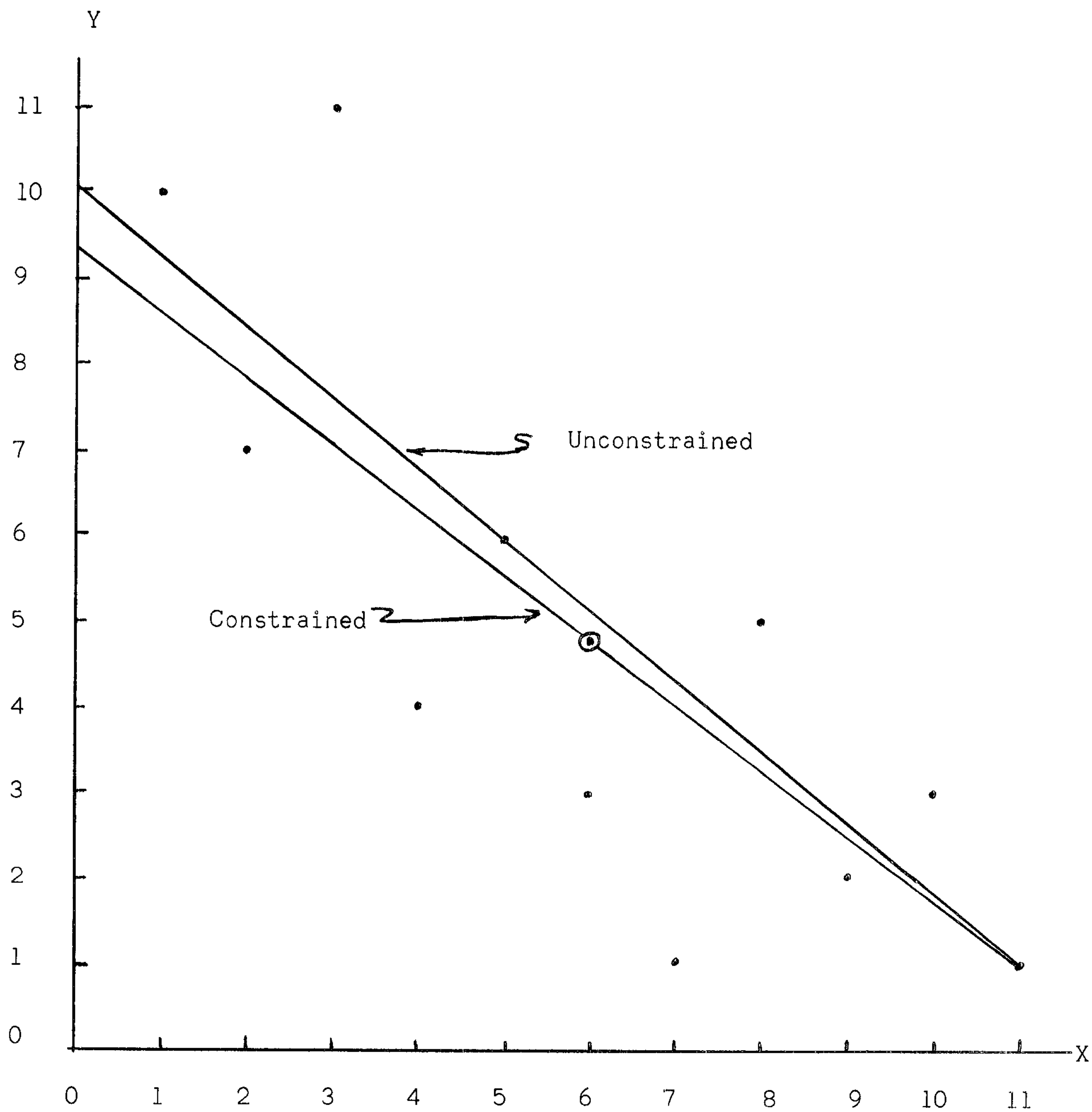
Then for the second iteration: $(b = y_5 / x_5 = -.833)$
This is the same as the result of the first iteration. The fitted line reflects back through X = 5, Y = 6 and is the final unconstrained solution.

$$b = -.833$$

$$a = 6 + (.833)(5) = 10.165$$

$$f = 1 - (17.167 / 29.000) = .408$$

Figure 1: Fitted MSAD lines and sample data for illustrative example



References

1. Ashar, V. G. and T. D. Wallace., "A Sampling Study of Minimum Absolute Deviations Estimators," Operations Research, No. 5, Vol. 1, September - October 1963, pp. 747-758.

2. Edgeworth, F. Y., "On a New Method of Reducing Observations Relating to Secular Quantities," Philosophical Magazine, Series 5, Vol. 25, No. 154, March 1888, pp. 184-191.

3. Fisher, W. D., "A Note on Curve Fitting with Minimum Deviations by Linear Programming," Journal of American Statistical Association, Vol. 56, 1961, pp. 359-362.

4. Jones, Harold B., and Jack C. Thompson., "Squared versus Unsquared Deviations for Lines of Best Fit," Agricultural Economics Research, Vol. 20, No. 2, April 1968, pp. 64-69.

5. Kaarst, Otto J., "Linear Curve Fitting Using Least Deviations," Journal of the American Statistical Association, No. 281, Vol. 53, March 1958, pp. 118-132.

6. Rhodes, E. C., "Reducing Observations by The Method of Minimum Deviations," Philosophical Magazine, Series 7, Vol. 9, No. 60, May 1930.

7. Singleton, R. R., "A Method for Minimizing The Sum of Absolute Values of Deviations," Annals of Mathematical Statistics, Vol. 11, 1940, pp. 301-305.

8. Waugh, A. E., Elements of Statistical Method, Third edition, (McGraw-Hill: 1952), p. 114.