THE REGRESSION METHOD IN ECOTOXICOLOGY

A BETTER FORMULATION USING THE GEOMETRIC MEAN FUNCTIONAL REGRESSION

By

Efraim Halfon

National Water Research Institute

Canada Centre for Inland Waters

Burlington, Ontario, Canada L7R 4A6

EXECUTIVE SUMMARY

The application of mathematical routines by applied scientists are often lacking in understanding of the underlying assumptions. This paper illustrates vividly the errors that can result in incorrect application of simple linear regression analysis when an assumption of an error-free independent variable is not fulfilled. A correct regression routine for this common case is proposed.

Résumé

Les scientifiques dans le domaine des sciences appliquées manquent souvent la comprehension des suppositions fondamentales des méthodes mathématiques. Cet article illustre les erreurs qui peuvent résulter de l'application inexacte de la méthode de la régression linéaire lors qu' on suppose incorrectement qu'une variable independente ne contient pas d'erreur. On propose une méthode correcte pour ce problème commun.

ABSTRACT

When both variables in a linear regression model are measured with errors, the geometric mean functional regression method should be used to compute the slope and intercept coefficients. Examples are taken from the literature and new equations are derived.

INTRODUCTION

The linear regression method commonly used in ecotoxicology to derive a regression line between two variables assumes that one variable is measured with error (the Y's) while the other is fixed (the X's). The regression equation is then used to find out what proportion of the variability in the Y's is due to the regression and what proportion to random errors of observations. In the literature (1, 2) many of these equations have been published during the years: examples may be the relations between Kow and bioconcentration and Kow and solubility, etc. What is not recognized is the fact that also the "fixed" X variables are usually measured with errors; for example for a given compound several measures of Kow are reported and they are often different (1). The method used to compute the coefficients of the linear regression should take this fact into account, i.e., that both the X's and the Y's are measured with error. This fact is important since the linear models are often used in practical applications and in modelling efforts. In the rest of the paper I suggest an alternative method to compute the regression coeffficients and suggest that this method should always be used when there is uncertainty in the measurement of the X's. The statistical method is not new (3) but as far as I know it has not been extensively used in ecotoxicology.

THE GEOMETRIC MEAN FUNCTIONAL REGRESSION METHOD

A statistical index often associated with paired ecotoxicological data is the correlation coefficient r. The correlation coefficient is computed as

$$r = \text{Covariance } (X,Y) \;/\; \text{sqrt } [\text{Variance } (X) \text{ Variance } (Y)]$$

which takes into account both the variability of the X's and the Y's and therefore the fact that both are measured with errors. These assumptions are the same in the geometric mean functional regression method.

The common way of computing the slope of a regression line $y = a + bx$ based on n observations is

$$b'' = \frac{Sxy}{Sx^2} \qquad (1)$$

where

$$Sxy = \sum xy - \frac{\sum x \sum y}{n}$$

and

$$Sx^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

The computation of the slope takes into account the covariance of the X's and the Y's assuming that the X's are measured with no error. When, however, the X's are measured with error, the slope b must be computed with the following formula:

$$b = \pm \text{sqrt} \left(\frac{\sum Sy^2}{\sum Sx^2} \right) \qquad (2)$$

where

$$Sy^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

or depending on what statistics are at hand:

$$b = \frac{b''}{|r|}$$

where b" is computed according to Equation 1 and r is the correlation coefficient. The sign of b is of course the same as the sign of b and r (4).

This formulation takes into account that both the X's and the Y's are measured with errors and do not depend on each other's presence: statistically they must be considered random variables and if a linear relationship is hypothesized, the regression analysis must be done accordingly. This functional regression minimizes both the vertical and horizontal distances of each point from the line. In the normal regression analysis, since the X's are considered fixed, only the distances for the Y's from the regression line are minimized using the least squares criterion.

EXAMPLES

Geyer et al. (1) presented some equations, computed by linear regression, to estimate the bioaccumulation of organic chemicals by the alga Chlorella using the n-octanol/water partition coefficient Kow. Geyer et al. realized that several Kow measures were available for the 41 compounds they used to compute a regression equation; therefore they included them in their Table 1 and Figure 4, but they used an average of the measured Kow as X's. As it is well known in statistics, to use means instead of raw data implies losing useful information; if they had derived the equation using all 70 data points, instead of 41, the resulting linear model would have been

$$\log BF_{1w} = 0.680 \log Kow + 0.171 \quad n = 70,\ r = 0.913 \qquad (3)$$

rather than

$$\log BF_{1w} = 0.681 \log Kow + 0.164 \quad n = 41, r = 0.902 \qquad (4)$$

where BF_{1w} is the bioaccumulation factor by Chlorella on a wet weight basis after one day and Kow is the partition coefficient of the chemical between n-octanol and water (1). The two equations are quite similar, but they do not take into account the fact that all Kow's are measured with error. With the geometric mean functional regression method the equation would have been

$$\log BF_{1w} = 0.737 \log Kow - 0.046 \quad n = 41, r = 0.902 \qquad (5)$$

using the 41 data points used by Geyer et al. (1), or better

$$\log BF_{1w} = 0.745 \log Kow - 0.053 \quad n = 70, r = 0.913 \qquad (6)$$

using all 70 data points. In this case the slope of Equation 6 is higher than those of Equations 4 and 5.

Mackay (2) recently reanalyzed data by Veith et al. (5) which show an obvious correlation between log Kow and log K_b. In his Table II Mackay reported 71 values stating in the text that he excluded 27 points because of data inconsistency. Figure 1 in the

same paper, however, shows that he used 51 points and excluded 20. I therefore could not reproduce his calculations to obtain the linear model

$$\log K_B = \log Kow - 1.32 \quad n = ??, r = 0.975 \qquad (7)$$

even if my calculations show a similar

$$\log K_B = 0.978 \log Kow - 1.239 \quad n = 44, r = 0.977 \qquad (8)$$

Mackay claims that a slope of one should be forced between log K_B and log Kow because of theoretical considerations. This forcing is not necessary if the geometric mean method is used to compute the linear model. In this case the model is

$$\log K_B = 1.000 \log Kow - 1.336. \quad n = 44, r = 0.977 \qquad (9)$$

and no a priori constraints are necessary to compute a higher and theoretically correct slope from the data for the regression line

DISCUSSION

When the relationships between chemical characteristics of contaminants and biological properties such as bioaccumulation are studied with statistical methods, often the correlation coefficient r

and the linear regression method are used. The correlation coefficient takes into account the variability of both the X's and the Y's but often ecotoxicologists use the regression method (Equation 1) without realizing that the independent variables, the X's, are also often measured with errors. In this paper I have suggested that the geometric mean functional regression method (Equation 2) should be used instead when calculating the coefficients of linear models. This usage is particularly important when regression results are extrapolated to new compounds and when the linear models are used for predictive purposes. Given the fact that the geometric mean line takes into account errors in the X's and Y's, the inverse equation can also be calculated easily, since b of X on Y is equal to 1/b, where b is the slope of Y on X. When the normal method of computing the regression slope is used (Equation 1), $b' = r^2/b$, thus the two slopes are not the inverse of each other.

Mackay (pers. comm. May, 1984) noted that, in some instances, theoretical calculations show that the slope of a regression line between two variables, such as solubility and Kow and Kow and K_B, should be one, while published regression lines are always less than one. However, as shown in Equation 9, the correct use of a regression formula allows the derivation of the theoretical expected slopes without any a priori constraints.

ACKNOWLEDGEMENTS

I am indebted to Drs. J.M. Ribo and D. Mackay for fruitful discussions and comments.

REFERENCES

1. Geyer, H., Politzki, G., Freitag, D. 1984. Chemosphere. 13, 269.

2. Mackay, D. 1982. Environ. Sci. Technol. 16, 274.

3. Teisser, G. 1948. Biometrics. 4, 14.

4. Richer, W.E. 1973. J. Fish. Res. Board Can., 30, 409.

5. Veith, G.D., DeFoe, D.L., Bergstedt, B.V., 1979. J. Fish. Res. Board Can., 36, 1040.