Correcting parameter bias caused by taking logs of exponential data

William J. Thompson and J. Ross Macdonald
Department of Physics and Astronomy, University of North Carolina, Chapel Hill, North Carolina 27599-3255

(Received 23 July 1990; accepted for publication 26 October 1990)

Exponential growth and decay are ubiquitous in physics, and when teaching techniques of data analysis in experimental physics we show students how the simple device of taking logarithms can reduce a highly nonlinear problem to a linear one, from which estimates of the slope (exponent) and intercept (preexponential) can be readily obtained, either graphically or by using a linear least-squares-fitting program. Here, we show that this seemingly innocuous procedure of taking logs usually results in biased values of fitting parameters but that such biases can often be simply corrected. This problem is mentioned but not solved in, for example, Ref. 1.

A moment of reflection will show why the biasing occurs. Consider the example of data that range from $1/M$

through unity up to M. For large M there are two subranges, with lengths roughly unity and M, respectively. After taking (natural) logs, the subranges are each of length $\ln(M)$, so that the smallest values of the original data have been unnaturally extended relative to the larger values. The effect of this is to make derived parameter estimates different from the true values.

In the course of a recent extensive analysis[2] of data-fitting methods for nonlinear data of wide range and nonuniform error variance ("heteroscedasticity"), we discovered that the existence of such parameter bias is known to statisticians but it has not been quantified by them, and that texts on statistical and data analysis methods for the physical sciences usually do not even mention it. In the following, we present a simplified version of our analysis that physics students can follow and that is realistic.

We first define the fitting function, Y, in terms of the independent variable, x, by the exponential relation

$$Y(x) = A\exp(Bx), \tag{1}$$

in which the fitting parameters are the preexponential A and the exponent B (positive for growth, negative for decay). Suppose that the data to be described by the function in Eq. (1) are $y(x_i)$; that is,

$$y(x_i) = Y(x_i) + e_i, \tag{2}$$

in which e_i is the unknown random error for the ith datum. Under log transformation, Eqs. (1) and (2) result in

$$\ln(y_i) = \ln(A) + \ln[1 + e_i/Y(x_i)] + Bx_i, \tag{3}$$

which, if the e_i were ignored, would be a linear relation between the transformed data and the independent variable values x_i. If Eq. (1) is substituted into Eq. (3), A and B appear in a very complicated nonlinear way that prevents the use of linear least-squares methods.

To proceed requires an "error model," that is, a model for how the distribution of the errors e_i depends upon the x_i. The only possibility in Eq. (3) that allows straightforward estimates of bias and that is independent of the x_i is to assume proportional random errors, that is,

$$e_i = \sigma Y(x_i)P(0,I_i), \tag{4}$$

in which σ is the same standard deviation of $e_i/Y(x_i)$ at each i. The notation $P(0,I_i)$ is to be interpreted as follows. In statistics $P(0,I)$ denotes an independent probability distribution, P, having zero mean and unity standard deviation at each i. Since I is a unit vector, $I_i = 1$ for all i, and $P(0,I_i)$ is a random choice from $P(0,I)$ for each i. For example, a Gaussian distribution has $P(0,I_i) = \exp(-t_i^2/2)/\sqrt{(2\pi)}$, where t_i is a random variable parametrizing the distribution of errors at each data point. Proportional errors (constant percentage errors from point to point) are common in many physics measurements by appropriate design of the experiment. An exception is radioactivity measurements, which have Poisson statistics[3] with square-root errors, unless counting intervals are steadily increased to compensate count rates that decrease with time.

Before Eqs. (3) and (4) can be used for fitting, we have to take expectation values, E in statistical nomenclature,[4] on both sides, corresponding to many repeated measurements of each datum. We assume that each x_i is precise, so that we obtain

$$E\{\ln(y_i)\} = \ln(A_b) + Bx_i, \tag{5}$$

in which the biased estimate of the intercept, A_b, is given by

$$A_b = A\exp(E\{\ln[1 + \sigma P(0,I)]\}). \tag{6}$$

The use of I rather than I_i is a reminder that the expectation value is to taken over all the data. Even when only a single set of observations is available, it is still most appropriate to correct the bias in the estimate of A by using Eq. (6) as described below. An estimate of the fractional standard deviation σ can be obtained either experimentally by choosing a representative x_i and making repeated measurements of y_i, or computationally it can be obtained from the standard deviation of the least-squares fit.[2]

Equation (6) shows that a straightforward least-squares fit of the log-transformed data will give a biased value for A, namely A_b, and that the amount of bias will depend both upon the size of the error (σ) and its distribution (P) but, most importantly, not at all on the x_i. Note also that in this error model the exponent B (which is often of primary interest) is unbiased.

The bias in A can be estimated by expanding the logarithm in Eq. (6) in a Maclaurin series, then evaluating the expectation values term by term. The unbiased value, A, can be estimated from the extracted biased value A_b in Eq. (5) by solving for A in Eq. (6) to obtain

$$A = A_b\exp[L_b(P)], \tag{7}$$

where the bias term, $L_b(P)$, is given by

$$L_b(P) = \sigma^2/2 + S(P), \tag{8}$$

with

$$S(P) = \sum(-1)^m\sigma^m E_m(P)/m, \tag{9}$$

where the sum starts at $m = 3$ and E_m denotes the mth moment of the distribution P. The first term in the Maclaurin series vanishes because P is to have zero mean, while the second term contributes $\sigma^2/2$, since P is to have unity standard deviation (second moment about the mean). The remaining sum, Eq. (9), depends on the error distribution P. For example, for the commonly assumed Gaussian (normal) distribution $P = P_G$, its third moment vanishes because of its symmetry and its fourth moment[5] gives $L_b(P_G) \approx \sigma^2/2 + 3\sigma^4/4$, while for the uniform (rectangular) distribution $P = P_U$, one obtains $L_b(P_U) \approx \sigma^2/2 + 9\sigma^4/20$.

Table I gives examples of the bias induced in the preexponential A by a logarithmic transformation. Note that A needs to be corrected upward by 0.5% for data with 10% standard deviation ($\sigma = 0.1$) and upward by about 5% for data with 30% random errors ($\sigma = 0.3$).

Table I. Logarithmic bias estimate exponents in Eqs. (7) and (8) for lowest order, for Gaussian (P_G), for Monte Carlo simulation estimated ($L_{b\,MC}$) from the distributions displayed in Fig. 1, and for uniform (P_U) error distributions.

	σ		
Bias estimate	0.1	0.2	0.3
$\sigma^2/2$	0.005 00	0.0200	0.045
$L_b(P_G)$	0.005 08	0.0214	0.053
$L_{b\,MC}$	0.005 08	0.0214	0.054
$L_b(P_U)$	0.005 05	0.0208	0.049

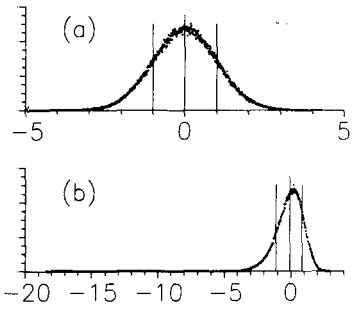

Fig. 1. (a) The Gaussian (normal) distribution, $P_G(0,I)$, with zero mean and unity standard deviation randomly sampled 200 000 times. (b) The distribution of $\ln[1+\sigma P_G(0,I)]$, as used in Eq. (6), with the same abscissa as in (a), for $\sigma = 0.2$. In both plots the central vertical shows the mean and the outer verticals show 1 s.d. from the mean. By including all the points in the sample shown in (a), the ln distribution acquires the very long negative tail shown in (b).

As a way of confirming the above analysis, we made a Monte Carlo simulation of the random error distributions, as follows. We used a computer random number generator to provide a sample of 200 000 values from a Gaussian distribution, and we forced this sample to have zero mean and unity standard deviation, as the above analysis uses. The sample was then sorted into 800 bins, producing the distribution shown in Fig. 1 (a). Choosing $\sigma = 0.1$, 0.2, or 0.3, we then formed $\ln[1+\sigma P(0,I)]$, as in Eq. (6). The corresponding distribution for $\sigma = 0.2$ (20% error) is shown in Fig. 1 (b), where the long negative tail that induces the bias in A is evident. The Monte Carlo estimate of the bias is just the negative of the mean value of this distribution, which we call $L_{b\,\mathrm{MC}}$. Table I shows that the agreement with our analytic estimate, $L_b(P_G)$, is very close.

The mathematically punctilious reader should object to our analysis for the Gaussian distribution, because the argument of the ln function in the error model may become negative, even though this is very improbable if σ is small. (For $\sigma = 0.2$ the probability is about 3×10^{-7}.) Therefore, in the analytical treatment Eq. (9) represents an asymptotic series which eventually diverges, while in the Monte Carlo simulation if the sample size is very large the chance of getting a negative argument increases. Formally, this problem can be circumvented by defining suitably truncated distributions whose low-order moments are nearly the same as the complete distributions, so that there are no practical consequences. For a uniform distribution, the problem arises only if $\sigma > 1/\sqrt{3} = 0.58$, which would usually be considered too large an error to justify anything more than a cursory fit.

Clearly, the simple corrections suggested by Eqs. (7)–(9) are worth making if the assumption of proportional random errors, Eq. (4), is realistic. It is also reassuring that the exponent B is unbiased under this assumption. For any other error model the logarithmic transformation induces biases in both the exponent B and the preexponent A which cannot be easily corrected.

ACKNOWLEDGMENTS

We thank Timothy C. Black for thoughtful remarks on the manuscript and the referee for suggestions on tailoring the presentation for physicists.

[1] John R. Taylor, *An Introduction to Error Analysis* (University Science Books, Mill Valley, CA, 1982), pp. 166, 167.
[2] J. Ross Macdonald and William J. Thompson, "Strongly heteroscedastic nonlinear regression," submitted to Communications in Statistics.
[3] Reference 1, Chap. 11.
[4] For example, K. A. Brownlee, *Statistical Theory and Methodology* (Wiley, New York, 1965), 2nd ed., Chap. 1.15.
[5] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1965), Chap. 26.