SPURIOUS CORRELATION AND SOCIAL SCIENCE RESEARCH: THE EFFECT OF USING RATIO VARIABLES*

James E. Prather, Georgia State University

The theoretical and analytical treatment of what Pearson (1897) called "spurious correlation" has import in and of itself. However, to illustrate its impact upon actual research applications is the intent of this paper. Lacking empirical antecedents showing the practical implications of spurious correlation, the mathematics of spurious correlation is but an academic exercise. This paper presents examples which clearly illustrate that research models which contain deflated variables (whether called percentages, ratios or per capita measures) are not inconsequential or transparent data transformations. It will be shown that deflation does change the assumptions about the model an investigator is using.

The first section of the paper contains a brief introduction of the mathematics of the spurious correlation, of which a detailed treatment may be found in Prather (1975). The next section consists of a collection of empirical examples of how ratio variables have been dealt with in the social sciences. Schuessler (1974) describes the way ratio variables are used in social research, but this paper aims at looking beyond correlation of variables pairs, toward analyzing the model specification process, i.e., the role of ratio variables in the structure of the analytic approaches often employed by social scientists.

CALCULUS OF SPURIOUS CORRELATION

The "spurious" correlation question, as developed by Pearson (1897), was that, given

$a = X/Z$

$b = Y/Z$

and even if X, Y, and Z are stochastically independent, then,

$E(r_{ab}) \neq 0.$

This non-zero expected correlation has troubled observors in the decades since Pearson's exposition, and the practical import of spurious correlation has often been debated without satisfactory resolution.

It is the thesis of this paper that spurious correlation is a result of poor statistical model construction. The specification of linear models to include nonlinear, multiplicative variables expressed as ratios, percentages, or per capita measures is often not the product of theoretical demands, but rather a lack of appreciation for the purpose of the linear model. Thus, the remainder of this section focuses upon the deflation question from the least squares (regression) perspective.

The use of analysis models with deflated variables is found throughout social science literature. The models most frequently employed take a form similar to the following:

$$Y/Z = a + b_1 X_1/Z + \ldots + b_n X_n/Z + e/Z \quad (1)$$

where Y, X_1, . . ., X_n and Z are fixed measures and e is random. If equation (1) is expressed in its undeflated form, then:

$$Y = a Z + b_1 X_1 + \ldots + b_n X_n + u. \quad (2)$$

It is noted that equation (2) lacks a constant term. If the aim is to correct for a disturbance term with a multiplicative factor, i.e., $e = u/Z$, then the deflation of (2) to yield (1) will result in best, linear unbiased estimators (BLUE), given that the other assumptions of the linear model are met (Johnston, 1972, p. 122). If the problem is that Var $(u) = \sigma^2 Z^2$ then the more reasonable deflated equation would be,

$$Y/Z = A/Z + b_1 X_1/Z + \ldots + b_n X_n/Z + e/Z. \quad (3)$$

This specification is given by Johnston (1972, p. 216) to correct for heteroscedastic disturbances. Belsley (1974) has noted that the careless specification of the constant term, such as found in (1) can lead to inefficient, if not biased, estimators. Also, if the measures Y, X's and Z are stochastically independent, the b's will not have expected values of zero, i.e., $E(b_1), \ldots, E(b_n) \neq 0$. However, in equation (2) the null values for the b's would be zero (Prather, 1975).

Deflation may be viewed as an application of generalized least squares (GLS) or, more specifically, the type of GLS labelled weighted least squares. Equation (3) is in the form of GLS, where the deflation is for purposes of developing efficient estimators for equation (2). Belsley (1972) argues that if Var $(u) \neq \sigma^2 Z^2$, then deflation is not appropriate in that it will result in inefficient estimators if Z is fixed, and if Z is random, then spurious (bias) correlation is likely to be introduced.

There is a debate in the social sciences as to whether equation (1) is a structural model (a causal model), or just "fitting" an equation. The structural equation debate about ratio measures is a long standing one that has generated more emotional reaction than enlightenment. (Freeman and Kronenfeld, 1973; Fugitt and Lieberson, 1974; Schuessler, 1974). It is hoped that the following examples will demonstrate that deflation used for theoretical reasons is not as straightforward as some have maintained.

EXAMPLES

Professor Neyman has, for several decades, warned researchers of the pitfalls of ratio variables:

> In more modern times, spurious methods of studying correlations were involved in a great variety of empirical research; in astronomy, in farm economics, in biology, the study of elasticity of demand, in the problems of drunkenness and crime, of railroad traffic, and of racial segregation. On occasion, they were used in arguments about public policy matters. This applies to the health-pollution literature (Neyman, 1973, p. 31).

The following examples are presented with the aim of noting how deflation impacts on model specification itself. The implicit goal of each of the following examples is to increase the cumulative knowledge of these fields, freeing them from methodological questions of spurious correlation which interfere with the substantive pursuits of the discipline.[1]

Interstate Commerce Commission and Railroad Cost Studies. A case where deflation had policy implications may be found in the railroad cost studies performed for the Interstate Commerce Commission (ICC). Neyman (1952) made note of these data describing them as a possible example of deflated variables that could be misspecified. Neyman critiqued (pp. 151-154) the use of constructing railroad cost models with miles-of-track as a deflator. He expressed amazement that no one objected to this model.

It was not until 1972 that Griliches published a critique of deflated equations in cost studies. This example emphasizes the importance of the deflator used in a railroad cost model. Griliches (p. 29) pointed out that "it is very difficult to proceed to a discussion of the correct measurement of 'percent variable' unless it is possible to agree on which percent variable one is interested in--which 'average,' for what railroads, and at what traffic density."

The Griliches analysis reviewed the Cost Section's model, which contained these components: total costs (C), total gross ton-miles (X), and miles of track (M), the deflator. Griliches commented that, "to use deflation to eliminate the size component, one must assume that miles of road are in fact the relevant size measure and that the cost relationship is homogeneous in output and size, i.e., that there are no costs which are independent of size" (p. 31). The cost section used this model:

$$C/M = a + bX/M. \quad (4)$$

Griliches noted, however, that this assumed the undeflated model to be

$$C = aM + bX. \quad (5)$$

But Griliches proposed this model:

$$C = aM + bX + c, \quad (6)$$

and by deflating:

$$C/M = a + bX/M + c/M. \quad (7)$$

"The assumptions underlying the Cost Section procedure imply that the coefficient a is 'significant' while the coefficient c is insignificantly different from zero" (p. 32).

Griliches then fitted the equations using data from 97 railroads with the variables which are averages for the 1957-1968 period. The standard errors of the estimators for M in equations (6) and (7) showed that since the estimator for M is "insignificant," there is no evidence that M belongs in the equation. This is because the total miles of road variable is a poor measure of the "size" of the railroad. Obviously, division by an irrelevant variable is unneeded. Griliches then considers the problem of using deflation for stabilizing transformations. "Since the argument for any deflation must be made on efficiency grounds, it is not unreasonable, other things being equal, to prefer that procedure which yields the highest precision (lowest standard error) in estimating the parameter of interest (percent variable)" (p. 34).

Crime and Arrest Data Expressed in Ratios.

Spurious correlation has involved not just simple mathematical relationships, but it also has involved the foundations of some researchers' arduous work, combined with their personal perceptions. An illustration of conceptual needs exceeding a linear model is found in Logan (1971a, 1971b, 1972) who studied the relationship between crime (C), population (P), and arrests (A), focusing upon the correlation $r_{C/P,A/C}$. Logan spent a whole chapter (pp. 102-116) justifying his correlation of ratio variables. The correlation $r_{C/P(A/C.C)}$, the Pearson formula, and a data simulation were employed. Logan wrote: "in summary, the correlations obtained between certainty of imprisonment and crime rate cannot be explained as spurious, indexical artifacts" (1971a, p. 111). It is interesting that Logan's model is expressed:

$$C/P = a + bZ/C + e.$$

Note that the variable of prime interest--crime--cannot be expressed in this model in a manner where it does not appear on both sides of the equation. If the Logan Model were given as the formulation,

$$C/P = a/C + bA/C,$$

then it could be reexpressed as,

$$C^2 + aP + bAP.$$

This is certainly not what Logan had conceptualized nor would he be likely to be willing to accept such a peculiar model.

Crime and Population Density. The consequences of careless deflation can have policy implications contrary to what appropriate analyses might imply. An example of this possibility is a recent newspaper headline--"Crime Linked to Population Density"--(Thornton, 1974), which reported a study by Kvalseth (1974), to the effect that, "the negative relationships between crime rates and population density established in the present study are highly conclusive and statistically significant as determined by both the multivariate regression analysis and simple correlation analysis" (Kvalseth, 1974, p. 31). A look at the model used might lead us to doubt his conclusion: $R/P = a + bA + cA/P + b_1 X_1/P + \ldots + b_n X_n/P + e$, where robbery incidents (R), square acres (A), total population (P) and X's represents additional exogenous variables (all deflated) and their estimators. The estimator b is always negative, and the estimator c is always positive. The correlation $r_{A,A/P}$ would be expected to have a high value which introduces the problem of multicollinearity, as the author noted (p. 34). This leads to inefficient estimator variance and possibly unstable estimators (Deegan, 1972). This may account for the significant t-test of the estimator c. Another model used by Kvalseth (1974, p. 17), reveals yet another deflation problem in which acreage in commercial use (C) is found:

$$R/P = a + b\ (A/P)\quad (C/A) + e.$$

That expression (A/P) (C/A) equals C/P is obvious, simply showing the R^2 of .73 with robbery is indicative of commercial use being related to robberies. However, Kvalseth claims that the results ". . . indicate again the substantial influence exerted on the robbery rate in an area by its population density . . ." (p. 17).

The general problem of how to conceptualize the role of density in theory building was reviewed by Lawrence (1974, p. 712):

> While it is difficult to draw firm conclusions from the existing meager

> body of data, the balance of evidence appears not to support any simple causal relationship between density and socio-or psychopathology.

Deflation in Economic Models.

> Applied econometricians have long been careless in the way they introduce deflators into linear models (Belsley, 1972, p. 923).

The economics discipline is as advanced as any social science discipline in quantitative applications to its models. The economics literature supplies a number of examples of how many economists deal with the problem of deflated data.

A recent exchange of journal communications provides an example. This controversy over deflation was originated by Sato (1971) who criticized an article by Vanek and Studenmund (1968) for "spurious correlations [which] arise because the equation is divided by a variable whose value is much dispersed" (Sato, 1971, p. 625). Studenmund (1974, p. 497) rebutted that 'sound' theoretical and econometric reasons exist that show that the equational forms . . . are not subject to spurious correlations." As a further defense, Studenmund cited, from Sato's work, a model in the form of equation (3):

$$Y/Z = a/Z + Y/Z + e/Z.$$

Studenmund (p. 497) comments on the Sato model: "As can be clearly seen, Sato's equation is--according to his own technique--just as subject to spurious correlation as are those of our article because both sides of his equation are divided by [Z]." Studenmund uses as his reference the Kuh and Meyer (1955) work.

Sato (1974) replied to Studenmund (1974) by defending his own model specification and reemphasizing that the Vanek and Studenmund (1968) effort was amiss because " . . . as the growth rate appears on both sides of the equation, it is subject to spurious correlations in a high degree" (p. 499). Further, Sato argues that the Incremental Capital-Output Ratio (referred to as Y/Z above "is a single variable in theory, [but] it is derived in fact as a ratio of the investment share to the growth rate" (p. 499). Sato (p. 500) also argues that Vanek and Studenmund have used a deflator that is "largely stochastic." Sato concluded that "I believe that my results do not suffer from spurious correlations in spite of the Studenmund argument to the contrary" (p. 502).

In summary, from this exchange, it is readily apparent that deflation per se was not causing spurious correlation. Rather, all the authors seemed to express little if any awareness of what deflation does to their models, although Sato cites Kuh and Meyer's (1955) observation that a random deflator is of more concern than one that is non-random.

Administrative Intensity. In the field of management, one area of study that has treated the problem of using ratios at length is the analysis of administrative intensity (Freeman and Kronenfeld, 1973, and Freeman, 1973, p. 761). The problem was first noted by Akers and Campbell (1970), and Freeman and Kronenfeld devoted an entire article to the problem, calling it one of "definitional dependency." The frequently used model is that of

$$A/T = a + bT + e,$$

with numbers of administrators (A) and total employees including administrators (T). Inflating by T, it becomes

$$A = aT + bT^2 + eT.$$

Akers and Campbell looked at the model as being

$$\ln(A) = a + b \ln(T) + e.$$

They used a sample of 197 national membership associations and compared the numbers of staff members with the numbers of total members in the associations. Their equation was (p. 246) as follows:

$$\ln(A) = 1.974 + .784 \ln(T)$$

with $R^2 = .686$. The authors noted that the elasticity (slope) of .784, "fall[s] short of a perfect proportional relationship between membership and staff size. Staff size increases with organizational growth but at a slightly decreasing rate" (p. 247). The undeflated data was expressed in this equation:

$$A = 9.619 + .0012T \quad R^2 = .692.$$

The undeflated model was in need of correction for heteroscedasticity, whether through log or deflation transformations. Freeman and Kronenfeld (p. 119) reviewed the use of the logarithm model, but doubted its conceptual rationale.

The work on administrative intensity by Evers, Bohlen and Warren (1976) is an interesting case where many types of transformations and ratios are used and they result in unexpected and most unappealing specifications. However, Freeman and Hannan (1975) analyzed organizational structure using a model in the form of equation (3) with GLS.

CONCLUSIONS

If "spurious" correlation were a methodological or technical problem, the debate concerning the subject would have ended decades ago. Yet observers, e.g., Sockloff (1976), continue to pursue a technical "solution," hoping to adjust the Pearson formula of 1897 to yield an "exact" measure of the correlation of ratio variables. These efforts might be categorized as statistical exercises which really cannot answer the underlying question of what is the "true" correlation.

The correlation of ratio variables is not really a "bivariate" correlation. There are three variables in non-linear combinations. The constant term of equation (1) must be accounted for in one's theoretical and conceptual formulation. That is, the problem is normative or conceptual, not statistical. The introduction of deflation into a model transforms the model, and it is necessary for the researcher to be aware of the implications of this transformation.

There are numerous researchers who would strenuously argue that ratio variables are "structural" measures, having genuine meaning as ratios. Does per capita income really exist as a causal variable? The linear model "fits" linear relationships. The implied multiplicative functions of ratio variables should be introduced into linear models only with the greatest caution. When theory requires ratio variables, it is advised that the constant term be carefully

specified and that goodness-of-fit as measured by R^2 be viewed with caution. Buse (1973) notes that the R^2 in GLS may be derived in several ways--all of which differ from the R^2 of OLS.

If one approach to data analysis is that of the "exploratory" mode (recommended by Tukey and others) where the objective is to "fit" equations using the linear model, then all variables formed by division or multiplication should be avoided for these reasons: (1) the constant term would not be altered; (2) each of the variables can be tested (estimated as to their functional relationships--using deflation assumes a proportional or non-linear relationship); and because (3) the coefficient of determination (R^2) is not subjected to the ambiguity resulting from correlating ratio variables.

If heteroscedasticity is thought to be present (or is tested for by Harvey's [1976] general procedure), then deflation may be a solution for achieving efficient estimators. However, the double-logarithm transformation deals with this source of estimation inefficiency, typically better than does deflation (Prather and Hutcheson, 1976).

* I wish to thank John D. Hutcheson, Jr., Georgia State University, for his helpful comments and criticism.

NOTES: 1) An interesting, but purely historical example from the psychology literature is found in the debate over the impact of spurious correlation on the IQ ratio. Relevant studies are Thomson and Pinter (1924), Douglass and Huffaker (1929), Jackson (1940), DuBois (1948), McNemar (1946, p. 136) and Guilford (1965, p. 351).

Examples of the ratio conundrum in applied natural science may be seen in Sutherland (1965) who appeared most confused on the question of ratio data in experimental research; and Katch and Katch (1974) attempted, erroneously, to use ratio variables as a form of partial correlation.

REFERENCES

Akers, R. L. and Campbell, F. L. 1970. "Size and Administrative Component in Occupational Association." Pacific Sociological Review. 13(Fall): 241-251.

Belsley, D. A. 1972. "Specification with Deflated Variables and Specious Spurious Correlation." 40(September): 923-927.

Belsley, D. A. 1974. "The Constant Term and Deviations About the Mean." American Economist. 18(Fall): 109-112.

Buse, A. 1973. "Goodness of Fit in Generalized Least Squares Estimation." American Statistician. 27(June): 106-108.

Deegan, J. 1972. The Effects of Multicollinearity and Specification Error on Models of Political Behavior. Unpublished Ph.D. Dissertation, University of Michigan.

Douglass, H. R. and Huffaker, C. L. 1929. "Correlation Between Intelligence Quotient and Accomplishment Quotient." Journal of Applied Psychology. 13: 76-80.

DuBois, P. H. 1948. "On the Statistics of Ratios." Paper presented at the Annual Meeting of the Psychometric Society.

Evers, F. T., Bohlen, J. M., and Warren, R. D., 1976. "The Relationship of Selected Size and Structure Indicators in Economic Organizations." Administrative Science Quarterly. 21(June): 326-342.

Freeman, J. H. 1973. "Environment, Technology, and Administrative Intensity of Manufacturing Organizations." American Sociological Review. 38(December): 750-763.

Freeman, J. H. and Hannan, M. T. 1975. "Growth and Decline Processes in Organizations." American Sociological Review. 40(April): 215-228.

Freeman, J. H. and Kronenfeld, J. E. 1973. "Problems of Definitional Dependency: The Case of Administrative Intensity." Social Forces. 52(September): 108-121.

Fuguitt, C. V. and Lieberson, S. 1974. "Correlation of Ratios or Difference Scores Having Common Terms." Sociological Methodology, 1973-1974, edited by H. Costner. San Francisco: Jossey-Bass.

Griliches, Z. 1972. "Cost Allocation in Railroad Regulation." Bell Journal of Economics and Management Science. 3(Spring): 26-41.

Guilford, J. P. 1965. Fundamental Statistics in Psychology and Education. (4th ed.) New York: McGraw-Hill.

Harvey, A. C. 1976. "Estimating Regression Models with Multiplicative Heteroscedasticity." Econometrica. 44(May): 461-465.

Jackson, R. W. B. 1940. "Some Pitfalls in the Statistical Analysis of Data Expressed in the Form of IQ Scores." Journal of Educational Psychology. 31: 677-685.

Johnston, J. 1972. Econometric Methods (2nd ed.). New York: McGraw-Hill.

Katch, V. L. and Katch, F. I. 1974. "Use of Weight-adjusted Oxygen Uptake Scores that Avoid Spurious Correlations." Research Quarterly. 45(December): 447-451.

Kuh, E. and Meyer, J. R. 1955. "Correlation and Regression Estimates When the Data are Ratios." Econometrica. 23(October): 400-416.

Kvalseth, T. O. 1974. "Statistical Models of Urban Crime." Paper at Joint National Meeting of the Operations Research Society of America and the Institute of Management Sciences, October. San Juan, Puerto Rico.

Logan, C. H. 1971a. Legal Sanctions and Deterrence from Crime. Ph.D. Dissertation, Indiana University.

Logan, C. H. 1971b. "On Punishment and Crime (Chiricos and Waodo, 1970): Some Methodological Commentary." Social Problems. 19(Winter): 280-284.

Logan, C. H. 1972. "General Deterrent Effects of Imprisonment." Social Forces. 51(September): 64-73.

McNemar, Q. 1949. Psychological Statistics. New York: Wiley.

Neyman, J. 1952. Lectures and Conferences on Mathematical Statistics and Probability (2nd Ed.). Washington, D. C.: U. S. Department of Agriculture.

Neyman, J. 1973. "Epilogue of the Health-Pollution Conference." Bulletin of the Atomic Scientists. 29(September): 25-34.

Pearson, K. 1897. "Mathematical Contributions to the Theory of Evolution--On a Form of Spurious Correlation Which May Arise When Indices are Used in the Measurement of Organs." Proceedings of the Royal Society of London. 70: 489-497.

Prather, J. E. 1975. The Methodology of Spurious Correlation: Implications for Public Policy Analysis. Unpublished Ph.D. Dissertation, Political Science Department, University of Georgia, Athens.

Prather, J. E. and Hutcheson, J. D. 1976. "A Critique of Aggregate-Level Urban Policy Research." Paper presented to the Annual Conference of the Public Choice Society, Roanoke, Virginia.

Sato, K. 1971. "International Variations in the Incremental Capital-Output Ratio." Economic Development and Cultural Change. 19(July): 621-640.

Sato, K. 1974. "Spurious Correlation and Incremental Capital-Output Ratio: Reply." Economic Development and Cultural Change. 22(April): 499-502.

Schuessler, K. 1974. "Analysis of Ratio Variables: Opportunities and Pitfalls." American Journal of Sociology. 80(September): 379-396.

Sockloff, A. L. 1976. "The Analysis of Nonlinearity Via Linear Regression with Polynominal and Product Variables: An Examination." Review of Educational Research. 46(Spring): 267-291.

Studenmund, A. H. 1974. "Spurious Correlation and Incremental Capital-Output Ratio." Economic Development and Cultural Change. 22(April): 496-498.

Sutherland, T. M. 1965. "The Correlation Between Feed Efficiency and Rate of Gain, A Ratio and Its Denominator." Biometrics. 21(September): 739-749.

Thomson, C. H. and Pinter, R. 1924. "Spurious Correlation and Relationship Between Tests." Journal of Educational Psychology. 15(October): 433-444.

Thornton, L. 1974. "Crime Linked to Population Density." Atlanta Constitution. (August, 26): 7-A.

Vanek, J. and Studenmund, A. H. 1968. "Towards a Better Understanding of the Incremental Capital-Output Ratio." Quarterly Journal of Economics. 82(August): 435-451.