Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy

Eric A. Hanushek, John E. Jackson, and John F. Kain

INTRODUCTION

This paper is motivated by our observation that many researchers have confused the interrelated, but analytically separable, problems of data aggregation, of model specification, and of statistical bias in parameter estimation. This confusion has had such a profound and continuing impact upon social science research and analysis that a further attempt at clarification does not appear extravagant.

The origin of much of this confusion is a widely cited article by W. S. Robinson (1950) in which he states:

> The purpose of this paper will have been accomplished, however, if it prevents the future computation of meaningless correlations and stimulates the study of similar problems with the use of meaningful correlations between the properties of individuals. (Robinson, 1950:357)

Robinson's meaningless correlations are simple (bivariate) correlations based on aggregate (ecological) data, while his meaningful ones are similar simple (bivariate) correlations based on micro- or individual data. Robinson's article has been widely interpreted to demonstrate conclusively that:

(1) aggregate (ecological) data are unsuitable for analysis of individual and household behavior, and (2) statistics estimated with individual or microdata are unambiguously better. Both assertions are incorrect.

Robinson argues that aggregate data cannot be used to study the properties of individual members of these aggregates. But even Robinson's statement of the problem is incorrect, or misleading at best. Social scientists, except possibly psychologists, are rarely interested in individuals or the properties of individuals, per se. Social scientists' interests are the regularities in human behavior associated with the effects of various characteristics. Thus, they are interested in such questions as: the effect of income on consumption; the effect of family background on educational achievement; the effect of party affiliation on voting behavior; and the effect of income on residential location decisions. In fact, the empirical analysis of sample data, whether it is aggregate or individual, cannot be used to study "individual" behavior.

The objective of most empirical analyses is to determine the independent effects, in a probabilistic way, of some household or individual characteristic on the behavior of households or individuals' possessing that characteristic. Because of Robinson's misstatement of the problem, he and his interpreters further mistake the problem of the proper formulation and specification of statistical models for the appropriateness of particular statistical procedures and types of data. Robinson's article does identify a problem in statistical estimation, but its limited correct observation is virtually irrelevant to most of the questions and analysis for which it is cited as authority. Much of the persuasiveness of Robinson's article clearly arises from some empirical examples. In formulating these examples, Robinson commits a serious model specification error which dominates and biases his results at the aggregate or ecological level. As we demonstrate below, had he considered a more complete and accurately specified model, his empirical findings would have been much different and his conclusions, relating to the appropriate use of aggregate data, would have been much more limited and much less severe.

THE ROBINSON MODEL

Robinson in his article considers the effects of race and national origin on illiteracy. Formally he tests the hypotheses that: (a) Negroes were more illiterate than whites, and that (b) foreign-born whites were more illiterate than native-born individuals in 1930. He considers two types of evidence: first, he compares the estimated

proportion of Negroes who are illiterate with the estimated proportion of all whites who are illiterate, and similarly the foreign- and native-born individuals classified as illiterate in 1930. These proportions are: 16.1 percent of Negroes as contrasted to 2.7 percent of whites and 9.9 percent of foreign-born whites and 3.1 percent of all native-born persons (10 percent of whom are Negro). Second, he computes the Pearsonian (fourfold-point) correlation coefficient using individual data. The simple correlation between being illiterate and being Negro is 0.203, and between being foreign-born and being illiterate is 0.118.

Robinson then computes simple correlations between the percent illiterate in each state and the percent Negro and the percent foreign-born. These are 0.773 and -0.526 respectively.[1] From this evidence Robinson concludes: (1) that correlations computed with aggregate data (states) bear no consistent relationship to the correlations based upon the individuals, and that (2) they may lead to completely erroneous conclusions, as in the case of the simple correlation between percent illiteracy and percent foreign-born where the ecological correlation has a reversed sign from the individual correlation.

As descriptors of social science behavior, regression coefficients may be preferred to simple correlation coefficients. But, as other authors have pointed out (Goodman, 1953), the same problem exists if bivariate regression analysis is used rather than simple correlations, because a bivariate regression is a linear transformation of the simple correlation coefficient. Since the regression coefficients are often easier to interpret than the correlation coefficients, they will, without changing any conclusions, be used throughout this analysis. The simple regression models relating illiteracy to being Negro or being foreign for individuals are shown in equations 1 and 2, where I_i is a dichotomous variable which is one if the person is illiterate and zero if he is not. Similarly, N_i is one if the person is Negro, zero if he is not, and F_i is one for foreign-born individuals and zero for native-born. The e's represent the observed deviations from the estimated model. These same regressions computed for percentages at the state level are,

$$I_i = 0.03 + 0.14\ N_i + e_1 \tag{1}$$

$$I_i = 0.04 + 0.07\ F_i + e_2 \tag{2}$$

$$I = 0.02 + 0.22\ N + e_3 \tag{3}$$

$$I = 0.07 - 0.29\ F + e_4, \tag{4}$$

where I, N, and F are the proportions of the states' populations illiterate, Negro, and foreign-born respectively. The aggregate data implies a higher illiteracy rate among Negroes than is implied by the individual model. In equation 3, a unit increase in the percent of a state's population which is black yields a predicted increase in the state's population which is illiterate of 0.22 percent and implies that 24 percent of all blacks were illiterate. The model for foreign-born implies that a unit increase in the percent foreign-born white decreases the state's percent illiterate by 0.29 percent and that the foreign-born illiteracy rate was -.022 percent! These results lead to the same conclusions that Robinson made for correlation analysis, that aggregate regressions may bear no similarity to the individual regressions and that they may even have the incorrect sign, as in the case of foreign-born.

Robinson's error, and the similar difficulties with the regression model, however, result principally from an incorrect and incomplete model and from improper statistical methods rather than from the use of wrong data. The propositions can be illustrated by restating Robinson's analysis in terms of the estimation of the parameters in a more complete model of the determinants of illiteracy (literacy) at both the individual and state level. The problem is to identify and measure the influence of those factors which affect an individual's educational achievement or the probability that he will exceed some literacy norm.

The list of variables that might have some effect on the educational achievement of an individual is virtually limitless. It includes the circumstances of his schooling, the characteristics and attitudes of his parents, of their parents, and of their parents, his native intelligence, the attitudes of his classmates, and presumably such individual matters as his home discipline and sibling relationships. Even this partial enumeration of the variables influencing literacy illustrates the inadequacy of Robinson's models, in which illiteracy depends only on being black or foreign-born. This error is compounded, however, because several explanatory variables, such as the amount and quality of schooling, are systematically correlated with the characteristics of the individuals Robinson included in his model, particularly at the state level.

The importance of proper specification of statistical models can be illustrated by estimating an expanded version of Robinson's model of state illiteracy rates. Equation 5 presents a model for state illiteracy rates, based on the previous discussion of excluded variables, which can be estimated using the aggregate data to which Robinson objects.

$$(5) \qquad I = B_1 + B_2 N + B_3 F + B_4 M + B_5 \text{Ind} + B_6 S + e$$

where,

I = percent of the state's population which is illiterate;

N = percent of the state's population which is Negro;

F = percent of the state's population which is foreign-born (including non-white);

M = percent of the state's population which is Mexican;

Ind = percent of the state's population which is Indian;

S = percent of the state's elementary school age population (7-13) enrolled in school;

e = a stochastic error term representing the effects of random excluded variables, measurement errors, etc.

Like the model implicit in Robinson's ecological correlations, equation 5 uses the percent of the state's population that was illiterate in 1930 as the dependent variable. It departs from Robinson's analysis, however, in that it simultaneously examines the influence of several explanatory (independent) variables. Equation 5 expresses the per capita illiteracy rate among the forty-eight states and the District of Columbia in 1930 as a linear function of the percent of the population that is Negro (N) and the percent foreign-born (F), which are Robinson's hypotheses. In addition, the percent of the population which are Mexican (M), the percent Indian (Ind), and the percent of the school age population enrolled in school (S), are included in the model. The Mexican and Indian variables are included because they constitute minorities which are discriminated against in much the same way as Negroes. The variable measuring school attendance rates is included to measure the average educational experiences of the individual residents of each state.[2]

The estimates obtained from the 1930 state data, shown in equation 6, indicate that if school attendance rates were to fall to zero, the illiteracy rates for whites, Negroes, foreign-born whites, Mexicans, and Indians would be 86 percent, 96 percent, 98 percent, 89 percent, and 100 percent respectively.

$$(6) \quad I = \underset{(6.23)}{0.86} + \underset{(2.90)}{0.10}\,N + \underset{(2.82)}{0.12}\,F + \underset{(0.30)}{0.03}\,M + \underset{(0.80)}{0.14}\,Ind - \underset{(-6.13)}{0.88}\,S$$

(t-statistics) $\qquad R^2 = 0.86$

We can compute the illiteracy rates for native whites and the four minority groups using equation 6, the 1930 school attendance rates for the school-age members of each group, and by assuming a population which is 100 percent native, Negro, foreign-born, etc. These computed illiteracy rates and the actual rates are shown in table 1 along with the misspecified estimated Negro and foreign-born rates from equations 3 and 4. The significant aspect of these results is that including a measure of the availability of educational services yields estimates of minority group illiteracy rates which approach the actual rates. In fact, the differences between our estimated and actual rates are less than the standard error of each of the minority group coefficients. Inferences about the illiteracy rates of these groups relative to native whites is much different in the better specified model than in the misspecified model.

TABLE 1:

ESTIMATED AND ACTUAL ILLITERACY RATES

Group	Estimates (Equation 6)	Actual	Misspecified Estimates (Equations 3 & 4)
Negro	19.1%	16.3%	24.4%
Foreign-Born	12.6	10.3[a]	-21.9
Mexican			
Native Born	19.8	21.8	
Foreign-Born	28.9	31.5	
Indian	28.8	25.7	
Native White	1.3	1.5	

[a]Includes all non-white foreign-born except Mexican.

It is not hard to explain Robinson's apparently contradictory results. The quantity and quality of schooling available to the residents of different states varies greatly. For example, in 1930 enrollments among the elementary school-aged population varied from below 90 percent in many southern states to over 98 percent in New England and the Upper Midwest, while in 1910, the respective rates were 70 percent and 95 percent with one state (Louisiana) below 60 percent. Part of the observed differences in illiteracy among the states are explained by these differences in school attendance. In 1930, the U. S. Negro population was heavily concentrated in states with historically low education levels for both whites and blacks. Conversely the foreign-born population was concentrated in the Northeast and North Central States, where school attendance was relatively high among all population groups. Robinson's Negro and foreign-born variables then measure the combined effects of minority-group illiteracy rates and the quantity of education services on the state's illiteracy rate. The net effect is to overwhelm the foreign-born influence, reverse the sign, and provide a biased estimate of the independent influence on illiteracy of being foreign-born.

BIAS IN STATISTICAL MODELS[3]

Model misspecification affects the relationship of the sample estimate of a given coefficient to the true or population value of the coefficient. In general, when relevant variables are excluded from the estimation of a statistical model, the expected value of the sample estimate will not equal the true coefficient. In statistical terminology, the coefficient is biased. The magnitude of bias, the difference between the expected value of the estimate and the true coefficient, is a multiplicative function of: (1) the strength of the omitted variables, and (2) the correlation within the sample between the omitted and the included variables. Assume that the true relationship is:

$$Y_i = B_0 + B_1 X_{i,1} + B_2 X_{i,2} + u_i \tag{7}$$

and X_2 is omitted during the model estimation. The expected value of b_1 (the estimate for B_1) will differ from B_1 by an amount equal to $B_2 * b_{21}$ where b_{21} is the coefficient from the regression of X_2 on X_1 in the sample and B_2 is the population parameter relating X_2 to Y. Within the framework of

Robinson's model, imagine that Y is illiteracy, X_1 is race (or nativity), and X_2 is schooling. A characteristic of the 1930 aggregate sample is the high correlation between X_1 and X_2 because of geographic and locational considerations; neither population group nor educational services were uniformly distributed across the country. This implies a high b_{21} and thus, a large bias.

Two points need to be emphasized. Bias is introduced by a misspecified model regardless of whether the observations are based on individual or aggregate data. In addition, the term expressing the bias in a misspecified model contains one element which is a population parameter, B_2, and one element which is characteristic of the particular sample being used in the statistical study, b_{21}. This suggests that it is possible to reduce, or even eliminate, the specification bias by careful sample selection. If a sample of data can be obtained in which X_2 does not vary, i.e., is held constant, then b_{21} must be zero and the bivariate estimate of B_1 will be unbiased. Physical scientists usually are able to accomplish this objective by proper sample stratification. In other circumstances they are able to obtain a data sample in which X_1 and X_2 are uncorrelated usually through appropriate experimental design. This lack of correlation between X_1 and X_2 is usually much harder with aggregate data, since individuals and households generally do not group themselves randomly across geographic or social areas. Thus, units with high or low values of X_1 are also systematically likely to have high or low values for X_2 due to this nonrandomness.

Although bias will be present in a misspecified model whether individual or aggregate data are used for estimation, it is true that the amount of the bias will usually be less for regressions estimated for microdata. For example, the bias is smaller in equation 1, the regression on microdata, than in equation 3, the regression on aggregate data, because the correlation between being Negro or foreign-born and the amount and level of education is less for individuals than the correlation between percent Negro and amount of education at the state level. (However, aggregation could conceivably reduce bias rather than increase it if the grouping is done appropriately. For example, if our aggregate units were cities in the 1930 census, and if some chosen cities had high educational expenditures with a large number of blacks and some cities with low expenditures had

a small percentage of black citizens, then the correlation between percent Negro and educational services could be small and our estimate of B_1 would be less biased.)

Specific cases which imply that B_2 or b_{21} is small or zero have been used by some authors (Goodman, 1959; Blalock, 1965; Shively, 1969) to define situations in which ecological regression is appropriate. Goodman (1959:612-13), in commenting on the relationship between illiteracy rates, percent Negro, and aggregate data, says that the aggregate estimates will be unbiased,

> when the probability of illiteracy, say, is more a function of color . . . rather than a function of the ecological area being considered. Where the phenomenon under investigation is more a function of the area . . . than a function of color, the methods presented here are not recommended . . . .

Goodman concludes that aggregate regressions are inappropriate when an excluded variable, such as education, has a significant effect on illiteracy (i.e., B_2 is large) and this omitted variable is not uniformly distributed across the aggregate units. In this case illiteracy has become largely a function of area due to the differences in educational services at the state level. We could still avoid the bias problem, however, if the distribution of blacks and foreign-born individuals was uncorrelated with the distribution of educational services, that is, if b_{21} were zero.[4] This is obviously not true of the data analyzed by Robinson.

There is another route open to the researcher. Social scientists typically must rely on data obtained for other purposes and seldom have the opportunity to reduce b_{21} through original sample selection or stratification. Use of multivariate models is, therefore, usually a more convenient and practical method of obtaining unbiased estimates than acquiring samples in which X_1 and X_2 are uncorrelated. Multiple regression analysis incorporating better model specification can be used to eliminate or to minimize bias in parameter estimation.

As demonstrated in the appendix, the problem of bias results from the correlation within the sample between the included independent variable (X_1), and the error term or residual in the model. This correlation in the misspecified case arises because the error term includes X_2. If all important explanatory variables are included in a multivariate model, then the error term will not be correlated

with any variable or variables the researcher is investigating, and the estimates of the relevant coefficients will be unbiased.[5] We have shown already that when logical explanatory variables are added to Robinson's bivariate models, much more realistic estimates of the coefficients are obtained. These estimates, moreover, tend to agree more with the results from the individual data.

SPECIFICATION AND AGGREGATION

To illustrate these problems of aggregation and model specification more clearly, it is useful to compare the four permutations of well-specified/poorly specified and aggregated/disaggregated models for a consistent sample of data. For this purpose we now consider a model of educational achievement analyzed by one of the authors (Hanushek, 1972). This model, which hypothesizes that an individual's educational achievement at any point in time depends on his family background, on school inputs, and on his entering achievement level, was tested with a 1969 sample of 1061 third-grade students attending twenty-five separate schools. Fourteen percent of the students were Mexican-Americans; the remainder were Anglos. For each student, data are available for third-grade reading achievement, father's occupation, teacher's verbal ability, recentness of teacher's education and first grade achievement.[6] The results of the four model estimates of the independent achievement effects associated with being Mexican-American are shown in table 2. Each cell includes the estimated coefficient for the Mexican-American variable along with the R^2 for the model.

TABLE 2:

AGGREGATION AND MISSPECIFICATION EFFECTS ON ESTIMATED MEXICAN-AMERICAN COEFFICIENT

Level of Aggregation		Model Specification	
		Bivariate	Multivariate
Individual	b_1	-11.60	-3.20
	R^2	0.04	0.54
School	b_1	-28.70	-8.40
	R^2	0.47	0.82

The first attribute of the different estimates is that the R^2 (and, thus, r) is always higher in the aggregate case than in the individual case. However, we wish to emphasize the importance of obtaining accurate estimates of the coefficients, rather than explaining the variance of the dependent variable. The coefficient tells us how much the dependent variable can be expected to change with changes in the explanatory variable. The coefficient of determination (R^2) merely tells us how well sample values of Y are "explained" by the statistical model. The R^2 usually increases in the aggregate case, but this apparent increase in explanatory power is largely irrelevant.[7] The behavioral content of the model is contained in the coefficients, and these are clearly the most important output of a statistical analysis.

When considering the coefficients, several things are striking about table 2. First the misspecified model is much more sensitive to aggregation than the well-specified model; the change in the Mexican-American coefficient was -17.1 in the misspecified model compared to -5.2 in the multivariate equation. Second, the individual multivariate coefficient is more similar to the coefficient in the aggregate multivariate equation than to the one in the individual bivariate model. This is a graphic demonstration of the central theme of this paper. The mere existence of micro-data does not insure that coefficient estimates will be more accurate, more interesting, or more useful. Even micro-data are subject to interpretive difficulties and statistical bias when used in a misspecified model.

A word of caution is, however, necessary. Admonitions about obtaining the correct model specification are not universally helpful. Similarly, mere reference to the fact that correct equation specification can alleviate many of the problems associated with aggregate data will not insure unbiased estimates. It is often difficult in the social sciences to know the correct specification of behavioral relationships or to obtain the data required for an appropriate multivariate model. Incomplete theoretical structures and lack of appropriate data even at the aggregate level require that analysts have considerable skill and imagination to obtain even approximately correct specifications.

Aggregate data tends to compound specification problems. First many social science theories relate to behavior at the microlevel, and these theories do not always lead to simple aggregations. For example, models of mass voting behavior often require measures of individual attitudes on different issues or toward the competing candidates. It is difficult to see how a model incorporating such influences could be specified and an appropriate variable measured at the aggregate level. Secondly, the process of aggregation may increase the sample correlations

between appropriate variables, because individuals are not grouped randomly. We have already mentioned this problem and the difficulty produced when multicollinear variables are included in the estimated equation. The increased correlations also mean that the bias introduced by any model misspecification will be more severe with aggregate data. This was illustrated with the educational achievement example where the effects of misspecification were more severe in the aggregate than in the individual model. The important point is that problems of specification are inherent in all statistical analyses regardless of the type of data being used, and they deserve the most serious consideration on the part of the researcher and reader.

CONCLUSIONS

There are many advantages to having microdata. One aspect of most natural aggregation schemes is that the independent effects of different variables are obscured and that it becomes difficult to disentangle these separate effects. Microdata often provide considerably more information than do aggregate data in the sense that there is more independent variation among different explanatory variables. Also, microdata allow the analyst more flexibility by allowing him to decide on any aggregations or stratifications, rather than having them imposed by the form of the data presentations.[8] Thus, there is a clear preference for microdata, and we would never argue that there are not advantages to having microdata.

Microdata are not, however, a panacea. Microdata estimates are subject to the same strictures about proper model specifications as aggregate data, although the penalties for violation of these strictures may be less severe with microdata. It is simply not true, however, that any simple correlation using microdata is superior to the coefficient estimates from a similar, but well-specified, multivariate aggregate model. Multivariate models usually are more interesting in terms of behavior content, and they often have better (less biased) coefficient estimates.

Finally, microdata are not always available for the problems which concern social scientists. This is an important consideration, and researchers should not let their inquiries be constrained by a reluctance to use aggregate data. Aggregate data relevant to many current questions are readily available and amenable to useful analysis. This is not the case with microdata; nor is it likely to be the case in the near future. For example, it is now impossible to collect additional individual data on the

1972 presidential election, to say nothing of the elections of 1928 and 1896. If we rely solely on microdata to analyze social science problems, many important questions will be made inaccessible to research. Aggregate data are also usually less expensive to collect, store, and analyze than individual data. Consequently, currently limited budgets may, in some instances, be more profitably spent examining additional questions through aggregate data. What this paper has tried to do is facilitate such efforts by identifying the difficulties associated with aggregate data and suggesting ways to deal with these through model specification.

APPENDIX

The biases introduced by trying to estimate a misspecified model are very easy to show in a formal way.[9] Assume that the true model explaining some behavior measured by the variable Y is,

$$Y_t = B_1 X_{t,1} + B_2 X_{t,2} + u_t, \tag{A-1}$$

where all variables are measured as deviations about their means so that there is no need to include a constant term. The u_t of course measures the effects of any random elements included in our measurement of Y and is assumed to be independent of both the X variables. If the misspecified model in A-2 is estimated with a sample of data,

$$Y_t = B_1 X_{t,1} + e_t \text{ where } e_t = B_2 X_{t,2} + u_t \tag{A-2}$$

certain assumptions about the relationship between X_1 and X_2 in the data sample must be made if we are to assume that the results based on equation (A-2) are to be unbiased. If ordinary least squares techniques are used to estimate the parameter in the misspecified model, b_1, the least squares estimator of B_1 is given by:

$$b_1 = \frac{\sum_{t=1}^{T} (X_{t1} Y_t)}{\sum_{t=1}^{T} X_{t1}^2}. \tag{A-3}$$

$E(b_1)$, the expected value of b_1, is found by substituting the true model, equation (A-1), into equation (A-3) and taking the expected value of both sides. Under the assumptions that X_1 and u and X_2 are independent and that $E(u_t) = 0$,

(A-4)

$$E(b_1) = \frac{E\sum\left[X_{1t}(B_1X_{t1} + B_2X_{t2} + u_t)\right]}{\sum X_{t1}^2} =$$

$$\frac{E\left[B_1\sum X_{t1}^2 + B_2\sum X_{t1}X_{t2} + \sum X_{t1}u_t\right]}{\left[\sum X_{t1}^2\right]} =$$

$$B_1 + B_2 * E\left[\frac{\sum X_{t1}X_{t2}}{\sum X_{t1}^2}\right] \quad \text{since } E\sum X_{t1}u_t = 0.$$

The last part of equation (A-4) indicates that the expected value of b_1 equals the true value of B_1 plus a term involving the true coefficient B_2 and the relationship between X_1 and X_2. In fact this latter term, $\frac{\sum X_{t1}X_{t2}}{\sum X_{t1}^2}$, is simply the regression coefficient that would be obtained by regressing the sample values of X_2, the excluded variable, on the sample values of X_1, the included variable and is the covariance of X_1 and X_2 divided by the variance of X_1. We will call the regression coefficient obtained in this fashion b_{21}. Clearly, X_2 is often unmeasured at this point, or presumably it would have been included in the analysis. However, there is a value for X_2 implicitly associated with each one of our sample observations so that conceptually this regression could be computed. In terms of Robinson's model, this X_2 would measure the amount of education received by each person in the 1930 census. Theoretically, then, the Census Bureau could have presented a table

showing the education level or amount of education received by whites and Negroes. If this had been done, the value of b_{21} could have been computed for both the individual and the aggregate levels.

When the term containing the sum of the products of X_1 and X_2 is rewritten as b_{21} equation (A-4) can be restated as,

$$E(b_1) = B_1 + B_2 b_{21}. \tag{A-5}$$

The term $B_2 b_{21}$ shows the bias in the expected value of b_1. The bias will disappear only if B_2 is zero (i.e., if X_2 has no effect on Y) or if b_{21} is zero (i.e., if there is no systematic covariation between X_1 and X_2 in our sample). The greater the effect of X_2 on Y and the larger the relationship between X_2 and X_1 as measured by b_{21}, the larger the bias will be. This bias will be positive, meaning that b_1 will overestimate B_1, if both B_2 and b_{21} are positive or if both are negative. If one is negative and the other is positive, the estimate of B_1 will be underestimated.

In the case of Robinson's models at the state level, the major excluded variable is the amount of education provided to the citizens of each state. This variable should have a negative effect on illiteracy rates; that is, B_2 is negative. This variable is also very negatively related to the percent Negro in each state; that is, b_{21} is also negative. The effect, then, of excluding the provision of educational services from the Negro model is to overestimate the coefficient on percent Negro in the misspecified model, which is exactly what we found when we compared equations 3 and 6. In the case of the foreign-born model, B_2 is still negative, but now b_{21} is positive since foreign-born individuals tended to live in areas of greater educational services. This causes b_1 to be badly underestimated. The magnitude of the bias in this case being sufficient to make the estimated coefficient negative.

FOOTNOTES

[1]These correlations are 0.946 and -0.619 respectively when computed on the basis of nine census regions rather than the 48 states.

[2]The data for the illiteracy, Negro, and foreign-born variables were taken from the Bureau of the Census (1933). Data for the Mexican and Indian variable were from the U. S. Dept. of Commerce (1932:16-17). Enrollment data came from the Bureau of the Census (1933:1104).

Although the schooling variables pertain to 1930, they are highly correlated with state values for previous years. The simple correlation between 1930 and 1910 state values for enrollment are on the order of 0.80. The illiteracy rate in any given state could be a function of the type of in and out migration that state has experienced. For example, if a state with a low attendance rate has had an influx of people from states with high attendance rates, the illiteracy rate in this state will be lower than predicted by the model in equation 5. This is indicated by a negative value for e. Likewise a state with a high in-migration of people from states with poorer educational services will have a higher than expected illiteracy rate. If the total number of migrants constitutes a significant portion of a state's population and these characteristics of the migrants' educational backgrounds are correlated with the variables included in the model, it will bias the estimated coefficients just as Robinson's excluded variables biased his results. We are assuming here that the proportion of a state's population in 1930 which was educated in a state providing a significantly different amount of education is small and that the characteristics of these migrants are not correlated with percent Negro, percent foreign-born, etc.

[3]This section presents a nontechnical explanation of the preceding empirical results. The problems are presented in a more formal manner in a mathematical appendix, and they are developed in detail elsewhere. See, for example, Draper and Smith (1966:81-85) and Theil (1971:540-56).

[4]Blalock (1965) discusses a different method of reducing bias. He argues that in any analysis where the aggregation of units is either random or by groupings of the independent variable, the bivariate regression Y on X_1 will yield unbiased estimates of B_1. However, if the grouping is by the values of Y, or at least approximates this grouping,

then the bivariate estimation will produce biased estimates of B_1. Blalock is simply pointing out that in the former cases, the aggregation is accomplished in such a way that X_1 and X_2 will be uncorrelated in the grouped observations unless they are also systematically correlated at the individual level. If we assume that X_2 has a positive effect on Y, then a grouping by similar values of Y combines into a single observation the units with high values for both x's or low values for both x's. This then insures a positive correlation between X_1, the included variable, and X_2 the omitted variable, with the obvious result.

[5]Efforts to eliminate bias in this way will often be stymied by multicollinearity, a rather common statistical problem. Multicollinearity arises when two or more independent (explanatory) variables in an equation are too highly correlated to allow precise estimates of the individual regression coefficients. The coefficients subject to serious multicollinearity are unbiased, but the estimation error is large. Serious problems of multicollinearity are more often encountered using aggregate than microdata. For a discussion of problems of multicollinearity and its treatment, see Farrar and Glauber (1967).

[6]The conceptual basis for this model and variable definitions are found in Hanushek (1972). For the individual models, regression equations were estimated using 1061 observations; the well-specified model was the regression of individual achievement on each of the variables mentioned and a dummy variable for Mexican-American students, and the misspecified model was the regression of individual achievement only on the Mexican-American dummy variable. The aggregate models were the regressions of mean achievement in each of the twenty-five schools on the aggregates of the independent variables used in the individual models, and the percent Mexican-American in each school.

[7]Aggregating the predicted values from micro-equations will often provide better aggregate estimates (larger R^2 for the aggregate observations) than are obtained from models estimated from aggregate data. In other instances, particularly when the specification of the micro-models is in doubt the aggregate models will out perform micromodels in predicting aggregates. See Grunfeld and Griliches (1960) and Orcutt et al. (1968).

[8]Appropriate model specification and stratification are closely linked. There are many instances where data for

all of the important independent variables are not available. This arises both from not collecting all of the correct data and from not knowing all of the intricacies of a behavioral problem. In such a case, it is often possible to salvage a model through appropriate sample stratification. The model of educational achievement discussed in the previous section is a case in point. When a model of the educational process is estimated for the entire student sample, one implicit assumption is that all of the behavioral coefficients are the same for each individual. However, if some of the important factors are unspecified, the coefficients for different individuals could take on different values. In this example, some of the important differences between Mexican-Americans and Anglos are not known and stratification of the sample by ethnic background is called for. See Hanushek (1972).

There are also problems with the aggregation of data, although these are usually discussed in a different context. See Theil (1971) and Orcutt et al. (1968).

[9]Multivariate extensions of this development can be found in Draper and Smith (1966) and Theil (1971).

REFERENCES

Blalock, H. M.
1965 Causal Inferences in Non-Experimental Research. Chapel Hill: University of North Carolina Press.

Draper, N. R., and H. Smith.
1966 Applied Regression Analysis. New York: Wiley.

Farrar, D. E., and R. R. Glauber.
1967 "Multicollinearity in regression analysis: the problem revisited." Review of Economics and Statistics 49 (February):92-107.

Goodman, Leo.
1953 "Ecological regression and the behavior of individuals." American Sociological Review 18 (December):663-64.

1959 "Some alternatives to ecological correlation." American Journal of Sociology 64 (May):610-25.

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

Hanushek, E. A.
1972 Education and Race. Lexington, Mass.: D. C. Heath.

Orcutt, G. W., H. N. Watts, and J. B. Edwards.
1968 "Data aggregation and information loss." American Economic Review 58 (September): 773-87.

Robinson, W. S.
1950 "Ecological correlations and the behavior of individuals." American Sociological Review 15 (June):351-57.

Shively, W. P.
1969 "Ecological inference: the use of aggregata data to study individuals." American Political Science Review 63 (December):1183-96.

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

U. S. Bureau of the Census.
1933 Fifteenth Census of the United States: 1930, Volume 2. Washington, D. C.: U. S. Government Printing Office.

U. S. Department of Commerce.
1932 Statistical Abstract of the United States 1932. Washington, D. C.: U. S. Government Printing Office.