# On Fitting a Straight Line to Data when the "Noise" in Both Variables Is Unknown*

ALLAN J. CLARKE AND STEPHEN VAN GORDER

*Department of Earth, Ocean and Atmospheric Science, The Florida State University, Tallahassee, Florida*

(Manuscript received 30 March 2012, in final form 19 June 2012)

ABSTRACT

In meteorology and oceanography, and other fields, it is often necessary to fit a straight line to some points and estimate its slope. If both variables corresponding to the points are noisy, the slope as estimated by the ordinary least squares regression coefficient is biased low; that is, for a large enough sample, it *always* underestimates the true regression coefficient between the variables. In the common situation when the relative size of the noise in the variables is unknown, an appropriate regression coefficient is plus or minus the ratio of the standard deviations of the variables, the sign being determined by the sign of the correlation coefficient. For this case of unknown noise, the authors here obtain the probability density function (pdf) for the true regression coefficient divided by the appropriate regression coefficient just mentioned. For the case when the number of data is very large, a simple analytical expression for this pdf is obtained; for a finite number of data points the relevant pdfs are obtained numerically. The pdfs enable the authors to provide tables for confidence intervals for the true regression coefficient. Using these tables, the end result of this analysis is a simple practical way to estimate the true regression coefficient between two variables given their standard deviations, the sample correlation, and the number of independent data.

## 1. Introduction

In meteorology and oceanography, and other fields, it is often necessary to check whether two quantities $x$ and $y$ are linearly related. Usually this involves the calculation of a correlation coefficient and a regression coefficient. Both coefficients are useful; the sample correlation coefficient $\hat{r}$ provides a test of the linearity of the relationship and the regression coefficient provides the proportionality constant between the two variables. By its sign and size, the regression coefficient often informs us whether the relationship makes sense physically.

In practice, "noise" reduces the correlation coefficient and affects the accuracy of the regression coefficient. Noise can be due to measurement error or to different physical processes in $x$ and $y$ that affect the linearizing process common to both. While measurement error can sometimes be estimated for each variable, error due to possible physical influences in the real data is often not known. So there is a large class of regression problems in meteorology, oceanography, and other fields in which the signal-to-noise ratio in both variables is unknown.

When there is noise in both variables, the regression coefficient between the variables is underestimated even if an infinite sample of data were available. To illustrate this well-known result and to introduce our notation, consider the problem of finding, from a sample of $M$ points, the true linear regression coefficient between two random variables $X$ and $Y$ that are both subject to random noise. Mathematically, we want to find the regression coefficient $\alpha$ in the linear relationship

$$Y = \alpha X + a, \tag{1}$$

given the observations

$$x_i = X_i + \varepsilon_i \quad i = 1, \ldots, M, \tag{2}$$

$$y_i = Y_i + \delta_i = \alpha X_i + a + \delta_i \quad i = 1, \ldots, M. \tag{3}$$

In (2) and (3), the error or noise terms $\varepsilon$ and $\delta$ are taken to be normally distributed zero-mean random variables that are independent of each other and $X$.

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/2012JTECH00067.s1.

*Corresponding author address:* A. J. Clarke, Department of Earth, Ocean and Atmospheric Science, The Florida State University, Tallahassee, FL 32306-4320.
E-mail: aclarke@fsu.edu

DOI: 10.1175/JTECH-D-12-00067.1

One estimate for $\alpha$ is the ordinary least squares estimate

$$\hat{\alpha}_{\text{OLS}} = s_{xy}/(s_x)^2, \tag{4}$$

where

$$s_{xy} = \frac{1}{M}\sum_{i=1}^{M}(x_i - \overline{x})(y_i - \overline{y}) \tag{5}$$

is the sample covariance of $x$ and $y$,

$$s_x = \left[\frac{1}{M}\sum_{i=1}^{M}(x_i - \overline{x})^2\right]^{1/2} \tag{6}$$

is the sample standard deviation of $x$, and

$$\overline{x} = \frac{1}{M}\sum_{i=1}^{M}x_i, \quad \overline{y} = \frac{1}{M}\sum_{i=1}^{M}y_i \tag{7}$$

are the sample means of $x$ and $y$, respectively. Since $X$, $\varepsilon$, and $\delta$ are all independent of each other, as $M \rightarrow \infty$,

$$s_x^2 \rightarrow \sigma_X^2 + \sigma_\varepsilon^2 \quad \text{and} \quad s_{xy} \rightarrow \alpha\sigma_X^2, \tag{8}$$

where $\sigma_X$ and $\sigma_\varepsilon$ are the (true) standard deviations of $X$ and $\varepsilon$. Thus it follows from (4) that as $M \rightarrow \infty$,

$$\hat{\alpha}_{\text{OLS}} \rightarrow \alpha_{\text{OLS}} = \alpha\sigma_X^2/(\sigma_X^2 + \sigma_\varepsilon^2) = \alpha/(1 + n_x), \tag{9}$$

where the noise-to-signal ratio $n_x$ is defined as

$$n_x = \sigma_\varepsilon^2/\sigma_X^2. \tag{10}$$

Therefore, whenever $x$ has nonzero noise and hence $n_x > 0$, (9) implies the well-known result that $\hat{\alpha}_{\text{OLS}}$ underestimates the true regression coefficient $\alpha$ even when the number of observations $M \rightarrow \infty$. If $x$ has zero noise, then $n_x = 0$ and an unbiased estimate $\alpha$ for the regression of $y$ is available; conversely, if $y$ has zero noise ($\delta = 0$), then a regression of $x$ on $y$ yields an unbiased estimate for $\alpha^{-1}$. Confidence intervals in either case can then be determined in the finite case using a Student's distribution with $M - 2$ degrees of freedom. However, when both $x$ and $y$ have nonzero noise, an unbiased estimate of $\alpha$ or $\alpha^{-1}$ is not available and confidence intervals for the true regression coefficient are unknown.

Many methods (Ricker 1973; Jolicoeur 1975; Riggs et al. 1978; McArdle 1988, 2003; Frost and Thompson 2000) have been devised to obtain a best estimate for $\alpha$ when both variables are noisy. When the relative size of the noise for each variable is known, an explicit estimate for $\alpha$ can be given (Kendall and Stuart 1973). However, as discussed above, often this ratio is unknown and then the way forward is murky and ''definitive recommendations are difficult to make'' (Sokal and Rohlf 1995).

In this paper, we overcome this difficulty by noting that when the relative size of the noise for each variable is unknown, we have no basis for choosing between the variables. Consequently, we must assume that the noise in each variable is equally likely subject to the constraint that the sample correlation coefficient for the given set of data is known. This equal likelihood noise assumption, subject to the known sample correlation coefficient, enables us to determine explicitly the confidence intervals in the $M = \infty$ case analytically and in the finite $M$ case numerically.

The rest of the paper is organized as follows: in the next section we discuss an unbiased estimate for the true regression coefficient when nothing is known about the noise in each variable. Then, in section 3, we examine the $M = \infty$ case and determine a probability density function for the true regression coefficient analytically. Confidence intervals for the finite $M$ case are found numerically in section 4, and an example of their use is then given in section 5. A final section 6 contains some concluding remarks.

## 2. An unbiased estimate for the true regression coefficient

Consider the problem of estimating the true regression coefficient between $x$ and $y$ when nothing is known or assumed about the noise in $x$ and $y$ except for the constraint provided by the sample correlation coefficient. One estimate for the true regression coefficient $\alpha$ of $y$ on $x$ is $\hat{\alpha}_{\text{est}} = f_1\hat{\alpha}_{\text{OLS}}$, where the nondimensional factor $f_1$ is used to correct the bias in $\hat{\alpha}_{\text{OLS}}$ discussed in the introduction. We could equally as well have regressed $x$ on $y$ to obtain $\hat{\beta}_{\text{OLS}} = s_{xy}/(s_y)^2$, where $s_y$ is the sample standard deviation of $y$ defined analogously to $s_x$ in (6). The bias in $\hat{\beta}_{\text{OLS}}$ can similarly be corrected by writing $\hat{\beta}_{\text{est}} = f_2\hat{\beta}_{\text{OLS}}$ using another nondimensional factor $f_2$. But note that there is no basis for distinguishing $f_1$ and $f_2$. In both regression cases there are the same number of points $M$ and the same sample correlation coefficient. Also in both cases we are carrying out the same procedure, namely, using a factor to scale the ordinary least squares regression coefficient so that it gives an unbiased estimate of the true regression coefficient. Therefore, the factors $f_1$ and $f_2$ should be the same and we write $f_1 = f_2 = f$.

Since $\hat{\alpha}_{\text{est}} = f\hat{\alpha}_{\text{OLS}}$ is our estimate for the true regression coefficient $\alpha$ and $\hat{\beta}_{\text{est}} = f\hat{\beta}_{\text{OLS}}$ is our estimate for $\alpha^{-1}$, and we have no way of distinguishing these estimates, we must have $\hat{\beta}_{\text{est}} = (\hat{\alpha}_{\text{est}})^{-1}$ and hence

$$1 = f^2 \hat{\alpha}_{\rm OLS} \hat{\beta}_{\rm OLS} = f^2 (s_{xy})^2/(s_x)^2(s_y)^2, \qquad (11)$$

that is, $f^{-1} = |\hat{r}|$, where $\hat{r}$ is the sample correlation coefficient $s_{xy}/s_x s_y$. Thus an appropriate $\hat{\alpha}_{\rm est} = \hat{\alpha}_{\rm OLS}/|\hat{r}|$, and from (4) and the formula for $\hat{r}$ we have

$$\hat{\alpha}_{\rm est} = \hat{\alpha}_{\rm OLS}/|\hat{r}| = {\rm sgn}(\hat{r}) s_y/s_x, \qquad (12)$$

where sgn ($\hat{r}$) is +1 if $\hat{r}$ is positive and −1 if $\hat{r}$ is negative. From (12), $|\hat{\alpha}_{\rm est}|$ is simply the ratio of the standard deviations of $y$ and $x$. In the oceanography literature, $\hat{\alpha}_{\rm est}$ has been called the "neutral" regression coefficient (Garrett and Petrie 1981), but in other fields, for example, the life sciences, fisheries, and statistics (Ricker 1973; Riggs et al. 1978; Sprent and Dolby 1980; Barker et al. 1988; Sokal and Rohlf 1995), it is known as the geometric mean regression coefficient. The latter nomenclature follows because $|\hat{\alpha}_{\rm est}|$ is the geometric mean of the estimates $\hat{\alpha}_{\rm OLS}$ and $\hat{\beta}_{\rm OLS}^{-1}$. Henceforth, we write

$$\hat{\alpha}_{\rm est} = \hat{\alpha}_{\rm GMR} = \hat{\alpha}_{\rm OLS}/|\hat{r}| = {\rm sgn}(\hat{r}) s_y/s_x. \qquad (13)$$

Note that $\hat{\alpha}_{\rm GMR}$ is also the regression coefficient obtained when $x$ is normalized by $s_x$, $y$ is normalized by $s_y$, and the *perpendicular* distance from the regression line is minimized rather than the vertical distance as in an ordinary least squares fit. A principal component analysis of the normalized variables also has its principal axis as the $\hat{\alpha}_{\rm GMR}$ regression line. There has been considerable controversy and confusion over the use of this regression coefficient (Jolicoeur 1975; Ricker 1975; Sprent and Dolby 1980; Barker et al. 1988; Emery and Thomson 2001; McArdle 2003). Emery and Thomson (2001) comment that the geometric mean coefficient, "though appealing, rests on shaky statistical ground and its use remains controversial." As suggested by (12), one way to think of $\hat{\alpha}_{\rm GMR}$ is as an appropriate estimate for $\alpha$ when signal-to-noise ratios are not known (see also Barker et al. 1988).

## 3. The probability density function for $\alpha/\alpha_{\rm GMR}$ for the limiting $M = \infty$ case

Given $M$ data pairs, we can calculate both a correlation coefficient $\hat{r}$ and the regression coefficient $\hat{\alpha}_{\rm GMR}$ above using the sign of $\hat{r}$ and the ratio of the standard deviations. In what follows, we analyze a large data point limiting case using $M = \infty$ values $r$ and $\alpha_{\rm GMR}$.

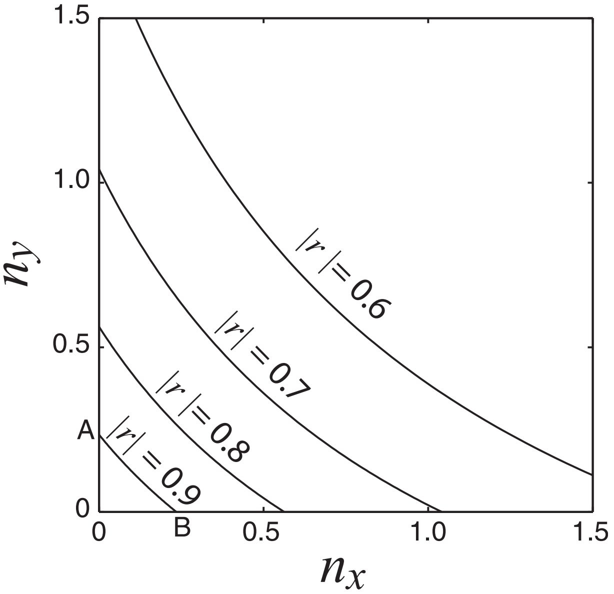

FIG. 1. The hyperbolic curve segments defined by $r^2 = (1 + n_x)^{-1}(1 + n_y)^{-1}$ and $n_x, n_y \geq 0$ for $|r|$ in increments of 0.1. The points $A$ and $B$ are shown when $r$ = 0.9; in general the curves intersect the $n_y$ axis at $A[n_x = 0, n_y = (r)^{-2} - 1]$ and the $n_x$ axis at $B[n_x = (r)^{-2} - 1, n_y = 0]$.

As $M \rightarrow \infty$, the sample coefficient

$$\hat{r} = s_{xy}/s_x s_y \qquad (14)$$

approaches

$$r = \alpha\sigma_X^2/[(\sigma_X^2 + \sigma_\varepsilon^2)^{1/2}(\alpha^2\sigma_X^2 + \sigma_\delta^2)^{1/2}], \qquad (15)$$

where $\sigma_\delta$ is the true standard deviation of $\delta$. By dividing both numerator and denominator by $|\alpha|\sigma_X^2$, (15) may also be written in the form

$$r = (\alpha/|\alpha|)(1 + n_x)^{-1/2}(1 + n_y)^{-1/2}, \qquad (16)$$

where

$$n_y = \sigma_\delta^2/[\alpha^2\sigma_X^2]. \qquad (17)$$

Note that $\alpha^2\sigma_X^2$ is the (true) variance of the signal $Y$ in $y$ [see (1) and (3)], and so, analogous to the noise-to-signal ratio $n_x$, the parameter $n_y$ in (16) is the noise-to-signal ratio for $y$. Squaring $r$ gives, from (16), the constraint

$$r^2 = (1 + n_x)^{-1}(1 + n_y)^{-1} \qquad (18)$$

on $n_x$ and $n_y$ since $r$ can be determined from the data.

As $n_x \geq 0$ and $n_y \geq 0$, it follows from (18) that $n_x$ and $n_y$ both vary between 0 and $r^{-2} - 1$; geometrically, the points ($n_x$, $n_y$) lie along a hyperbolic curve (see Fig. 1). Because we know nothing about the relative sizes of $n_x$ and $n_y$, we assume that the points ($n_x$, $n_y$) are uniformly distributed along the curve between the limiting values at $A$ ($n_x = 0, n_y = r^{-2} - 1$) and $B$ ($n_x = r^{-2} - 1, n_y = 0$). If $l$ denotes arc length and $l_{AB}$ is the length of the curve

$AB$, then the required uniformly distributed probability density function (pdf) is

$$g = l_{AB}^{-1} = \left(\int_A^B dl\right)^{-1}, \tag{19}$$

since this is the same value for any point on the curve $AB$ and satisfies $\int_A^B g\,dl = 1$.

Note that other assumptions about the distribution of $n_x$ and $n_y$ along the curve (18) are not justifiable. For example, if we assume $n_x$ is uniformly distributed along the curve, then because of the hyperbolic form of (18), $n_y$ is not uniformly distributed along the curve. Different distributions would be inappropriate, since we have no way of distinguishing the noise in $x$ and $y$; all we have is the knowledge that the noise pair $(n_x, n_y)$ lies somewhere along the curve defined by (18) and illustrated in Fig. 1.

The cumulative distribution function corresponding to (19) is

$$G = \int_0^l dl/l_{AB}, \tag{20}$$

or, in terms of $n_x$,

$$G = \int_0^{n_x} \sqrt{1 + (dn_y/dn_x)^2}\, dn_x/l_{AB}. \tag{21}$$

From (18) we can find $n_y$ in terms of $n_x$ and $r$ and hence calculate $dn_y/dn_x$ to obtain, from (21),

$$G = \int_0^{n_x} [1 + (1 + n_x)^{-4} r^{-4}]^{1/2}\, dn_x. \tag{22}$$

From (9),

$$\alpha/\alpha_{\mathrm{OLS}} = 1 + n_x, \tag{23}$$

and from (13), $\alpha_{\mathrm{GMR}} = \alpha_{\mathrm{OLS}}/|r|$. Therefore,

$$\alpha/\alpha_{\mathrm{GMR}} = (1 + n_x)|r| = \zeta, \tag{24}$$

where we have introduced the variable $\zeta$ in (24) for notational convenience. Changing variables in (22) to $\zeta$, we have

$$G(\zeta) = |r|^{-1} \int_{|r|}^{\zeta} (1 + \zeta_*^{-4})^{1/2} d\zeta_*/l_{AB}. \tag{25}$$

Thus,

$$g(\zeta) = dG/d\zeta = (1 + \zeta^{-4})^{1/2}/(l_{AB}|r|). \tag{26}$$

Since $\zeta = \alpha/\alpha_{\mathrm{GMR}}$ from (24), $g$ in (26) is the pdf for $\alpha/\alpha_{\mathrm{GMR}}$. It enables us to calculate confidence intervals for $\alpha$ given a knowledge of the parameters $r$ and $\alpha_{\mathrm{GMR}}$, which we can calculate from the data.

Figure 2 shows examples of $g$ for various values of $r$. It follows from (24) and $0 \le n_x \le r^{-2} - 1$ that

$$|r| \le \alpha/\alpha_{\mathrm{GMR}} \le |r|^{-1}, \tag{27}$$

so plots of $g$ as a function of $\alpha/\alpha_{\mathrm{GMR}}$ vary over this domain. The density function $g$ has a long tail for small $|r|$; that is, the confidence intervals are wide, because for small $|r|$ not much of the variance is explained by the regression fit. From (20), the median of $g$ occurs when $l = (1/2)l_{AB}$, that is, halfway along the hyperbolic curve $AB$ (see Fig. 1). By symmetry this corresponds to $n_x = n_y$, or by (18), when $(1 + n_x)|r| = 1$. From (24) this is equivalent to $\alpha = \alpha_{\mathrm{GMR}}$. This differs from the most likely value for $\alpha/\alpha_{\mathrm{GMR}}$, which occurs when $g$ is a maximum. Since $g$ is a monotonically decreasing function of $\alpha/\alpha_{\mathrm{GMR}}$ [see (26)], from (27) its maximum value occurs when

$$\alpha = |r|\alpha_{\mathrm{GMR}} = \alpha_{\mathrm{OLS}}. \tag{28}$$

Note from (18) that the curve $AB$ in Fig. 1 of length $l_{AB}$ depends on $r^2$. As $r^2 \to 1$, the intercepts at $A(0, r^{-2} - 1)$ and $B(r^{-2} - 1, 0)$ approach 0 and so the length of the curve $l_{AB}$ approaches zero; as $r^2 \to 0$, $l_{AB} \to \infty$. Figure 3 shows $l_{AB}$ as a function of $r^2$. Note that if we approximate the length of $l_{AB}$ using a straight line between $A$ and $B$, then $l_{AB}$ is the length of the hypotenuse of a right-angled isosceles triangle with equal sides $r^{-2} - 1$, and so

$$l_{AB} = \sqrt{2}(r^{-2} - 1) = \sqrt{2}(1 - r^2)/r^2. \tag{29}$$

Figures 1 and 3 show that this approximation becomes increasingly accurate as $r^2 \to 1$, and that it is a reasonable approximation even when $0.3 \le r^2 \le 1$. Thus (26) can also be written approximately, with increasing accuracy as $r^2 \to 1$, as

$$g(\alpha/\alpha_{\mathrm{GMR}}, r^2) = 2^{-1/2}(1 + (\alpha/\alpha_{\mathrm{GMR}})^{-4})^{1/2}|r|/(1 - r^2). \tag{30}$$

While it is useful to have obtained results for the limiting large $M$ case, in practice we are usually faced with the problem of determining $\alpha$ given a finite number $M$ of hard-won data points and finite $M$ estimates $\hat{r}$ and $\hat{\alpha}_{\mathrm{GMR}}$ of the correlation and regression coefficients. In the $M = \infty$ case, the confidence intervals follow directly from the pdf (26), but in the finite $M$ case we know only

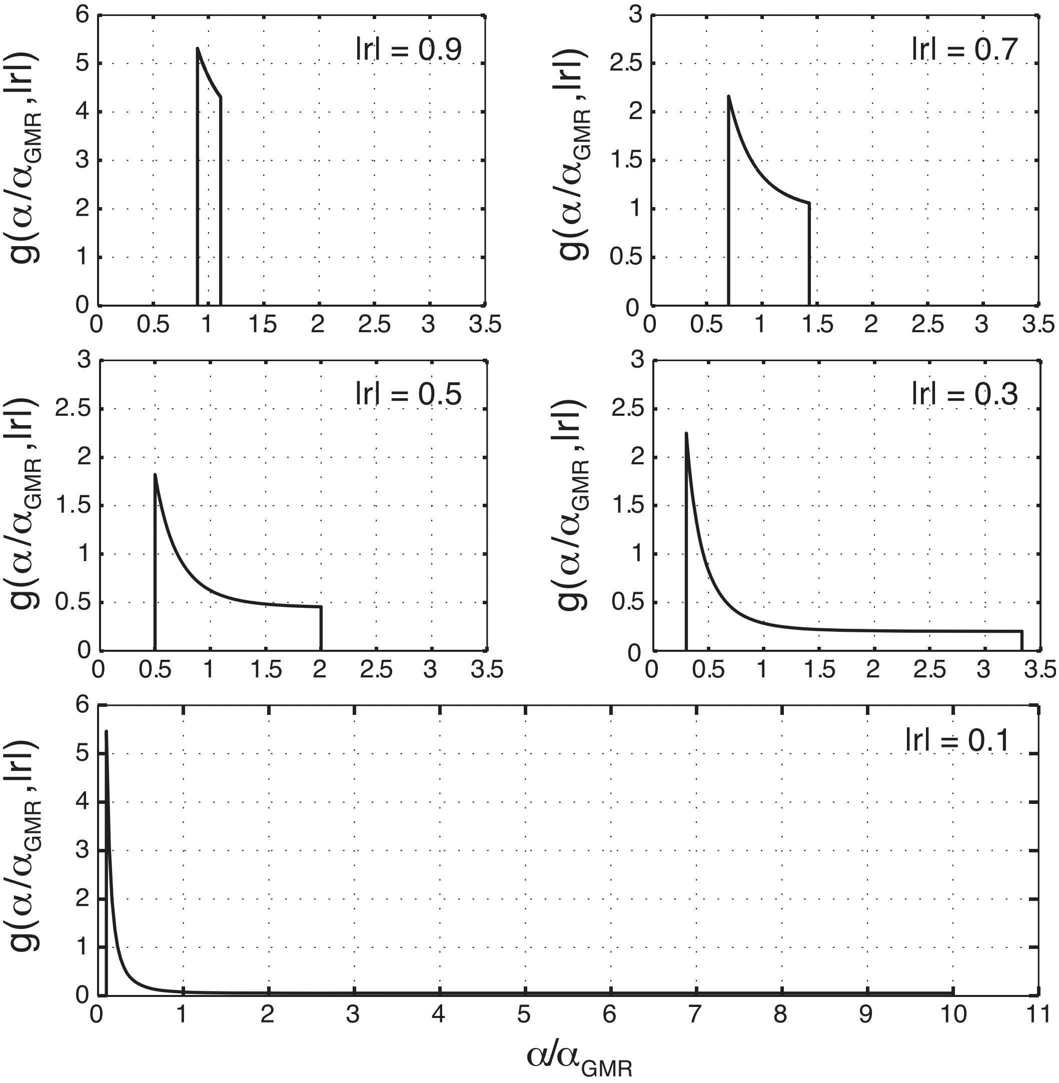

FIG. 2. The $M = \infty$ pdf $g(\alpha/\alpha_{\mathrm{GMR}}, r^2)$ for five values of $|r|$. As noted in the text, the domain of $g$ in each case is $[|r|, |r|^{-1}]$.

$\hat{r}$ and $\hat{\alpha}_{\mathrm{GMR}}$ rather than $r$ and $\alpha_{\mathrm{GMR}}$. In the next section, we show how to calculate the relevant finite $M$ pdf and corresponding confidence intervals numerically.

## 4. Confidence intervals for the true regression coefficient for finite *M*

To calculate the confidence intervals for finite $M$ and a given $\hat{r}$ and $\hat{\alpha}_{\mathrm{GMR}}$, begin by defining

$$x_* = (x - \overline{x})/\sigma_X, \quad y_* = (y - \overline{y})/|\alpha|\sigma_X. \tag{31}$$

The sample correlation coefficient for $x_*$ and $y_*$ is the same as that for $x$ and $y$ while the sample regression coefficient $\hat{\alpha}_{*\mathrm{GMR}}$ of $y_*$ on $x_*$ is

$$\hat{\alpha}_{*\mathrm{GMR}} = \hat{\alpha}_{\mathrm{GMR}}/|\alpha|. \tag{32}$$

Noting from (2) and (3) that $\overline{x} = \overline{X} + \overline{\varepsilon}$ and $\overline{y} = \alpha\overline{X} + a + \overline{\delta}$, we have

$$x - \overline{x} = X - \overline{X} + \varepsilon - \overline{\varepsilon} \tag{33}$$

and

$$y - \overline{y} = \alpha(X - \overline{X}) + \delta - \overline{\delta}. \tag{34}$$

If we now divide (33) by $\sigma_X$ and (34) by $|\alpha|\sigma_X$ and use (31), we may write (33) and (34) as

$$x_* = (X - \overline{X})/\sigma_X + \varepsilon/\sigma_X - \overline{\varepsilon}/\sigma_X \tag{35}$$

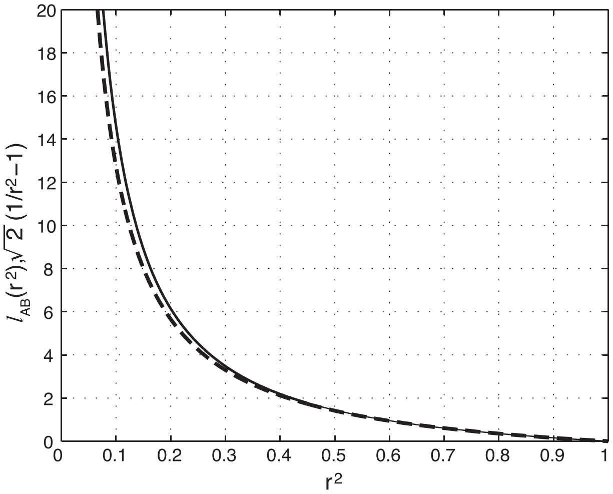

FIG. 3. The length of the hyperbolic curves $l_{AB}(r^2)$ in Fig. 1 as a function of $r^2$ (solid line). The dashed line is $\sqrt{2}(r^{-2}-1)$, the length of the straight line approximation to those curves.

and

$$y_* = \mathrm{sgn}(\alpha)(X-\overline{X})/\sigma_X + \delta/|\alpha|\sigma_X - \overline{\delta}/|\alpha|\sigma_X. \qquad (36)$$

The right-hand sides of (35) and (36) may be written in terms of $n_x$, $n_y$, and the zero-mean, unit variance variables

$$\varepsilon_* = \varepsilon/\sigma_\varepsilon, \qquad (37)$$

$$\delta_* = \delta/\sigma_\delta, \qquad (38)$$

and

$$X_* = (X-\mu_X)/\sigma_X, \qquad (39)$$

where $\mu_X$ is the true mean of $X$. To see that (35) and (36) can indeed be written in this way, first note that

$$\begin{aligned} X_* - \overline{X}_* &= (X-\mu_X)/\sigma_X - (\overline{X}-\mu_X)/\sigma_X \\ &= (X-\overline{X})/\sigma_X. \end{aligned} \qquad (40)$$

Also, from (10), (17), (37), and (38) we have

$$\varepsilon/\sigma_X = \varepsilon_* n_x^{1/2} \qquad (41)$$

and

$$\delta/|\alpha|\sigma_X = \delta_* n_y^{1/2}. \qquad (42)$$

The substitution of (40), (41), and (42) into (35) and (36) yields

$$x_* = X_* - \overline{X}_* + (\varepsilon_* - \overline{\varepsilon}_*)n_x^{1/2} \qquad (43)$$

and

$$y_* = \mathrm{sgn}(\alpha)(X_* - \overline{X}_*) + (\delta_* - \overline{\delta}_*)n_y^{1/2}. \qquad (44)$$

We found the required confidence intervals numerically from (43) and (44), sampling the zero-mean, unit-variance variables $X_*$, $\varepsilon_*$, and $\delta_*$ from independent zero-mean, unit-variance normal distributions. The numerical calculations proceed by first obtaining a random $r$ from a uniform distribution over the interval $[-1, 1]$. For that $r$ we then randomly sample a point $(n_x, n_y)$ from a uniform distribution along the curve $AB$ corresponding to $r^2$. Now that we have $n_x$ and $n_y$ we can use the independent zero-mean, unit-variance normal distributions of $X_*$, $\varepsilon_*$, and $\delta_*$ to obtain $M$ random samples of $x_*$ and $y_*$ from (43) and (44) and hence obtain the correlation coefficient $\hat{r}$, which, as we noted earlier, is the same as that for $x$ and $y$. We can also calculate the regression coefficient $\hat{\alpha}_{*\mathrm{GMR}}$, which, from (32), is $\hat{\alpha}_{\mathrm{GMR}}/|\alpha|$. But since we know $r$ for this sample, we also know $\mathrm{sgn}(\alpha)$ and hence $\hat{\alpha}_{\mathrm{GMR}}/\alpha$. Proceeding in this way we calculated 100 million samples of $\alpha/\hat{\alpha}_{\mathrm{GMR}}$ and thus calculated the probability density function of $\alpha/\hat{\alpha}_{\mathrm{GMR}}$ as a function of $\hat{r}$.

This pdf can be used to find the lower and upper limits $L$ and $U$ for confidence intervals for the true regression coefficient $\alpha$. For example, for a 95% confidence interval we can find $L$ and $U$ such that

$$\mathrm{probability}(L \le \alpha/\hat{\alpha}_{\mathrm{GMR}} \le U) = 95\%. \qquad (45)$$

Once $L$ and $U$ are known, then for a given sample $(x, y)$ of $M$ points we can calculate $\hat{\alpha}_{\mathrm{GMR}}$ and then use (45) to estimate the 95% confidence intervals for $\alpha$ as $(L\hat{\alpha}_{\mathrm{GMR}}, U\hat{\alpha}_{\mathrm{GMR}})$ for $\hat{\alpha}_{\mathrm{GMR}} > 0$ and $(U\hat{\alpha}_{\mathrm{GMR}}, L\hat{\alpha}_{\mathrm{GMR}})$ for $\hat{\alpha}_{\mathrm{GMR}} < 0$. Other confidence intervals can similarly be obtained.

Since we have an analytical solution for $M = \infty$ and numerical results for finite $M$, it is of interest to see how large $M$ has to be for the $M = \infty$ confidence limits to be approximately correct. Figure 4 shows the 95% $L$ and $U$ as a function of $\hat{r}$ for various $M$. For small $M$ it is only when $|\hat{r}| \rightarrow 1$ that the finite $M$ case resembles the $M = \infty$ case. As for the $M = \infty$ pdf (see Fig. 2), distributions are skewed right so that, as $|\hat{r}|$ decreases, $U$ increases more rapidly than $L$ decreases. In the finite case, plots end when $L = 0$, because for smaller $|\hat{r}|$ even the sign of $\alpha$ is not known with 95% confidence.

## 5. An example

Tables for $L$ and $U$ corresponding to the 80%, 90%, 95%, and 99% confidence intervals for various $M$ and $\hat{r}$

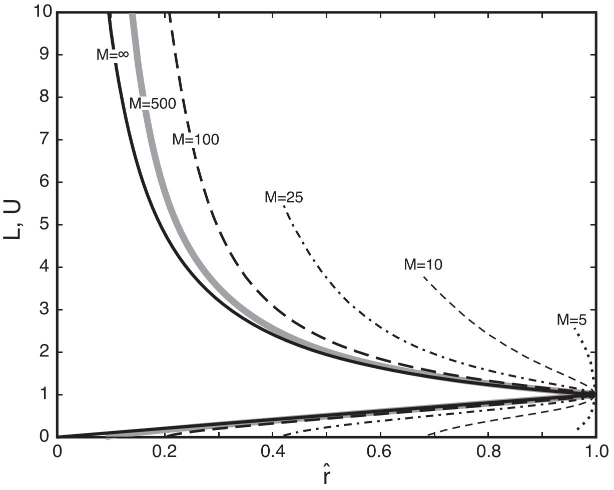

FIG. 4. Upper ($U$) and lower ($L$) 95% confidence interval limits as a function of $\hat{r}$ for the ratio $\alpha/\hat{\alpha}_{\mathrm{GMR}}$ for $M = 5$ (dotted line), $M = 10$ (short-dashed line), $M = 25$ (dashed–dotted line), $M = 100$ (long-dashed line), $M = 500$ (gray line), and $M = \infty$ (black solid line). The upper confidence intervals $U$ are all ≥1 and lower confidence intervals $L$ are all ≤1. In the finite cases the plots end when $L = 0$ because for lower $\hat{r}$, even the sign of $\alpha$ is not known with 95% confidence. The curves approach the analytical $M = \infty$ limit as $\hat{r} \rightarrow 1$ or $M \rightarrow \infty$.

are provided in the supplemental material that is online (http://dx.doi.org/10.1175/JTECH-D-12-00067.s1.). These tables can be used on a wide variety of scientific data and we here illustrate their use in an example relating monthly atmospheric anomaly time series strongly influenced by El Niño. These time series are anomalous in that they are monthly time series that are departures from the seasonal cycle. Theory (Bunge and Clarke 2009) for the equatorial Pacific atmospheric boundary layer suggests that the monthly surface equatorial atmospheric pressure anomalies in the western equatorial Pacific $p(o, t)$ and eastern equatorial Pacific $p(L, t)$ should be related to the zonal integral of the eastward equatorial wind stress anomaly $\tau$ by

$$\int_0^L \tau(\xi, t)\, d\xi = h(L)[p(L, t) - 1.67p(o, t)]. \qquad (46)$$

In (46), $t$ is the time in months, $\xi = 0$ refers to the western equatorial Pacific boundary, $\xi = L$ to the eastern equatorial Pacific boundary, and $h(L)$ is a measure of the height of the turbulent atmospheric boundary layer in the eastern equatorial Pacific. Note that both monthly time series on the left- and right-hand sides of (46) contain noise. This noise is due to the measurement error in the thousands of ships of opportunity wind and pressure observations, the measurement error in the station pressure measurements at Darwin, Australia, and their adjustment to the equator, and the noise due to theoretical approximations in the derivation of (46). The noise-to-signal ratios $n_x$ and $n_y$ are thus nonzero and unknown and this must be taken into account when estimating $h(L)$ by regression.

Bunge and Clarke (2009) tested the validity of (46) using monthly pressure and wind stress data from 1978 to 2003 inclusive. Since the data are autocorrelated, the number of degrees of freedom in the data is not the number of months of data, but rather the number of years of data because El Niño time series can be thought of as independent 12-month segments [see, e.g., Fig. 2.14 of Clarke (2008)]. As there are 26 yr of data, $M = 26$. Also, Bunge and Clarke found from the correlation of the time series in (46) and their standard deviations that $\hat{r} = 0.91$ and $\hat{\alpha}_{\mathrm{GMR}} = \hat{h}(L) = 188$ m.

The required 95% confidence interval limits $L$ and $U$ for $\alpha$ can be obtained for $M = 26$ by linearly interpolating the values for $M = 25$ and $M = 35$ in the supplement. For example, from Table S17, the 95% entry for $L$ for $\hat{r} = 0.91$ is 0.7845 for $M = 25$ while the corresponding value from Table S18 for the $M = 35$ case is 0.8149. Linear interpolation gives $L = 0.7875$ for $M = 26$ and a similar analysis gives $U = 1.2706$. This implies, from the discussion associated with (45), that the 95% confidence interval for the true value of $\alpha = h(L)$ is

$$(L\hat{\alpha}_{\mathrm{GMR}}, U\hat{\alpha}_{\mathrm{GMR}}) = (148\,\mathrm{m}, 239\,\mathrm{m}). \qquad (47)$$

Note that the above confidence interval takes into account both the finite number of points $M$ *and* our uncertainty about the noise. By comparison, the standard ordinary least squares regression of the left-hand side of (46) and on the right-hand side gives a 95% confidence interval (138 m, 204 m). This interval is smaller than that in (47), but it is a confidence interval for the $M = \infty$ regression coefficient $\alpha_{\mathrm{OLS}}$ rather than the true regression coefficient $\alpha$. The confidence interval for the true regression coefficient is larger because, in addition to the finite $M$ limitation, the confidence interval for the required true regression coefficient takes into account the unknown noise.

## 6. Concluding remarks

Two reviewers' comments made us think that we should point out here the difference between linear prediction and our goal of estimating the true regression coefficient $\alpha$. In linear prediction the aim is to predict $y$ from $x$ linearly. In that case, the ordinary least squares regression coefficient $\hat{\alpha}_{\mathrm{OLS}}$ is appropriate because it matches $y$ as best it can (in the least squares sense) given past data. When the noise is large, $r^2$ is small [see (18)], and such a linear predictor will only explain a small percentage of the variance. This results in all predictions being near the mean. Because of this, in some prediction circles the regression fit is ''inflated'' so that the regression estimates have the same variance as $y$ (i.e., effectively $\hat{\alpha}_{\mathrm{GMR}}$ is used as the regression coefficient). It is recognized that this is at the expense of the regression coefficient no longer giving the best least squares fit to $y$.

The preceding is related to, but separate from, our goal to find the true regression coefficient $\alpha$ between the variables $Y$ and $X$ given noisy realizations $y$ and $x$ as stated in (2) and (3). In our case, if it is known that the ''noise''-to-signal ratio in at least one of the variables $x$ and $y$ (say $x$) is small, then the ordinary least squares regression coefficient is nearly unbiased and can be used [see (9) with small $n_x$]. But when the noise-to-signal ratio is not known in both variables, the ordinary least squares regression coefficient is biased. In that case the confidence intervals that are calculated for $\hat{\alpha}_{\mathrm{OLS}}$ just inform you about the true least squares regression coefficient $\alpha_{\mathrm{OLS}}$ rather than the regression coefficient $\alpha$, the coefficient we really want. When the noise-to-signal ratio is not known in both variables, $\hat{\alpha}_{\mathrm{GMR}}$ and the tables in the supplement provide a simple practical way to estimate confidence intervals for the true regression coefficient. All that is needed is the number $M$ of independent samples, and, from the data, the standard deviations of each variable and the correlation coefficient between them.

*Acknowledgments.* We thank F. Huffer for helpful comments on an early version of our paper and the National Science Foundation for financial support (Grants OCE-0850749 and OCE-1155257).

## REFERENCES

Barker, F., Y. C. Soh, and R. J. Evans, 1988: Properties of the geometric mean functional relationship. *Biometrics,* **44,** 279–281.

Bunge, L., and A. J. Clarke, 2009: A verified estimation of the El Niño index Niño-3.4 since 1877. *J. Climate,* **22,** 3979–3992.

Clarke, A. J., 2008: *An Introduction to the Dynamics of El Niño & the Southern Oscillation.* Academic Press, 324 pp.

Emery, W. J., and R. E. Thomson, 2001: *Data Analysis Methods in Physical Oceanography.* 2nd rev. ed. Elsevier, 638 pp.

Frost, C., and S. G. Thompson, 2000: Correcting for regression dilution bias: Comparison of methods for a single predictor variable. *J. Roy. Stat. Soc.,* **A163,** 173–189.

Garrett, C., and B. Petrie, 1981: Dynamical aspects of the flow through the Strait of Belle Isle. *J. Phys. Oceanogr.,* **11,** 376–393.

Jolicoeur, P., 1975: Linear regressions in fisheries research: Some comments. *J. Fish. Res. Board Canada,* **32,** 1491–1494.

Kendall, M. G., and A. Stuart, 1973: *The Advanced Theory of Statistics.* 3rd ed. Vol. 2, Griffin, 723 pp.

McArdle, B. H., 1988: The structural relationship: Regression in biology. *Can. J. Zool.,* **66,** 2329–2339.

——, 2003: Lines, models, and errors: Regression in the field. *Limnol. Oceanogr.,* **48,** 1363–1366.

Ricker, W. E., 1973: Linear regressions in fishery research. *J. Fish. Res. Board Canada,* **30,** 409–434.

——, 1975: A note concerning Professor Jolicoeur's comments. *J. Fish. Res. Board Canada,* **32,** 1494–1498.

Riggs, D. S., J. A. Guarnieri, and S. Addelman, 1978: Fitting straight lines when both variables are subject to error. *Life Sci.,* **22,** 1305–1360.

Sokal, R. R., and F. J. Rohlf, 1995: *Biometry: The Principles and Practice of Statistics in Biological Research.* 3rd ed. W. H. Freeman and Co., 887 pp.

Sprent, P., and G. R. Dolby, 1980: The geometric mean functional relationship. *Biometrics,* **36,** 547–550.