THE EFFECT OF AGGREGATION ON THE INTERPRETATION OF R^2

James E. Prather, Georgia State University

The use of aggregate data in regression analysis is pervasive in such fields of study as public policy, demography, political science, economics, and sociology. For several decades, a debate on the proper specification of aggregate models, so that inferences could be made about micro-level relationships from the macro-level estimators, has permeated literature in these fields. For most investigators, this question remains unresolved or insoluble, though there have been continuous refinements of techniques designed to mitigate aggregation problems (Irwin & Lichtman, 1976; Smith, 1977).

This paper does not focus upon macro to micro inference directly, rather it is concerned with the interpretation of the standard measure of goodness-of-fit for regression analysis--the multiple-correlation-squared (R^2). The importance of the R^2 as a test statistic is the rationale for exploring its interpretation when using macro-level data for analyses employing least squares regression. However, it is acknowledged that it is not possible to divorce substantive problems of model formation from the methodological questions concerning technique. Thus, a review of previous work on aggregate allows one to view the question holistically, rather than as solely a problem of calculation or reading a computer printout.

The previous works on analyzing grouped data can for heuristic purposes be divided into two separate development paths. The two perspectives can be illustrated by the seminal work of Robinson (1950) in sociology and of Prais and Aitchison (1954) in economics. As has been previously noted, Robinson's "ecological correlation" approach and the grouping in linear models approach of Prais and Aitchison complement each other. A review of the aggregation issue from these two perspectives will be presented in the next two sections.

The importance of the R^2 is that it is often employed as a measure of the power and amount of explanatory worth of a particular specification. Even though this paper does not focus on model building, the use of R^2 in model selection with aggregate data does warrant considering specification impact on R^2.

Analysis of Covariance Approach to Aggregation

The analysis of variance method is illustrated by partitioning the sum of squares about the mean for Y (the dependent variable) into "explained" sums of squares and residual sum of squares. Following the notation of Johnston (1972:192-207) a simple model is defined as

$$y = X\beta + u \tag{1}$$

Where the sample y is a column vector (n x 1) of micro-level observations composed of p sub-vectors --i.e., the groups. The independent variables are the X matrix (n x k) divided into p groups and the first column is all ones to allow a constant term, while β is a vector (k x 1) of the estimators. The vector u contains stochastic noise values where $E(u)=0$. To incorporate the possible effect of the p groups, then an expanded model is

$$y = D\alpha + X\beta + u, \tag{2}$$

which allows the p groups to have different constant terms, thus α is a vector of (p - 1) elements. The D' matrix is of dummy variables with order (Mp x [p-1]), where $M = \sum_{i=1}^{p} m_i$; is the sum of the number of observations in each p, for instance:

$$D' = \begin{matrix} 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1 \end{matrix} \tag{3}$$

Remembering that D has p groups, with each p having m elements. To estimate (1) above, start with

$$y = X\hat{\beta} + s, \tag{4}$$

which can be estimated by

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{5}$$

where s gives the least square residuals. An additional relationship may be derived as,

$$y'y = \hat{\beta}'X'y + s's \tag{6}$$

Returning to (2) above, the estimation of

$$y = D\hat{\hat{\alpha}} + X\hat{\hat{\beta}} + e \tag{7}$$

becomes

$$\begin{matrix} \hat{\hat{\alpha}} \\ \hat{\hat{\beta}} \end{matrix} = \begin{matrix} D'D & D'X \\ X'D & X'X \end{matrix}^{-1} \begin{matrix} D'y \\ X'y \end{matrix} \tag{8}$$

and from (6) above

$$y'y = \hat{\alpha}'D'y + \hat{\hat{\beta}}'X'y + e'e. \tag{9}$$

The e vector contains residuals for (7).

To calculate the R^2 for analysis of the covariance problem, it is necessary to define (Thiel, 1971, p. 176)

$$1-R^2 = \frac{e'e}{y'Ay} \tag{10}$$

where

$$A = I - \frac{1}{N} V V' \tag{11}$$

with V a vector n ones. The A matrix is to transform to deviations from the mean.

If a standard analysis of covariance were desired, the terms given in Table 1 would be the appropriate residual sum of squares to use for an F-test after converting by degrees of freedom to determine mean squares. However, our interest is in the R^2's that would be associated with the differing levels. The micro-level R^2 is for the "Total" formula. Compare this to the macro-level or aggregate R^2 which has an additional factor of 'D'y -- indicating that the value of $\hat{\alpha}$ vector would inflate the R^2 to the extent that it is related to y. When $\hat{\bar{\alpha}}$ is vector of zeros or near zeros it could be concluded that the grouping factor had no independent effect on the dependent variables and the between groups R^2 would equal the total R^2. To restate the above, if the grouping is random, then the between groups R^2 is an unbiased estimate of the total R^2 -- though not as efficient as the total R^2 estimate (Cramer, 1964). The experimental statistician would note the treatment (i.e., groups) had no significant effect. There are undoubtedly many investigators using aggregated data whose research would be much easier if the grouping was random. Grunfeld and Griliches (1960) noted the phenomenon of the higher R^2 that was often found with grouped data and referred to it as a "synchronization" effect. As a historical note, Gehkel and Bichel (1934), Thorndike (1939), and Yule and Kendall (1950) observed the same problem. At the time there was no clear explanation except the intuitive one that "grouping" on substantive factors caused this to happen. The formula in Table 1 clearly shows that what is happening is that the additional variance is accounted for by the grouping estimators. Thus, the gain in the R^2 is not due to better data but simply the contribution of the grouping scheme -- D -- and not to the variables of interest in aggregate analysis -- the X matrix.

The Generalized Least-Squares Approach to Aggregation

In this section, if we start with (1) of the previous section the grouping of observations into p groups and taking means yields (Johnston, pp. 228-241):

$$\bar{y} = \bar{X}\beta + \bar{u} \tag{12}$$

Then the ungrouped data are related to the aggregated in these forms,

$$\bar{y} = Gy \tag{13}$$

$$\bar{X} = GX \tag{14}$$

$$\bar{u} = Gu \tag{15}$$

with G as the grouping matrix of (m x n). The form of G is, for instance,

$$G = \begin{matrix} 1/1 & 1/1 & 0 & 0 & 0 & 0 \ldots 0 \\ 0 & 0 & 1/2 & 1/2 & 1/2 & 1/2 \ldots 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \ldots 1/p \end{matrix} \tag{16}$$

While $E(\bar{u}) = 0$ it is also noted that

$$E(\bar{u}\bar{u}') = \sigma^2 I \tag{17}$$

which means that the estimators will be unbiased but inefficient. However, it is the case that

$$E(\bar{u}\bar{u}') = \sigma^2 GG' \tag{18}$$

which is efficient. To estimated B, the generalized least squares is

$$b = [\bar{X}'(GG')^{-1}\bar{X}]^{-1}\bar{X}'(GG')^{-1}\bar{y} \tag{19}$$

and

$$var(b) = \sigma^2[\bar{X}'(GG')^{-1}\bar{X}]^{-1} \tag{20}$$

Here, generalized least squares overcomes the heteroscedastic problem (17) by inserting the grouping factor G in (18). The expression $(GG')^{-1}$ is actually a weighting matrix which contains the numbers in each group. Note that the generalized least squares estimates are not as efficient as the ungrouped ones.

The R^2 question may now be approached, when recalling from Table 1 that R^2 for equation (1) is

$$1 - R^2 = \frac{s's}{y'Ay} \tag{21}$$

by simple reexpression. But what about the R^2 for the groups values? The R^2 for equation (12) could be of the form

$$1-\bar{R}^2 = \frac{\bar{e}'(GG')^{-1}\bar{e}}{\bar{y}'(GG')^{-1}A\bar{y}} \tag{22}$$

where

$$\bar{e}'(GG')^{-1}\bar{e} = \bar{y}(GG')^{-1}\bar{y} - b'\bar{X}(GG')^{-1}\bar{y} \tag{23}$$

By definition the sums of squares may be partitioned:

$$y'Ay = \bar{y}'(GG')^{-1}\bar{A}\bar{y} + y^{*}'y^{*} \tag{24}$$

with

$$y^{*} = y - DGy \tag{25}$$

referring back to equation (3). Thus, it must be the case that

$$y'Ay \geq \bar{y}'(GG')^{-1}\bar{A}\bar{y} \tag{26}$$

and we can see that the reduction of the denominator for between groups sum of squares is again a function of D -- the relationship of the grouping factor with y. As the association of y with D increases, the between sum of squares decreases - i.e., $\bar{R}^2$ for between groups must increase.

An Example of the Effect of Aggregation on R^2

Of substantive interest in political sociology has been voter participation in the electoral process. In light of the traditional democratic norms concerning the importance of citizen participation, researchers have, through the years, focused on this problem. Though much of what is known about the factors influencing voter participation derives from survey, micro-level

data, there have been numerous occasions when aggregate data have been employed to investigate voting behavior (Alford and Lee, 1968). Studies using aggregate data have most often used correlational methods, seldom attempting to estimate regression coefficients. This example will illustrate grouping data by census blocks and tracts (a common procedure in macro-level voting studies) as it compares with ungrouped responses. Kim, Petrocik and Enokson (1975) treat the problems of analyzing voting with aggregate data where micro and macro data are combined and systematically measure the interaction.

The model of voter participation used in this example is drawn from the literature based on micro-level survey data. It is hoped that this will lessen the likelihood of misspecification and thus avoid that additional handicap. Ben-Sira (1977) has suggested a model based on a thorough review of the previous research on voting and notes that there has been shown to be a strong association between socio-economic status and voting. The trend is for individuals to have a higher propensity to vote, given a higher social status. The components of the model are presented in Table 2.

The data is from a one percent survey of Atlanta and suburban Fulton County conducted in 1976 and yielding over 7,000 respondents. The substantive model and this data provide a background to test the methodological problem of the effect of grouping on correlational measures such as the R^2. Incomplete data were accounted for by the mean substitution technique which does not bias the regression coefficients but does lower variance and efficiency. No missing data cases for the dependent variable were included.

The illustration is in the form of three regression analyses: one each of three levels of aggregation -- census blocks, tracts, and total respondents. The specification remained the same for each level at which the data were grouped. The model for the grouped data was that of equation (12) and was estimated as a special case of generalized least squares, weighted least squares. Due to the limitations of the Statistical Package for Social Sciences (SPSS) software, a dummy constant term had to be included in grouped equations along with the actual constant term, but this does not affect these examples. The ungrouped data was simply estimated by the model in equation (4). The R^2 in the micro-level specification was found to be .14, a modest coefficient but not unrespectable, given that voting was coded as a dicotomy with having voted in the previous five years as a "1" and not having voted as "0". The strongest variable was the years of schooling.

The grouping by census blocks resulted in 2867 blocks for 7018 individuals and, as shown in Table 2, the resulting R^2 was .18, indicating a modest increase from the total R^2 of .14. The partitioning of sum of squares is presented in Table 3 and indicates that the within groups R^2 is .12, suggesting that controlling for the effects of grouping by blocks has a slight but measurable impact on the specification. Additionally, this implies that grouping by census blocks could possibly be done for purposes of confidentiality or if there were need to reduce data components (see Feige & Watts, 1972).

As is shown in Table 2, the data grouped by census tracts were analyzed. The partitioning of the sum of squares is given in Table 3. The R^2 for total, and also for within census tracts, was .14, however, the between census tracts R^2 was found to be .60 -- a clear example of the effect of grouping on the R^2. Here the increase in R^2 was due to the association between census tracts (the D matrix), as a proxy measure of contextual factors, with the dependent variable (the y vector). In addition, it should be noted that while the regression estimators were not seriously affected by the grouping by tracts, the standard errors of estimators were inflated along with the standardized regression coefficients (β). Thus, the typical measures of the importance of the regression were altered to a considerable degree by aggregation effect. But to restate, for the researcher who is hoping to estimate the micro specification with aggregate data, the R^2 and standardized coefficients are to a large degree a product of the grouping effect itself rather than the substantive in dependent variables.

Summary and Discussion

The purpose of this paper has been to approach the problem of the effect of grouping on R^2 from the analysis of covariance approach, and relate it to the clustering approach of generalized least squares. While through the years there have been warnings against over-reliance on R^2 with grouped data, there continues to be statements such as:

> An additional motivation for using grouped data, however, is that even with sophisticated operational definitions of income and prices, these explanatory variables alone appear to "explain" only a small part of the variations in demand for specific goods and services in individual household data. Grouping observations by the independent variables considerably increases the "explanatory power" of the estimating equation. (Michael and Becker, 1973, pp. 379-380).

It is hoped that the above authors were not seriously claiming that grouping increased the substantive "explanatory power" of their specification. What, in all likelihood, occurred was an artifactual increase in R^2 that the grouping factor induced. It should be noted that there does exist the possibility of an actual "aggregation gain" when, for instance, the micro equation is misspecified and the grouping factor (the D matrix) helps correct the poorly specified micro model (see Irwin and Lichtman, 1976, pp. 423-433). A similar point has been made by Hanuschek, Jackson, and Kain (1974).

The generalized least squares perspective can also be an aid in investigating both temporal and spatial autocorrelation. The G matrix can be used to correct for such misspecifications as heteroscedasticity and autocorrelation. Granger and Newbold (1974) have cautioned that high R^2 may be generated by a misspecified temporal autocorrelation structure. Spatial autocorrelation can result in inefficiency of the estimates of cross sectional studies (Lebanon and Rosenthal, 1975; Cliff, Haggett, Ord, Bassett and Davies, 1975).

The researcher cannot expect the R^2 determined from grouped data to be a robust measure for use in evaluating models unless the grouping procedure is random with respect to the dependent variable. The R^2 "inflation problem" is actually a specification issue where methodology and technique are, at best, only partial factors in a more complete solution.

References

Alford, R.R. and E.C. Lee (1968) "Voting Turnout in American Cities." American Political Science Review, 62: 796-813.

Ben-Sira, Z.C. (1977) "A facet theoretical approach to voting behavior." Quality and Quantity, 11: 167-188.

Cliff, A.D., P. Haggett, J.K. Ord, K.A. Bassett, and R.B. Davies (1975) Elements of Spatial Structure: A Quantitative Approach. New York: Cambridge University Press.

Cramer, J.S. (1964) "Efficient grouping: regression and correlation in Engle Curve Analysis." Journal of the American Statistical Association, 59: 233-250.

Feige, E.L. and H.W. Watts (1972) "An investigation of the consequences of partial aggregation of micro-economic data." Econometrica 40: 343-360.

Gehkle, C. and K. Biehel (1934) "Certain effects of grouping the size of the correlation coefficient in census tract material." Journal of the American Statistical Association Supplement 29: 169-170.

Granger, C.W.J. and P. Newbold (1974) "Spurious regression in econometrics." Journal of Econometrics 2: 111-120.

Grunfeld, Y. and Z. Griliches (1960) "Is aggregation necessarily bad?" Review of Economics and Statistics 42: 1-13.

Hanushek, E.A., J.E. Jackson and J.F. Kain (1974) "Model specification, use of aggregate data, and the ecological correlation fallacy." Political Methodology 1: 87-106.

Irwin, L. and A.J. Lichtman (1976) "Across the great divide: inferring individual level behavior from aggregate data." Political Methodology 3: 411-439.

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

Kim, J.O., J.R. Petrocik, and S.N. Enokson (1975) "Voter turnout among the American States: systemic and individual components." American Political Science Review 69: 107-123.

Lebanon, A. and H. Rosenthal (1975) "Least squares estimation for models of cross-sectional correlation." Political Methodology 2: 221-244.

Michael, R.T. and G.S. Becker (1973) "On the new theory of consumer behavior." Swedish Journal of Economics 75: 378-396.

Prais, S.J. and J. Aitchison (1954) "The grouping of observations in regression analysis." Review of the International Statistical Institute 22: 1-22.

Robinson, W.S. (1950) "Ecological correlations and the behavior of individuals." American Sociological Review 15: 351-357.

Smith, K.W. (1977) "Another look at the clustering perspective on aggregation problems." Sociological Methods and Research 5: 289-315.

Theil, H. (1971) Principles of Econometrics. New York: Wiley.

Thorndike, E.L. (1939) "On the fallacy of inputing the correlations for groups to the individuals or smaller groups." American Journal of Psychology 52: 122-124.

Yule, G.U. and M.G. Kendall (1950) An Introduction to the Theory of Statistics. London: Charles Griffin.

Table 2

COMPARISON OF REGRESSIONS FOR MICRO-LEVEL DATA WITH CENSUS BLOCK AND TRACT AGGREGATIONS

Dependent Variable: Voted in Previous Five Years

	Micro-Level			Macro-Level					
				Census Blocks			Census Tracts		
Independent Variables	Estimator	Standard Error of Estimator		Estimator	Standard Error of Estimator		Estimator	Standard Error of Estimator	
Schooling (years)	.040	.0016	.32	.040	.0024	.35	.065	.010	.74
Age (10 year units)	.038	.0032	.15	.033	.0046	.13	.045	.022	.16
Income ($10,000 units)	.022	.0050	.052	.031	.0079	.074	.054	.034	.16
Race (White)	-.029	.011	-.033	-.034	.015	-.043	-.081	.033	-.19
Political Efficacy	-.0019	.0036	-.0058	.0029	.0053	.0293	-.023	.026	-.051
Public Interest	.0068	.00062	.13	.0062	.00096	.12	.0074	.0045	.13
Governmental Salience (high or low)	.019	.010	.021	.022	.016	.024	-.073	.066	-.077
Constant	.017			.011	.019	.0095	.22	.14	.11
Dummy Constant	*			-.012			-.15		
R^2	.14			.18			.60		
Standard Error of Estimate	.41			.34			.11		
N	7018			2867			137		

*Not needed in the micro-level specification

Table 1

ANALYSIS OF COVARIANCE APPROACH TO GROUPED DATA

Source of Variation	Residual Sum of Squares	$1-R^2$
Between	$e'e = y'y - \hat{\hat{\alpha}}'D'y - \hat{\hat{\beta}}'X'y$	$\frac{y'y - \hat{\hat{\alpha}}'D'y - \hat{\hat{\beta}}'X'y}{y'Ay}$
Within	$s's - e'e = \hat{\hat{\alpha}}'D'y + \hat{\hat{\beta}}'X'y - \hat{\beta}'X'y$	$\frac{\hat{\hat{\alpha}}'D'y + \hat{\hat{\beta}}'X'y - \hat{\beta}'X'y}{y'Ay}$
Total	$s's = y'y - \hat{\beta}'X'y$	$\frac{y'y - \hat{\beta}'X'y}{y'Ay}$

Table 3

EFFECT OF GROUPING BY CENSUS TRACTS AND BLOCKS ON SUMS OF SQUARES FOR VOTING MODEL

Source of Variation	Sums of Squares: Regression	Residual	Total	R^2
Between Census Tracts	2.39	1.59	3.97	.60
Within Census Tracts	189.56	1183.77	1373.34	.14
Total	191.95	1185.36	1377.31	.14
Between Census Blocks	71.82	333.72	405.54	.18
Within Census Blocks	119.82	851.64	971.77	.12
Total	191.94	1185.36	1377.31	.14