Environmental Toxicology and Chemistry, Vol. 12, pp. 1129–1133, 1993
Printed in the USA. Pergamon Press Ltd.

0730-7268/93 $6.00 + .00

001728

Short Communication

REGRESSION ANALYSIS OF LOG-TRANSFORMED DATA: STATISTICAL BIAS AND ITS CORRECTION

MICHAEL C. NEWMAN
University of Georgia, Savannah River Ecology Laboratory,
P.O. Drawer E, Aiken, South Carolina 29801

(*Received* 4 *June* 1992; *Accepted* 29 *September* 1992)

Abstract – Power and exponential models are used frequently in environmental chemistry and toxicology. Such models can generate biased predictions if derived with least-squares, linear regression of log-transformed variables. An easily calculated but seldom used estimate of bias can enhance the accuracy of subsequent predictions. This prediction bias and means of correcting it are presented, along with several examples.

Keywords – Statistics   Regression   Bias   Log-transformed variables

INTRODUCTION

Power and exponential relationships are common in most quantitative disciplines. In environmental chemistry and toxicology, predictive applications range from flow-related variation in water quality [1] to factors influencing toxicity [2,3]. The most frequently used method of fitting such data is least-squares, linear regression using logarithms of the X and Y variables (power relationships) or the Y variable (exponential relationships). This procedure involves four steps. First, the variables are transformed to their logarithms with base 10 or e. Second, the variables are fit using least-squares, linear regression methods. Next, the correlation coefficient (r) and a plot of regression residuals vs. the independent variable (X or $\log X$) may be used to judge the adequacy of model fit to the data. Finally, the linear model is transformed back to the original arithmetic units.

The resulting power ($Y = mX^{b}$) or exponential ($Y = m\ 10^{bX}$) model may then be used to predict values of Y given X. However, inherent in the steps described above is a bias that detracts from the accuracy of associated predictions. This bias and means of minimizing its influence have been discussed elsewhere [1,3–6]; however, it remains ignored in most studies. There may be two reasons for this oversight in the fields of environmental chemistry and toxicology. First, if the immediate goals of the treatment did not include prediction, then the bias correction would be irrelevant. Unfortunately, many such published models are used by later workers for prediction in modeling or risk assessment activities. Alternatively, the bias may remain uncorrected because a general, straightforward statement of its prevalence and potential influence in environmental sciences has not been developed to date. The purpose of this paper is to provide such an assessment. The logic is identical to that of earlier, more restricted discussions [3,6]. However, the prevalence of the bias will be emphasized rather than specific application of bias correction. Previous discussions are also expanded to include prediction bias in exponential relationships.

THE PROBLEM

Power relationships

Conforming to the notation of Neter et al. [7], the regression model used to describe power relationships is

$$\log Y = \beta_0 + \beta_1 \log X + \epsilon \qquad (1)$$

where

β_0 = the regression intercept estimated by b_0
β_1 = the regression slope estimated by b_1
ϵ = the random error term.

Let ϵ_i represent the error term associated with the ith data pair (X_i, Y_i). Then the mean expected value of ϵ for any data pair, $E(\epsilon_i)$ is zero with a variance of σ_i^2. Variances of the error terms associated with all pairs of data are assumed to be equal, that is, $\sigma_i^2 = \sigma^2$.

Regression models using logarithmic transforms of variables are usually back-transformed to the following power model:

$$Y = b_{0a} X^{b_1} \tag{2}$$

where b_{0a} = the antilog of b_0.

For predictive purposes, Equation 2 is incomplete, as the transform of the error term has been omitted. This oversight is understandable as the error term does not appear to be incorporated when making similar predictions with least-squares, linear regression models involving untransformed variables. But, as mentioned previously, the mean of the ϵ_i terms is zero in such a model. In the regression employing transformed variables, the ϵ_i values have a mean of zero in logarithmic units but not in the original arithmetic units. Because the mean will not be zero after back-transformation, the error term must be retained during the back-transformation:

$$Y = b_{0a} X^{b_1} 10^{\epsilon} \tag{3}$$

Unless there is no error ($10^{\epsilon} = 1$), values of Y predicted from the back-transformed model (Eqn. 2) will be biased by the quantity 10^{ϵ}.

Exponential relationships

The exponential relationship can be written in terms similar to those used for the power relationship above.

$$\log Y = b_0 + b_1 X + \epsilon \tag{4}$$

Similar to the discussion associated with the presentation of Equation 3 for power relationships, unbiased estimates for exponential relationships can be obtained with the following equation:

$$Y = b_{0a}\, 10^{b_1 X}\, 10^{\epsilon} \tag{5}$$

If the natural logarithms were used then the relationship would be the following:

$$Y = b_{0a} e^{b_1 X} e^{\epsilon} \tag{6}$$

BIAS CORRECTION

Estimation of 10^{ϵ} (or e^{ϵ}) is all that is required to account for the bias in predictions from power (Eqn. 3) or exponential (Eqns. 5 or 6) relationships fit by the process described above. Two approaches are applicable [1,3–6]. If the regression residuals were normally distributed then the following estimate could be used (the base in Eqn. 7 would be e if the natural logarithms were used):

$$10^{\epsilon} = 10^{\mathrm{MSE}/2} \tag{7}$$

where MSE = the mean square of the error from the regression.

$$\mathrm{MSE} = \frac{\sum_{i=1}^{N} e_i^2}{N-2} \tag{8}$$

where

e_i^2 = regression residual from the ith data pair squared

N = the total number of pairs.

If the residuals were not normally distributed, then the "smearing estimate of bias" [8] would be recommended to determine the prediction bias (if the natural logarithm were used then the base in Eqn. 9 would become e, not 10):

$$10^{\epsilon} = \frac{\sum_{i=1}^{N} 10^{e_i}}{N} \tag{9}$$

where e_i = the ith regression residual.

Regardless of the normality of residuals or the type of relationship, a relatively straightforward estimation of bias is obtained. Predicted values are then obtained from Equations 3 or 5 by using estimates of 10^{ϵ} from Equation 7 or 9.

PERVASIVENESS OF TRANSFORMATION BIAS

The potential for transformation bias is high in environmental chemistry and toxicology. Table 1 presents selected publications using log–log or logarithmetic transformations. It is important to note that the intentions in many of the cited publications were to provide data description, not prediction. The publications were selected to demonstrate the pervasiveness of power and exponential relationships in environmental sciences, not the correctness of the cited work.

Regardless of the original intent, many power and exponential relationships derived by linear regression on transformed variables are eventually employed for predictive purposes. If insufficient information to estimate the bias were present in the original publication, the possibility of inaccurate prediction would be increased and the seriousness

Table 1. Selected examples from the literature illustrating the pervasive use of log-log and log-arithmetic transformations to describe power and exponential relationships, respectively

Relationship	Y	X	Ref.
Power			
Water quality	Conductivity Ionic proportions Sediment load	Average daily stream flow	[1]
Bioaccumulation	Metal body burden	Animal wt.	[6,13,14]
	Zinc in gills	Fish wt.	[15]
	Radiocesium concn.	Oxygen consumption	[16]
	Strontium BCF[a]	Calcium concn.	[17]
	BCF	Octanol/water partition coefficient (K_{ow})	[11,18,19]
	Hydrophobic chemical elimination	K_{ow}	[20]
	Zinc elimination and uptake	Animal wt.	[21]
	Food consumption rate	Animal wt.	[22]
	Copper accumulation rate	Seawater copper concn.	[23]
Trophic transfer	Radiocesium in consumer	Radiocesium in food	[24]
	Cadmium or copper in consumer	Cadmium or copper in food	[25]
Metabolism	Liver microsomal monooxygenase activity	Animal wt.	[2]
Sublethal effect	Larval protein content	RNA/DNA upon toxicant exposure	[26]
Toxicity	LC50	Liver microsomal monooxygenase activity	[2]
	LC50 of water-column species	LC50 of benthic species	[27]
	IC50 of bacteria	IC50 of standard species	[28]
	Total residual chlorine	Duration of survival	[29]
	Methoxychlor LC50	Exposure duration	[30]
	LC50 or LD50	Animal wt.	[31]
	LC50 of metals	Water hardness	[32]
Exponential			
Water quality	Ionic proportion	Average daily stream flow	[1]
Elimination	Proportion of radionuclide retained	Clearance time	[33,34]
	General exponential clearance	Clearance time	[35,36]
Toxicity	LC50 of free copper	pH	[37]
	LC50 of pentachlorophenol	Reciprocal of time	[38]
	LC50 of di-, triorganotin	Hansch π parameter	[39]
	Median resistance time	Oxygen concn.	[40]
	Median survival time during zinc exposure	Temperature	[41]

[a]Bioconcentration factor.

of the bias would remain undefined. If the bias were small, it might still have serious consequences in modeling efforts employing iterative methods. The small bias may be compounded such that the predicted outcome becomes worse as the simulation progresses.

SELECTED EXAMPLES

Influence of hardness on toxic impact

Prediction bias in back-transformed models from linear regressions of log toxicity vs. log hardness data may be significant. Newman [3] used data from U.S. and Canadian water-quality-criteria documents relating copper, cadmium, and zinc toxicity to water hardness to demonstrate this point. The bias in the selected cases ranged from 2% to an extreme of 57%. The biased prediction was as much as 57% higher than an unbiased estimate of the effect concentration.

Elimination rate constant estimation

Cutshall [9] measured ^{65}Zn elimination from oysters taken from below a nuclear facility. Data visually extracted from Figure 1a of his paper were used to demonstrate predictive bias in routine elimination kinetics techniques. Time and the natural logarithm of the ^{65}Zn activity were used as the independent and dependent variables, respectively, in linear regression.

The antilog of the Y intercept of such a model is routinely interpreted as predicting the concentration (or amount) of material in the organism at time = 0. (In multiexponential compartment models, antilogs of several predicted Y intercepts may be used to estimate additional parameters [10].) However, such predictions are biased for reasons described above. In this example, the extracted data had an MSE of 0.125. Regression residuals appeared to be normally distributed. The bias was estimated to be $e^{0.125/2}$ or 1.06, approximately 6%.

Bioconcentration factor prediction

Table 4 of Neely et al. [11] lists data pairs of bioconcentration factors (BCFs) and octanol/water partition coefficients (K_{ow}) for eight organic chemicals. Linear regression resulted in the model, log BCF = 0.542 log K_{ow} + 0.124. The MSE for the regression was 0.1173. If this model were back-transformed for predictive purposes, the bias would have been estimated to be $10^{0.1173/2}$ or 1.14. BCFs predicted from the back-transformed model would have been biased by 14%.

CONCLUSION

A predictive bias is associated with back-transformed power and exponential models derived using least-squares, linear regression of log-transformed data. The bias is easily estimated using Equations 7 or 9 and, consequently, should be corrected. Bias estimation should be made regardless of the original intent of the workers generating such relationships. Alternatively, sufficient information should be presented so that an estimation can be made by future users.

A statement concerning the normality of the regression residuals should also be included. As discussed in Newman and Heagler [6], the assumption of residual normality can be examined with the Kolmogorov D statistic or Shapiro-Wilk W statistic as implemented in the SAS® statistical package [12]. The MSE is sufficient if the regression residuals are normally distributed. The $\sum e_i^2$ and number of data pairs are needed if the residuals are not normally distributed.

Acknowledgements—This work was supported by contract DE-AC09-76SROO-819 between the U.S. Department of Energy and the University of Georgia. The author is grateful to Dr. John E. Pinder III, who first brought this bias to his attention; to Dr. Carl Strojan for reviewing early versions of the manuscript; and to Ms. Rose Jagoe, who provided technical support during manuscript preparation.

REFERENCES

1. **Koch, R.W.** and **G.M. Smillie.** 1986. Bias in hydrological prediction using log-transformed regression models. *Water Resour. Bull.* **22**:717–723.
2. **Walker, C.H.** 1978. Species differences in microsomal monooxygenase activity and their relationship to biological half-lives. *Drug Metab. Rev.* **7**:295–323.
3. **Newman, M.C.** 1991. A statistical bias in the derivation of hardness-dependent metals criteria. *Environ. Toxicol. Chem.* **10**:1295–1297.
4. **Beauchamp, J.J.** and **J.J. Olson.** 1973. Corrections for bias in regression estimates after logarithmic transformation. *Ecology* **54**:1403–1407.
5. **Sprugel, D.G.** 1983. Correcting for bias in log-transformed allometric equations. *Ecology* **64**:209–210.
6. **Newman, M.C.** and **M.G. Heagler.** 1991. Allometry of metal bioaccumulation and toxicity. In M.C. Newman and A.W. McIntosh, eds., *Metal Ecotoxicology. Concepts and Applications.* Lewis, Chelsea, MI, pp. 91–130.
7. **Neter, J., W. Wasserman** and **M.H. Kutner.** 1990. *Applied Regression Statistical Models. Regression, Analysis of Variance, and Experimental Designs.* Richard D. Irwin, Homewood, IL.
8. **Duan, N.** 1983. Smearing estimate: A nonparametric retransformation method. *J. Am. Stat. Assoc.* **78**:605–610.
9. **Cutshall, N.** 1974. Turnover of zinc-65 in oysters. *Health Phys.* **26**:327–331.
10. **Barron, M.G., G.R. Stehly** and **W.L. Hayton.** 1990. Pharmacokinetic modeling in aquatic animals. I. Models and concepts. *Aquat. Toxicol. (Amst.)* **17**: 187–212.
11. **Neely, W.B., D.R. Branson** and **G.E. Blau.** 1974. Partition coefficient to measure bioconcentration potential of organic chemicals in fish. *Environ. Sci. Technol.* **13**:1113–1115.
12. **SAS Institute.** 1988. *SAS® Procedures Guide. Release 6.03 Edition.* Cary, NC.
13. **Boyden, C.R.** 1974. Trace element content and body size in molluscs. *Nature (Lond.)* **251**:311–314.
14. **Boyden, C.R.** 1977. Effect of size upon metal content of shellfish. *J. Mar. Biol. Assoc. (U.K.)* **57**:675–714.
15. **Bradley, R.W.** and **J.B. Sprague.** 1985. Accumulation of zinc by rainbow trout as influenced by pH, water hardness and fish size. *Environ. Toxicol. Chem.* **4**: 685–694.
16. **Reichle, D.E.** 1967. Relation of body size to food intake, oxygen consumption, and trace element metabolism in forest floor arthropods. *Ecology* **49**:538–541.
17. **Blaylock, B.G.** 1982. Radionuclide data base available for bioaccumulation factors for freshwater biota. *Nucl. Saf.* **23**:427–437.
18. **Geyer, H., P. Sheehan, D. Kotzias, D. Freitag** and **F. Korte.** 1982. Prediction of ecotoxicological behavior of chemicals: Relationship between physico-chemical

properties and bioaccumulation of organic chemicals in the mussel, *Mytilus edulis*. *Chemosphere* **11**:1121-1134.

19. **Thomann, R.V.** 1989. Bioaccumulation model of organic chemical distribution in aquatic food chains. *Environ. Sci. Technol.* **23**:699-707.
20. **Connel, D.W.** and **D.W. Hawker.** 1988. Use of polynomial expressions to describe the bioconcentration of hydrophobic chemicals by fish. *Ecotoxicol. Environ. Saf.* **16**:242-257.
21. **Newman, M.C.** and **S.V. Mitz.** 1988. Size dependence of zinc elimination and uptake from water by mosquitofish *Gambusia affinis* (Baird and Girard). *Aquat. Toxicol. (Amst.)* **12**:17-32.
22. **Newman, M.C.** and **D.K. Doubet.** 1989. Size-dependence of mercury (II) accumulation kinetics in the mosquitofish, *Gambusia affinis* (Baird and Girard). *Arch. Environ. Contam. Toxicol.* **18**:819-825.
23. **Martin, J.L.M.** 1979. Schema of lethal action of copper on mussels. *Bull. Environ. Contam. Toxicol.* **21**: 808-814.
24. **Reichle, D.E.** and **R.I. Van Hook, Jr.** 1970. Radionuclide dynamics in insect food chains. *Manit. Entomol.* **4**:22-32.
25. **Lasknowski, R.** 1991. Are the top carnivores endangered by heavy metal biomagnification? *Oikos* **60**: 387-390.
26. **Barron, M.G.** and **I.R. Adelman.** 1984. Nucleic acid, protein content, and growth of larval fish sublethally exposed to various toxicants. *Can. J. Fish. Aquat. Sci.* **41**:141-150.
27. **Di Toro, D.M., C.S. Zarba, D.J. Hansen, W.J. Berry, R.C. Swartz, C.E. Cowan, S.P. Pavlou, H.E. Allen, N.A. Thomas** and **P.R. Paquin.** 1991. Technical basis for establishing sediment quality criteria for nonionic organic chemicals using equilibrium partitioning. *Environ. Toxicol. Chem.* **10**:1541-1583.
28. **Blum, D.J.W.** and **R.E. Speece.** 1991. A database of chemical toxicity to environmental bacteria and its use in interspecies comparisons and correlations. *J. Water Pollut. Control Fed.* **63**:198-207.
29. **Wang, M.P.** and **S.A. Hanson.** 1985. The acute toxicity of chlorine on freshwater organisms: Time-concentration relationships of constant and intermittent exposures. In R.C. Bahner and D.J. Hansen, eds., *Aquatic Toxicology and Hazard Assessment: Eighth Symposium.* STP 891. American Society for Testing and Materials, Philadelphia, PA, pp. 213-232.
30. **Heming, T.A., A. Sharma** and **Y. Kumar.** 1989. Time-toxicity relationships in fish exposed to the organochlorine pesticide methoxychlor. *Environ. Toxicol. Chem.* **8**:923-932.
31. **Anderson, P.D.** and **L.J. Weber.** 1975. Toxic response as a quantitative function of body size. *Toxicol. Appl. Pharmacol.* **33**:471-483.
32. **Brown, V.M.** 1968. The calculation of the acute toxicity of mixtures of poisons to rainbow trout. *Water Res.* **2**:723-733.
33. **Beasley, T.M., H.V. Lorz** and **D.L. Zahnle.** 1986. The biokinetic behavior of ^{95m}Tc in juvenile coho salmon, *Oncorhynchus kisutch* (Walbaum). *Mar. Environ. Res.* **19**:259-278.
34. **Gallegos, A.F.** and **F.W. Whicker.** 1971. Radiocesium retention by rainbow trout as affected by temperature and weight. *Proceedings,* Third National Symposium on Radioecology. CONF-710501. Oak Ridge, TN, May 10-12, pp. 361-371.
35. **Atkins, G.L.** 1969. *Multicompartment Models for Biological Systems.* Meuthuen, London, UK.
36. **Greenblatt, D.J.** and **R.I. Shader.** 1985. *Pharmacokinetics in Clinical Practice.* W.B. Saunders, Philadelphia, PA.
37. **Borgmann, U.** 1983. Metal speciation and toxicity of free metal ions to aquatic biota. In J.O. Nriagu, ed., *Aquatic Toxicology.* John Wiley & Sons, New York, NY, pp. 47-71.
38. **Adelman, I.R., L.L. Smith, Jr.** and **G.D. Siesennop.** 1976. Effect of size or age of goldfish and fathead minnows on use of pentachlorophenol as a reference toxicant. *Water Res.* **10**:685-687.
39. **Laughlin, R.B., Jr., R.B. Johannesen, W. French, H. Guard** and **F.E. Brinckman.** 1985. Structure-activity relationships for organotin compounds. *Environ. Toxicol. Chem.* **4**:343-351.
40. **Shepard, M.P.** 1955. Resistance and tolerance of young speckled trout (*Salvelinus fontinalis*) to oxygen lack, with special reference to low oxygen acclimation. *J. Fish. Res. Board Can.* **12**:387-446.
41. **Lloyd, R.** 1960. The toxicity of zinc sulphate to rainbow trout. *Ann. Appl. Biol.* **48**:84-94.