WATER RESEARCH 42 (2008) 4033–4040

Available at www.sciencedirect.com

ScienceDirect

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

An alternative to specious linearization of environmental models

Richard H. McCuen[a,*], Cristiane Q. Surbeck[b]

[a]Department of Civil and Environmental Engineering, University of Maryland, College Park, MD 20742-3021, USA
[b]Department of Civil Engineering, Carrier Hall, University of Mississippi, University, MS 38677, USA

ARTICLE INFO

Article history:
Received 26 October 2007
Received in revised form
7 May 2008
Accepted 20 May 2008
Available online 25 June 2008

Keywords:
Modeling
Ratio correlation
Linearization
Specious correlation
Spurious correlation
Monte Carlo analysis
Nonlinear least squares
Monod
Michaelis–Menten
Sludge filtration
Adsorption
Sediment rating curve

ABSTRACT

The solution of a number of environmental models is incorrectly obtained by linearizing a nonlinear analytical solution. The linearization can yield a model that includes a common variable on both sides of the equal sign (i.e., ratio analysis), which in calibration causes highly inflated goodness-of-fit statistics. These specious practices continue likely because of tradition, i.e., "that is the way it is done". Goodness-of-fit statistics that result from these erroneous practices do not accurately reflect the actual prediction accuracy of the model. Additionally, the linearly calibrated coefficients can be poor estimators of the true coefficients. The goal of this paper is to demonstrate the pitfalls of models based on ratio analyses. Several environmental models are used to demonstrate the erroneous procedure. Monte Carlo simulation is used to show the distribution of the true correlation coefficient and compare it to the distribution that results from the erroneous linearization. Linearization can produce correlation coefficients above 0.9 when the actual correlation is near 0. Nonlinear least squares algorithms can be used to more accurately fit nonlinear data to nonlinear models.

1. Introduction

Knowledge of modeling techniques has not kept pace with the growth of technical knowledge or technology itself. Methods of measurement have improved significantly over the last decade, yet specious (i.e., deceptively attractive) modeling methods applied to highly accurate measurements will not necessarily produce accurate analyses or designs. While we certainly want to continue with improvements in both measurement and technology, efforts must be made to upgrade the general understanding of methods of modeling, i.e., the extraction of information from measured data.

Ratio analyses of correlation and regression are one example of poor modeling practices. These problems were recognized in 1932 by Karl Pearson, the individual credited with many of the earlier advances of least squares and correlation (1932).

In general terms, ratio analyses involve combining random variables into a ratio (or a product), performing statistical analyses with the ratio, and then separating the ratio for the

* *Corresponding author*. Tel.: +1 301 405 1949; fax: +1 301 405 2585.
E-mail address: rhmccuen@eng.umd.edu (R.H. McCuen).
0043-1354/$ – see front matter
doi:10.1016/j.watres.2008.05.030

purpose of estimating one of the variables that was used to form the ratio. For example, assume y and x are two random variables, and we have a set of measurements on y and x that suggest a poor degree of correlation between the two variables. However, the formation of a ratio, such as y/x, produces a model that seemingly is highly accurate (i.e., the correlation coefficient is high) when regressed on x. As a simple linear case, the model might be $y/x = a + bx$. Then a predicted value of y is obtained by predicting the ratio y/x with a new value of x, and then multiplying the predicted value of y/x by x to yield a predicted value of y. A y/x vs. x regression will yield a high correlation, but unfortunately the correlation coefficient reflects the accuracy of estimates of y/x, but not the accuracy of estimates of y alone. Actually, predicted values of y from such analyses can be quite inaccurate. The high correlation coefficient was the result of x appearing on both sides of the prediction equation. It should be obvious that 1/x and x would be highly correlated, which yields a specious measure of the accuracy of predictions of y.

Articles, as well as a book, demonstrating the speciousness of such analyses have been published in many of the leading journals from a variety of disciplines (Benson, 1965; Chayes, 1972; Hao and Neethling, 1987; Kenney, 1982; Kritzer, 1990). Failure to recognize the pitfalls of self-correlation and self-regression has led to many inaccurate, may be misleading, results.

Pearson (1932) and Benson (1965) wrote about the problem of correlation of ratios. Ratio models also result from the linearization of a nonlinear relationship. This is the case in which scientific theory leads to a nonlinear model that includes one or more coefficients that must be calculated, i.e., fitted to data. This paper presents examples of these theory-based nonlinear relationships (the Michaelis–Menten or Monod model, an adsorption kinetic rate equation, and the sludge filtration equation) and one example of a linearized empirically based relationship (the sediment rating curve). All of these models have been applied using erroneous linearization because the model developer may not be aware of numerical methods of fitting coefficients of a nonlinear model. Consequently, the modeler rearranges the nonlinear model to a linear form, i.e., $y = a + bx$, that can be fitted with traditional analytical least squares regression. After linearizing the nonlinear model, the dependent y variable is actually a ratio.

The intent herein is to demonstrate the speciousness of ratio analyses and present an alternative to the procedure that directs individuals away from the antiquated modeling practices of forming ratios and linearization.

2. Examples of specious transformations

In this section, a number of commonly used examples that lead to specious linearization will be provided. Specific analyses of the more common examples will be provided in subsequent sections.

Kinetic rate modeling: consider the Michaelis–Menten enzymatic equation (Michaelis and Menten, 1913), or Monod growth model (Monod, 1949):

$$\mu = \frac{\mu_x S}{K + S} \quad (1)$$

where values of μ and S are measured and μ_x and K are fitting coefficients. By taking the reciprocal of both sides of Eq. (1) and rearranging, the following form results:

$$\frac{S}{\mu} = \frac{K}{\mu_x} + \frac{S}{\mu_x} \quad (2)$$

This is known as the Hanes plot (Hanes, 1932). Letting $y = S/\mu$, $x = S$, $a = K/\mu_x$, and $b = 1/\mu_x$, yields the traditional linear equation:

$$y = a + bx \quad (3)$$

Values of a and b are easily fitted with measured values of S and μ. The values of the fitting coefficients K and μ_x are easily retransformed by

$$\mu_x = 1/b \quad (4)$$

and

$$K = a/b \quad (5)$$

A major problem with this regression analysis is that the correlation coefficient based on the regression of S/μ on S does not reflect the accuracy of predicted values of μ. In fact, the correlation coefficient is quite likely to be a highly inflated indicator of the accuracy of predictions of μ. It likely indicates only that S on the left-hand side is related to S on the right-hand side of Eq. (2).

Adsorption modeling: the following is another example of linearization and typical regression calculations. The following kinetic rate equation is used for adsorption studies with the boundary condition that $q_t = 0$ at $t = 0$ (Lagergren, 1898):

$$\frac{dq_t}{dt} = k(q_e - q_t)^2 \quad (6)$$

in which q_t is the is the mass of divalent metal ion on the surface of the sorbent at any time t, and k and q_e are constants with the former being the rate constant of sorption and the latter being mass of a divalent metal ion sorbed at equilibrium. The differential equation (6) has the following analytical solution:

$$q_t = \frac{t}{(1/kq_e^2) + (t/q_e)} = \frac{tq_e}{(1/kq_e) + t} \quad (7)$$

For a given set of (q_t, t) measurements, values of k and q_e are required. One approach to this and similar nonlinear analyses has been to linearize the solution of Eq. (7). One linear model for solving Eq. (7) is:

$$\frac{t}{q_t} = \frac{1}{kq_e^2} + \frac{1}{q_e}t \quad (8)$$

Letting $y = t/q_t$ and $x = t$, Eq. (8) has the linear model form $y = a + bx$, which is easily solved by least squares linear regression. Once the coefficients a and b are estimated with paired measurements (q_t, t), the values of k and q_e are estimated using the regression values of a and b. Generally, even if q_t and t are not highly correlated, t/q_t and t are very highly correlated because the t of the t/q_t ratio has a perfect correlation with t. Variation in q_t may keep the correlation coefficient from being exactly 1, but it is often very high because of the large variation in the sample values of t. The correlation coefficient for