ERIC M. USLANER
University of Maryland

*The Pitfalls of Per Capita**

The use of per capita measures in aggregate data analysis has been given little theoretical justification. This paper argues that unless there is such a justification, the employment of such transformations can lead to greatly distorted results. It is shown that such transformations can lead to either spuriously high or low correlations among sets of variables. Furthermore, there is no precise formula for determining the extent of any such biases. Even when the distortions are not great, they may lead the researcher to erroneous conclusions about the strength of the relationships among his variables. It is argued that per capita transformations are not always undesirable, but that greater attention needs to be paid to the theoretical justifications for employing such indices.

Aggregate data analysis is irritating. In addition to the standard problems of inference, there is the difficult task of choosing a set of specific indicators for a particular analysis. Do we employ raw data, medians, means, modes, percentages or proportions, standard scores, per capita figures, or "more sophisticated" indices–either of our own invention or commonly employed ones such as the Gini index of inequality? A debate on standard scores (Marquette, 1972; Levine, 1973) has recently appeared in these pages. And there has been increasing attention to problems raised by per capita measures (cf. Schuessler, 1973, 1974). Indeed, Fernando Cortés and Adam Przeworski (1971) have argued that least squares regression estimates employing per capita measures of both the independent and dependent variables are inconsistent; they argue that raw data rather than per capita figures should always be employed when using statistical techniques based upon the general linear model (regression analysis, factor analysis, analysis of variance and covariance, etc.).[1]

*The comments of Alan Agresti, Herbert B. Asher, Paul M. Cohen, Richard K. Scher, Ronald E. Weber, and an anonymous referee for this journal are gratefully acknowledged.

[1] The arguments made by Cortés and Przeworski are based upon the assumption that it is the transformation of variables which produces the inconsistent regression estimates. However, there are other problems which per capita measures produce which yield the same conclusions: (1) specification error; (2) the problem of making misleading "ecological" inferences from per capitized variables (Schuessler, 1973, p. 227); (3) aggregation bias (Theil, 1971, pp. 560–562; Kuh and Meyer, 1955, pp. 406–408); and (4) the probability of multicollinearity. Which of these problems "causes" the distortions depends upon how one looks at the entire dilemma. For other studies on per capita (or ratio) variables not cited elsewhere in this paper, see Pearson (1897), Neifeld (1927), Kunreuther (1966), and Tufte (1969–70).

In this article, I do not maintain that per capita measures should *never* be employed, but only that there are serious problems with such indices which must not be overlooked. We rarely see any theoretical justifications for the use of such measures. Rather, it is generally assumed that we want "comparable" measures across a set of observations. However, per capita indices, unlike standardized scores (Levine, 1973) are nonlinear transformations of the original data.[2] It is thus not surprising that difficulties arise when per capita measures are employed in techniques encompassed in the general linear model. By discussing (1) two extreme cases in which the relationship between sets of variables can easily be predicted; and (2) one example of "middle range" biases in which the strength of the relationship rather than the existence of a statistical linkage might be incorrectly assessed, I shall point to some of the potential pitfalls of employing per capita measures without prior theoretical justification. This procedure is inelegant. However, no direct proof of the existence or the magnitude of the distortions arising from per capita transformations is available without knowledge of the means and variances of all of the variables considered (Madansky, 1964, p. 654); and estimation of these moments for the per capitized variables can itself be problematic (cf. Schuessler, 1973, p. 221). Thus, an approach which is based upon the analysis of data seems the most fruitful way to proceed.

The Extreme Cases

Theoretically Unrelated Variables

The clearest case of a set of unrelated variables is a group of observations, artificially created, and designed to have no relationship to each other. I thus

[2] If X_i is a vector of observations, a linear transformation on X_i is:

$$Y_i = a + bX_i$$

where a and b are any real numbers. If a = 0 and $b = (1/\sum_i X_i)$, then we have the familiar transformation which yields proportions. Here, b is the same quantity for each case. The per capita transformation employs a different denominator, b_i, for each case since the transformation itself is defined as:

$$Y_i = (X_i/P_i)$$

where P_i is the population of the *i*th group. The transformation is thus not necessarily order-preserving among the X_i's.

TABLE 1

Off-Diagonal Zero-Order Correlations among Eleven Simulated Random Variables

Variable Number	2	3	4	5	6	7	8	9	10	11
1	.013	–.067	–.070	–.008	.036	–.044	–.067	–.008	.059	–.107
2		.011	.013	.002	–.076	.092	–.031	–.064	–.046	–.063
3			.032	.057	–.116	.075	–.080	.081	.035	.034
4				.038	–.098	.034	–.086	–.023	.040	.035
5					–.018	.094	–.026	–.010	.098	.085
6						–.100	.074	–.011	–.056	–.056
7							.048	–.048	–.194	–.009
8								.063	.044	.002
9									.031	.009
10										.093

generated a set of 11 independent, normally distributed simulated variables.[3] Ten were selected for the original analysis and one was selected as the "control" variable for the second analysis. In the latter analysis, the original variables (one to ten) were "per capitized" by the respective values for each case of the "control" variable. The 11 variables are almost entirely uncorrelated, as Table 1 indicates. A factor analysis employing varimax rotation of the original ten variables produced five dimensions with eigenvalues above 1.00, accounting for 57.8 percent of the total variance. There was no clear-cut pattern to the factors and none of the variables had a communality of .500 or higher. Thus, the relative independence of the data is apparent.

Now, consider what happens when we "per capitize" the original variables by our control variable. We now have a new set of variables, 12 through 21. Every one of the 45 off-diagonal upper-triangular correlations is now positive (cf. Schuessler, 1974, p. 386)–see Table 2–and only four coefficients fall below .50. A factor analysis of these variables–presented in Table 3–produces only one dimension, accounting for 71.9 percent of the total variance. Four of the ten variables load at .90 or higher (in absolute value) on the single dimension, and only one variable has a communality less than .500. *Out of chaos, we have produced order.* Clearly, what is happening is that we are

[3] The variables created each had 200 cases with a mean of .500 and a standard deviation of .28.

TABLE 2

Off-Diagonal Zero-Order Correlations among Ten Transformed Random Variables

Variable Number	13	14	15	16	17	18	19	20	21
12	.897	.609	.688	.678	.795	.722	.831	.874	.827
13		.539	.611	.763	.653	.823	.848	.870	.785
14			.759	.322	.701	.584	.588	.508	.458
15				.482	.767	.594	.695	.644	.548
16					.463	.690	.601	.699	.606
17						.549	.776	.765	.650
18							.744	.669	.565
19								.920	.732
20									.806

introducing some common variance into all of the variables by employing the "per capitizing" transformation (cf. Schuessler, 1973, p. 220).

Theoretically Related Variables

We have seen that nonlinear transformations can lead to spuriously high relationships. Now, I present an analysis which shows that a similar transformation can lead to conclusions that the relationship between variables posited to be strongly correlated is rather low. In this example, I move from factor analysis to regression analysis. While the "switch" may appear somewhat confusing, both techniques do employ the least squares criterion of minimizing error variance from a linear hypothesis. In particular, Karl Schuessler (1973, pp. 219–220) has argued that the two techniques are susceptible to similar problems when per capita transformations are employed.

The starting point for the analysis of a strong theoretical linkage between a pair of variables is Ira Sharkansky's (1968, pp. 39ff.) observation that the best predictor of this year's expenditures across the American states is each state's expenditure in the previous year(s). I shall return to this example below. However, consider an even more powerful predictor of this year's expenditures: current revenues. Since most states have rather strict limitations–if not outright prohibitions–on borrowing power, we would expect an almost perfect linear fit between the two variables. Employing per

TABLE 3

Factor Analysis of Variables
12 to 21

Variable	Loading	Communality
12	–.943	.889
13	–.923	.852
14	–.663	.440
15	–.763	.582
16	–.708	.501
17	–.818	.670
18	–.792	.623
19	–.919	.845
20	–.925	.855
21	–.807	.651

capita data for both variables for 1970, we obtain the following regression equation:

$$Y = 218.33 + .442X, r = .471, r^2 = .22,$$

where Y is expenditures and X is revenues. The value of the coefficient of determination is surprisingly low, and the regression coefficient of only .442 suggests that most state governments are doing rather poor jobs in dispersing their revenues.

The explanation for these anomalous results is found in examining the use of the per capita transformation. Regressing the total revenues for each state against the total expenditures, we obtain the equation:

$$Y = 1387123.755 + 1.006X,$$

with a correlation coefficient of .998. The equation accounts for almost all of the variance in the level of expenditures, as expected, and the regression coefficient of 1.006 indicates that some states can borrow at least a little. Thus, in contrast to the simulated data experiment with factor analysis, this use of the per capita transformation creates *chaos out of order.* Through these extreme cases, we have seen that the inference problems associated with per capita transformations are potentially quite damaging.

Middle Range Biases

Next, let us consider how less pronounced distortions can lead to erroneous conclusions about the strength of a relationship. In Sharkansky's analysis of budgetary incrementalism in the American states, he found a zero-order correlation of .85 between per capita state expenditure levels in 1957 and 1962. Then why did the analysis of per capita state expenditures presented above fare so poorly in comparison? I attempted to replicate Sharkansky's analysis and was partially successful in so doing, being limited by not having the total expenditures available in the Sharkansky data file nor the overall per capita expenditures. By merging this file with another, I was able to approximate Sharkansky's analysis for the 48 states quite closely.[4] In comparison to Sharkansky's correlation of .85 for per capita expenditures for 1957 and 1962, I obtained a coefficient of .81.

Now, correlations in the range of .80 are hardly to be sneered at. Yet, a transformation of the per capita expenditure figures into actual dollar amounts increases the correlation between 1957 and 1962 expenditures to .969, accounting for 94 percent of the variance in 1962 expenditures as contrasted to only 66 percent for the per capitized variables. In this example, however, the researcher is not likely to reach an erroneous conclusion about the existence of the relationship, but only about its strength. The question remains, however, why the Sharkansky results are less sensitive to the per capita transformation than the example employing 1970 data I provided above.

The answer cannot be found by noting that the Sharkansky analysis employs different denominators (1957 and 1962 populations, respectively) whereas the example I provided uses the same denominators (1970 population). The 1957 and 1962 population variables correlate at .996. Thus, the only possible explanation is that the patterns of covariation in the two examples are different. The 1970 per capita expenditures and per capita revenues correlate with 1970 population at −.200 and −.129 respectively. The correlations of the respective population figures with 1957 and 1962 per capita expenditures are −.110 and −.069. In the 1970 example, the relation-

[4] The Sharkansky data set is the Inter-University Consortium for Political Research's (ICPR) "Sharkansky American State Data, 1956–1963." The other data set is the ICPR's "Hofferbert Public Policy Data for the United States–1890–1960." The 1970 data come from the Comparative State Data File, data generated at Syracuse and Indiana Universities by Frank J. Munger and Ronald E. Weber. Neither the ICPR nor Munger and Weber are responsible for the arguments made in this paper, but I am grateful for the availability of their data.

ships between population size and the per capita measures are stronger for both the independent and dependent variables than they are for the Sharkansky example. Thus, the per capita measures for this "extreme" case are more a component of population sizes than those in Sharkansky's analysis. Furthermore, the simple correlation between 1957 and 1962 total expenditures is slightly lower than the coefficient between 1970 revenues and expenditures. Thus, the use of the per capita transformation for the Sharkansky data could not "rob" as much predictive power from the raw data as the 1970 revenue-expenditure linkage could. The extent of the deflation—or possible inflation—of the correlation coefficient by employing this nonlinear transformation is thus a complex (and not always identifiable) function of the intercorrelations of all of the variables under consideration.

Whither Per Capita?

Cortés and Przeworski (1971, p. 1) base their criticism of per capita transformations on the assumptions that the theoretically relevant variables are the total figures and that the transformation is used only to "filter out" the possible confounding effects of the size of the geographic unit. For a variable such as state expenditures, this criticism is very appropriate. State legislators generally deal in absolute dollar and cent terms when deciding the level of state expenditures. There is no evidence that the expenditures of other states serve as a baseline for any individual state.

Clearly, in such cases we should avoid the use of per capita transformations. However, the argument of Cortés and Przeworski that these measures should never be employed is too strong. If the theoretically meaningful variables are per capita ones, then employing raw data can lead to the same types of spurious inferences that incorrectly using per capita measures does. Thomas R. Dye's (1966, p. 82) measure of "educational effort" (the total public school expenditures as a percentage of the total personal income) to determine "a state's willingness to sacrifice personal income for public education" is a legitimate use of a nonlinear transformation, because the question posed requires comparisons across cases. Per capita variables are also warranted in determining expenditure levels for such programs as Title I of the Elementary and Secondary Education Act of 1965 because allocations are in part based upon per capita characteristics of the states (Congressional Quarterly, 1969, p. 710). Per capita measures should not be shunned, but theoretically justified.

Thus, per capita measures should be employed *only* when: (1) the theoretical concern of the researcher involves explicit comparisons among cases, involving such measures as "*relative* deprivation" of one group of people contrasted to that of another population; and (2) one is attempting to predict scores on a dependent variable and one has prior knowledge that the relevant independent variables actually employed by decisionmakers in determining public policies include standardized measures. Otherwise, attempts to provide "comparative" measures across cases may lead to misleading results. The solution to problems of standardization is better met either by the use of direct linear transformations on the original variables, where possible, or by examining order-preserving transformations (such as beta weights) on the estimated coefficients.

Manuscript submitted November 7, 1974.
Final manuscript received April 25, 1975.

REFERENCES

Congressional Quarterly. 1969. *Congress and the Nation*, vol. II.

Cortés, Fernando, and Przeworski, Adam. 1971. "Per Capita or Sin Capita: A Note of Caution," a paper prepared for delivery at the Sixty-Seventh Annual Meeting of the American Political Science Association, Chicago, Illinois, September 7–11, 1971.

Dye, Thomas R. 1966. *Politics, Economics, and the Public*. Chicago: Rand McNally.

Kuh, Edwin, and Meyer, John R. 1955. "Correlation and Regression Estimates When the Data Are Ratios," *Econometrica*, October 1955, pp. 400–416.

Kunreuther, Howard. 1966. "The Use of the Pearsonian Approximation in Comparing Deflated and Undeflated Regression Estimates," *Econometrica*, January 1966, pp. 232–234.

Levine, Mark S. 1973. "Standard Scores as Indices: The Pitfalls of Not Thinking It Through," *American Journal of Political Science*, May 1973, pp. 431–440.

Madansky, Albert. 1964. "Spurious Correlation Due to Deflating Variables," *Econometrica*, October 1964, pp. 652–655.

Marquette, Jesse F. 1972. "Standard Scores as Indices: The Pitfalls of Doing Things the Easy Way," *Midwest Journal of Political Science*, May 1972, pp. 278–286.

Neifeld, M. R. 1927. "A Study of Spurious Correlation," *Journal of the American Statistical Association*, September 1927, pp. 331–338.

Pearson, Karl R. 1897. "Mathematical Contributions to the Theory of Evolution: On A Form of Spurious Correlation When Indices Are Used in the Measurement of Organs," *Proceedings of the Royal Society of London*, pp. 489–498.

Schuessler, Karl. 1973. "Ratio Variables and Path Models," in Arthur S. Goldberger and Otis Dudley Duncan, eds., *Structural Equation Models in the Social Sciences.* New York: Seminar Press, pp. 201–228.

____. 1974. "Analysis of Ratio Variables: Opportunities and Pitfalls," *American Journal of Sociology,* September 1974, pp. 379–396.

Sharkansky, Ira. 1968. *Spending in the American States.* Chicago: Rand McNally.

Theil, Henri. 1971. *Principles of Econometrics.* New York: Wiley.

Tufte, Edward R. 1969–70. "A Note of Caution in Using Variables That Have Common Elements," *Public Opinion Quarterly*, Winter 1969–70, pp. 622–625.