Atmospheric Environment 43 (2009) 749–752

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

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.

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

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

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).

1352-2310/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2008.10.005

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

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

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

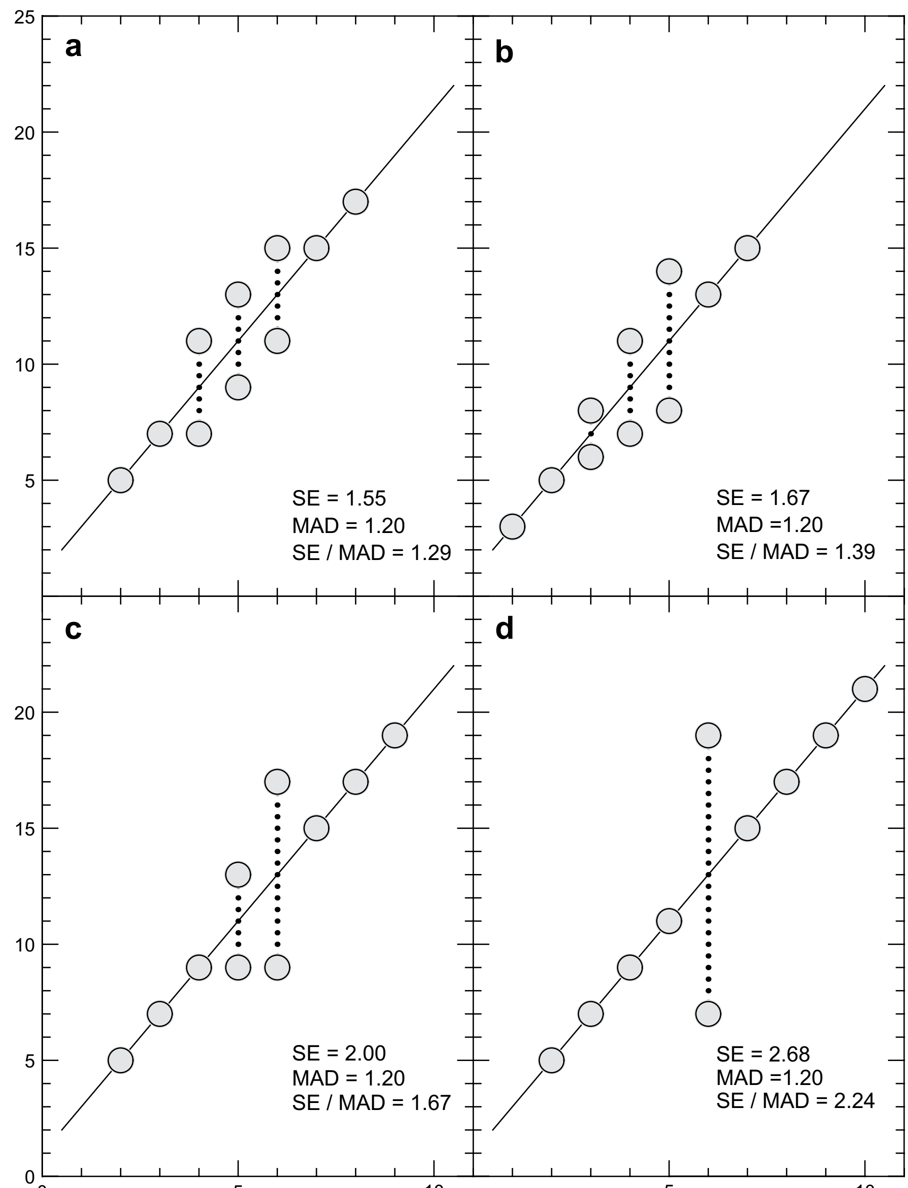

Fig. 1. Four least-squares-fit regression lines, each drawn through its corresponding set of ten pairwise hypothetical observations. The vertical axes represent y. Each of the four hypothetical sets of errors or deviations (from the regression line) has the same average error-magnitude (MAD); however, 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.