

Technical note

Ambiguities inherent in sums-of-squares-based error statistics

Cort J. Willmott^{a,*}, Kenji Matsuura^a, Scott M. Robeson^b^a Center for Climatic Research, Department of Geography, University of Delaware, Newark, DE 19716, USA^b Department of Geography, Indiana University, Bloomington, IN 47405, USA

ARTICLE INFO

Article history:

Received 25 August 2008

Accepted 8 October 2008

Keywords:

Error statistics

Standard deviation

Standard error

Mean-absolute deviation

ABSTRACT

Commonly used sums-of-squares-based error or deviation statistics—like the standard deviation, the standard error, the coefficient of variation, and the root-mean-square error—often are misleading indicators of average error or variability. Sums-of-squares-based statistics are functions of at least two dissimilar patterns that occur within data. Both the mean of a set of error or deviation magnitudes (the average of their absolute values) and their variability influence the value of a sum-of-squares-based error measure, which confounds clear assessment of its meaning. Interpretation problems arise, according to Paul Mielke, because sums-of-squares-based statistics do not satisfy the triangle inequality. We illustrate the difficulties in interpreting and comparing these statistics using hypothetical data, and recommend the use of alternate statistics that are based on sums of error or deviation magnitudes.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Sums-of-squares-based error statistics—such as the standard error, the root-mean-square error and the coefficient of variation—are often considered to be unambiguous indicators of average deviation, average error and average variability. Perusal of *Atmospheric Environment* or virtually any other applied-science journal reveals the widespread use of these and related statistics (e.g., see Case et al., 2008 and Krudysz et al., 2008). Little-known and undesirable properties of these statistics, however, foster their frequent misuse and misinterpretation. Difficulties typically arise because it is assumed, usually tacitly, that a sum-of-squares-based measure can faithfully represent average error or average deviation or average variability. It cannot; in fact, there is no clear-cut scientific interpretation of the values of these statistics, because sums-of-squares-based measures vary in response to both central tendency and variability within a set of error or deviation magnitudes.

Our goals within this note are to point out the ambiguities inherent within sums-of-squares-based error or deviation statistics, and to illustrate problems that can arise in their interpretation and comparison. We also recommend the use of alternate, absolute-error- or absolute-deviation-based measures. Our critique here is restricted to minimized or fit sums-of-squares measures, such as the standard deviation or standard error, because they are among the most commonly used sums-of-squares-based measures

and because sums-of-squares-based measures that are unconstrained by fit (most notably the root-mean-squared-error, RMSE) have been discussed elsewhere (Pontius et al., 2008; Willmott and Matsuura, 2005; 2006). Our hope is that—since the air-quality modeling community is at the forefront of the atmospheric sciences in assessing accuracy and precision statistically—the readers of *Atmospheric Environment* will find our observations and recommendations of value.

2. Background and context

Our assessment of fit sums-of-squares-based error or deviation measures uses the standard error (SE) to help illustrate the issues, since the SE is both well known and representative of many related statistics. Standard error can be written as

$$SE = \left\{ [d(n)]^{-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right\}^{1/2}, \quad (1)$$

where n is the number of errors or deviations, $d(n)$ is a degrees-of-freedom function [$d(n) < n$], $\hat{y}_i = f(\mathbf{X})$, and \mathbf{X} is a set of one or more independent variables. The units of the SE are the units of y . Note that, when $\hat{y}_i = \bar{y}$, the SE is an estimate of the standard deviation; otherwise, it represents the minimized variability (the minimized sum of the squared deviations) around a “best-fit” function (Draper and Smith, 1998). Indeed, minimization of the sum of squares or “least-squares” is the most popular way of fitting a function to data. It is not surprising then that minimized sums-of-squares-based measures (especially the SE and its very close relative the root-mean-square error, RMSE) tend to be reported and

* Corresponding author.

E-mail address: willmott@udel.edu (C.J. Willmott).

then interpreted as measures of average error, deviation or inaccuracy. It is worth noting that, with least-squares-fit functions, the sum of the errors or deviations is zero. For the purpose of simplifying our discussion below, we let $d(n) = n$.

An alternate estimate or representation of average error or average deviation can be obtained from the absolute values (magnitudes) of the errors or deviations. This measure is called the mean-absolute deviation (MAD) or occasionally the mean deviation (MD), although the MD is a less precise designation and easily confused with the average of the actual (signed) deviations about the mean which is always zero. Within this note, then, we refer to the MAD—the average of the magnitudes of the errors or deviations—which can be written as

$$\text{MAD} = n^{-1} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (2)$$

As with the SE, the units of MAD are the units of y . While the MAD is conceptually straightforward, its minimization during fitting is more cumbersome (an iterative solution or additional constraints are required) than is the analytically based minimization of a sum-of-squares-based measure, such as SE. Recall that the SE is simply a scaled version of the minimized sum-of-squared errors or deviations associated with a least-squares fit (Draper and Smith, 1998). Consider also that, like the standard deviation, the SE is reported by most statistical software. Our sense is that these are among the primary reasons why the SE is widely reported and interpreted, often as a measure of average error or average deviation, and MAD is not.

Our points are illustrated below by comparing the responses of the SE and the MAD to varying patterns that can occur within data. Hypothetical data are used, in order to isolate and illuminate factors to which the SE and the MAD are sensitive. It is useful to remember that the only difference between the SE and the MAD is in the way

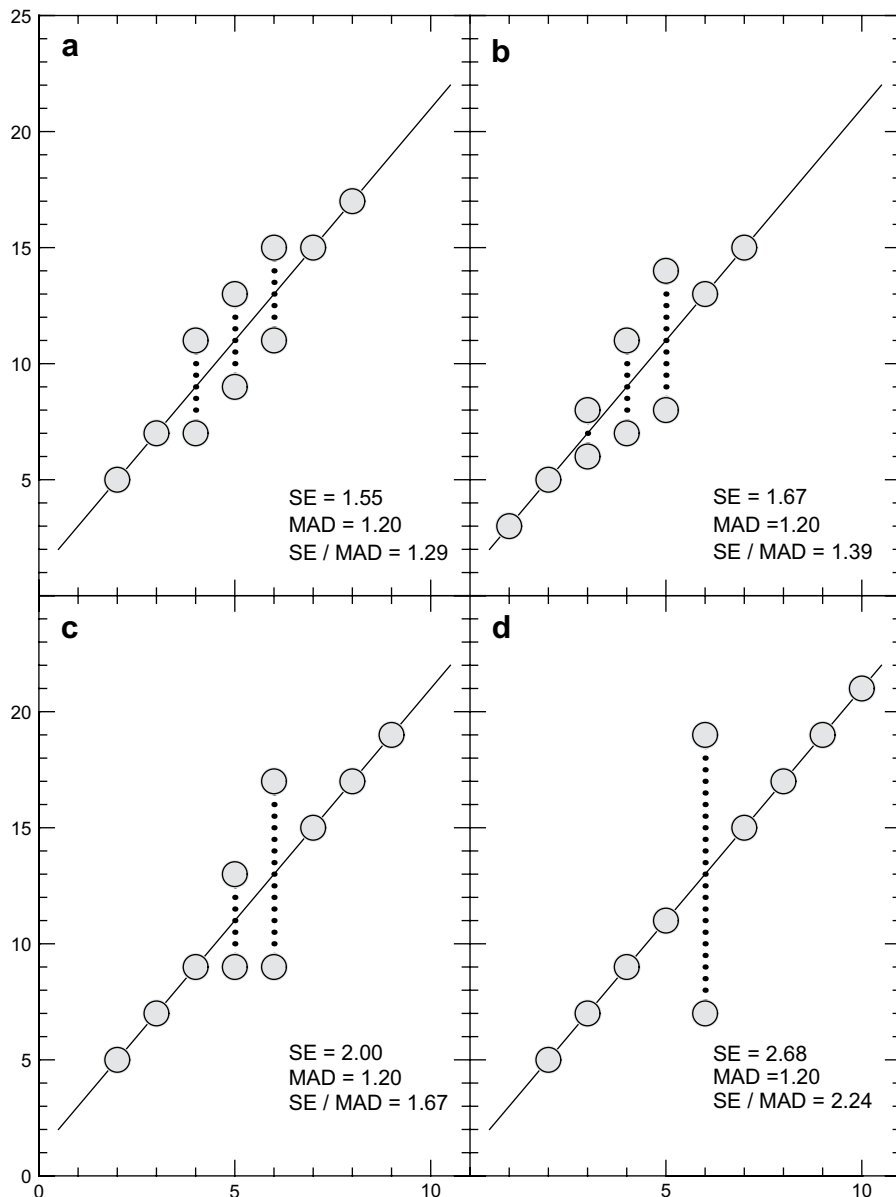


Fig. 1. Four least-squares-fit regression lines, each drawn through its corresponding set of ten pairwise hypothetical observations. The vertical axes represent y . Each of the four hypothetical sets of errors or deviations (from the regression line) has the same average error-magnitude (MAD); however, the variability within each of the four sets of errors or deviations increases from Fig. 1a–d, and the SE increases correspondingly as does the ratio of SE to MAD.

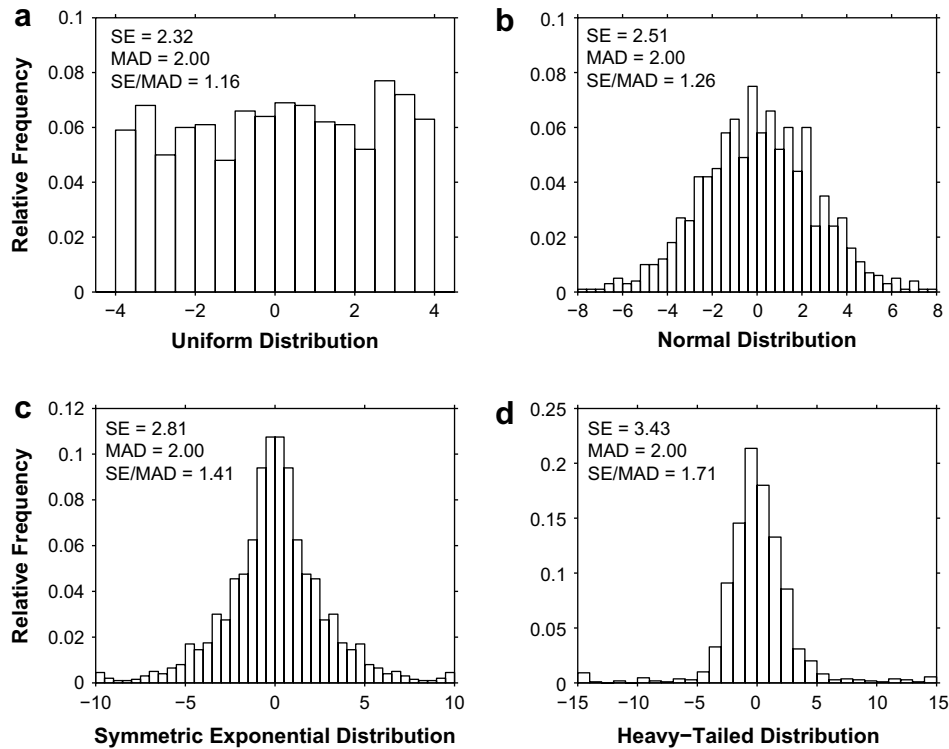


Fig. 2. Standard error (SE), mean-absolute deviation (MAD), and their ratio (SE/MAD) for four different distributions of randomly generated values: (a) uniform distribution, (b) normal distribution, (c) symmetric exponential distribution, and (d) a heavy-tailed distribution derived from a combination of two different normal distributions. Note that the SE to the MAD ratios differ slightly from the analytical solutions, since the individual error or deviation values are simulated.

in which the signs of the individual errors or deviations are removed, although it also should be kept in mind that, in fact, both the SE and the MAD are entirely based on the error or deviation magnitudes. This can be seen clearly when the equation for the SE is rewritten as $SE = [n^{-1} \sum_{i=1}^n |y_i - \hat{y}_i|^2]^{1/2}$ and compared to Equation (2).

3. Ambiguities within the standard error (SE)

Our critique of the SE—and, by extension, other statistics that have a sum-of-squared deviations embedded within them—uses the MAD as a benchmark for two related reasons. It unambiguously represents the arithmetic average of the magnitudes of the errors, and it satisfies the triangle inequality of a distance metric¹ (Mielke and Berry, 2001). Four relatively simple, hypothetical examples illustrate how values of the SE can be inconsistently related to the MAD (Fig. 1). It is apparent that, even though the MAD is stationary (from Fig. 1a–d), the SE is increasing because the variability among the error magnitudes is increasing. It is even possible for SE to be increasing at the same time that the MAD is decreasing.

Conjecture based on Fig. 1 suggests that the lower limit of the SE is the MAD—which occurs when all of the errors have the same magnitude—and the upper limit is $(n/2)^{1/2} \times MAD$, which is reached when all of the error or deviation is contained within two countervailing (same magnitude, opposite signs) errors or

deviations. Analogous proof of this, for cases within which the sum of the squared errors or deviations is not minimized, is contained within Willmott and Matsuura (2005, 2006). With the upper limit of the SE increasing with $(n/2)^{1/2}$, while the lower limit is fixed at MAD, it also is likely that the SE will increase with $(n/2)^{1/2}$ or, when spatial fields are being compared, with the square root of half the area covered by the grid (Willmott and Matsuura, 2006). All of this indicates that the SE has no consistent relationship with the average of the error or deviation magnitudes, other than being greater than or equal to the MAD. It also indicates that the SE has no consistent relationship with the variance within a set of error or deviation magnitudes. When computed from “real” data, however, it is not uncommon for increases and decreases in corresponding values of the SE and the MAD to be correlated (cf, Willmott and Matsuura, 2005), because of the partial dependence of the SE on the MAD.

Practical implications of the SE’s dependence on the variability within the distribution of error magnitudes, MAD and $(n/2)^{1/2}$ are worth pointing out. When the SE is calculated from a set of errors and reported in the literature, but the MAD is not, it is impossible to discern to what extent the SE reflects central tendency (MAD) and to what extent it represents variability within the distribution of error or deviation magnitudes. When two or more SEs are obtained from different sets of errors and reported, but the corresponding MADs and n s are not, meaningful comparisons of the SEs are confounded, because each SE may be a function of a different n (or spatial or temporal domain), as well as of a different error-magnitude variability and MAD.

An assessment of the dependence of SE values on the probability distribution (variability) of errors or deviations provides additional insight. Values of the SE can be confounded by this dependence, as we illustrate for four well-known relative frequency distributions (Fig. 2). For a set of errors or deviations, forming a ratio of the SE to MAD is a useful way to examine the influence of the error or

¹ A “metric” is a distance function $[d(a,c)]$ that satisfies four properties: $d(a,a) = 0$; $d(a,c) \geq 0$; $d(a,c) = d(c,a)$; and $d(a,c) \leq d(a,b) + d(b,c)$. Both $d(a,c)$ and its square satisfy the first three properties, but its square does not satisfy the fourth property, the “triangle inequality.” Imagine, for example, that $d(a,c) = 4$, $d(a,b) = 2$ and $d(b,c) = 3$. It is clear that $4 \leq 2 + 3$; whereas, $4^2 \not\leq 2^2 + 3^2$. This means that the relative influence of each squared error on the sum of the squared errors often is counterintuitive, which undermines meaningful scientific interpretation of the sum of the squared errors.

deviation distribution on the SE. When the probability distribution of the errors is known, the ratio of the SE to the MAD can be calculated explicitly and, for some distributions, an analytical solution of the ratio can be found. It can be shown analytically, for example, that—for a set of errors that follow a normal distribution—the ratio of the SE to the MAD is $\sqrt{\pi/2}$ (≈ 1.25). Similarly, the analytical solution for a uniform distribution of errors produces a SE-to-MAD ratio of $\sqrt{4/3}$ (≈ 1.15). As the distribution of errors becomes increasingly heavy-tailed, the ratio of the SE to the MAD becomes increasingly large (Fig. 2).

Inconsistencies between the magnitudes of the SE and the MAD arise from the inefficiency of the SE (relative to the MAD) when the underlying distribution of errors is increasingly non-normal, a problem that has been recognized for quite some time and was very clearly illustrated by Tukey (1960). Indeed, the inherent limitations of sums-of-squares approaches provide the motivation for much of the field of robust statistics (Huber, 1981). It is somewhat surprising, therefore, that the undesirable sensitivity of the SE to even slight departures of an error or deviation distribution from normal is not more widely known and appreciated. The relative advantages of the MAD should be especially valued in the atmospheric and environmental sciences, where outliers and deviations from normality are commonplace.

4. Summary and recommendations

Sums-of-squares-based error or deviation statistics are inappropriate measures of the average or typical error or deviation because their values are often counterintuitive (Mielke and Berry, 2001) or ambiguous (Willmott and Matsuura, 2005, 2006). The source of the problem is that, unlike the actual error magnitudes, each squared error may no longer be meaningfully comparable to other squared errors that populate the set of squared errors; that is, the triangle inequality may not be satisfied (Mielke and Berry, 2001). Interpretational difficulties ensue because the SE (or related sum-of-squares-based measure) depends not only on the average of the error or deviation magnitudes (the MAD), but also on the variability within the set of error or deviation magnitudes. The SE additionally may be inflated by (pulled in the direction of) its upper limit, which is a partial function of $n/2$. Consider that a particular fit model may be erroneously identified as “superior” simply because the distribution of error or deviation magnitudes is more homogeneous or because a smaller sample size was used. For these reasons, there is no clear-cut scientific interpretation of the SE or related sum-of-squares-based statistic.

Our analysis of the SE indicates that the SE and related measures should not be reported and interpreted routinely in the literature. It

also suggests that previous evaluations and comparisons of averages of error or deviation magnitudes, which were based primarily on the SE, are questionable and should be reconsidered. Unlike the SE, the MAD is a natural and unambiguous measure of the average of the error or deviation magnitudes and, therefore, should always be computed, reported and interpreted. Within particular circumstances (e.g., $\hat{y} = \bar{y}$), it also may be useful to compute, report and interpret related magnitude-based measures, such as the average variability of error or deviation magnitudes about the MAD ($n^{-1} \sum_{i=1}^n ||y_i - \hat{y}_i| - \text{MAD}|$) or, analogous to the coefficient of variation, the size of the average error or deviation magnitude (MAD) relative to the dependent-variable mean (MAD/\bar{y}). While our analysis is confined to simple measures of the average of error or deviation magnitudes (the SE and the MAD), the problems that we see within the SE also are present within more complex statistics that have one or more sums-of-squares embedded within them.

Acknowledgments

Much of this work was made possible by NASA Grant NNG06GB54G to the Institute of Global Environment and Society (IGES) and we are most grateful for this support.

References

- Case, M.W., Williams, R., Yeatts, K., Chen, F.-L., Scott, J., Svendsen, E., Devlin, R.B., 2008. Evaluation of a direct personal coarse particulate matter monitor. *Atmospheric Environment* 42, 4446–4452.
- Draper, N.R., Smith, H., 1998. *Applied Regression Analysis*, third ed. Wiley, New York, NY.
- Huber, P., 1981. *Robust Statistics*. John Wiley and Sons, New York, NY.
- Krudysz, M.A., Froines, J.R., Fine, P.M., Sioutas, C., 2008. Intra-community spatial variation of size-fractionated PM mass, OC, EC, and trace elements in the Long Beach, CA area. *Atmospheric Environment* 42, 5374–5389.
- Mielke, Jr., P.W., Berry, K.J., 2001. *Permutation Methods: a Distance Function Approach*. Springer-Verlag, New York, NY.
- Pontius, Jr., R.G., Thonteh, O., Chen, H., 2008. Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics* 15 (2), 111–142.
- Tukey, J., 1960. A survey of sampling from contaminated distributions. In: Olkin, I., Ghurye, S., Hoeffding, W., Madow, W., Mann, H. (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Stanford, CA, pp. 448–485.
- Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30, 79–82.
- Willmott, C.J., Matsuura, K., 2006. On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science* 20 (1), 89–102.