TECHNICAL WORKING PAPER SERIES

ON THE VALIDITY OF USING CENSUS GEOCODE CHARACTERISTICS TO PROXY INDIVIDUAL SOCIOECONOMIC CHARACTERISTICS

Arline T. Geronimus
John Bound
Lisa J. Neidert

Technical Working Paper 189

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 1995

This work was supported by the National Institute of Child Health and Development, Grant HD24122. The authors would like to thank Al Hermalin, Gary Solon, Douglas Wolfe and three anonymous reviewers for comments on an earlier version of this manuscript, Nancy Krieger for helpful conversations, and Sandra Crump and Carol Crawford for help in the preparation of the manuscript. This paper is part of NBER's research program in Labor Studies. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Technical Working Paper 189
December 1995

ON THE VALIDITY OF USING CENSUS GEOCODE CHARACTERISTICS TO PROXY INDIVIDUAL SOCIOECONOMIC CHARACTERISTICS

ABSTRACT

Investigators of social differentials in health outcomes commonly augment incomplete micro data by appending socioeconomic characteristics of residential areas (such as median income in a zip code) to proxy for individual characteristics. However, little empirical attention has been paid to how well this aggregate information serves as a proxy for the individual characteristics of interest. We build on recent work addressing the biases inherent in proxies and consider two health-related examples within a statistical framework that illuminate the nature and sources of biases. Data from the Panel Study of Income Dynamics and the National Maternal and Infant Health Survey are linked to census data. We assess the validity of using the aggregate census information as a proxy for individual information when estimating main effects, and when controlling for potential confounding between socioeconomic and sociodemographic factors in measures of general health status and infant mortality. We find a general, but not universal, tendency for aggregate proxies to exaggerate the effects of micro-level variables and to do more poorly than micro-level variables at controlling for confounding. The magnitude and direction of these biases, however, vary across samples. Our statistical framework and empirical findings suggest the difficulties in and limits to interpreting proxies derived from aggregate census data as if they were micro-level variables. The statistical framework we outline for our study of health outcomes should be generally applicable to other situations where researchers have merged aggregate data with micro data samples.

Arline T. Geronimus
Department of Health Behavior
and Health Education
School of Public Health
University of Michigan
Ann Arbor, MI 48109-2029

John Bound
Population Studies Center
University of Michigan
1225 South University Avenue
Ann Arbor, MI 48104-2590
and NBER

Lisa J. Neidert
Population Studies Center
University of Michigan
1225 South University Avenue
Ann Arbor, MI 48104-2590

1. INTRODUCTION

To overcome limitations in data, researchers commonly augment micro data with aggregate proxies to measure quantities of interest. If the micro data set is incomplete, the researcher attributes to each missing variable a group average derived from the corresponding variable in an alternative data set. For example, economists have often used characteristics of a person's three-digit census occupation as a proxy for the characteristics of the individual's job (e.g., Lucas 1977; Olson 1981; Thaler and Rosen 1976; Viscusi 1978). Economists and education researchers have used aggregate characteristics of the schools in an individual's childhood state or school district to proxy for the characteristics of that person's particular classroom (e.g., Coleman et al., 1966; Card and Krueger 1992). Recently, it has become increasingly common for investigators of health outcomes to augment data sets by appending individual records with socioeconomic characteristics of residential areas drawn from census data (such as median income in a zip code) to proxy the individual's socioeconomic characteristics.

The growing tendency of health researchers to employ this approach is understandable. Social inequalities in health have been persistent and well-documented, but they are poorly understood (Preston and Haines 1991; Antonovsky 1967; Kitagawa and Hauser 1973; Williams 1990; Pappas et al. 1993). A pressing need exists to estimate the contribution of socioeconomic factors to health status and to disentangle their effects from those of sociodemographic factors, most prominently race. However, many sources of health data have inadequate individual socioeconomic information. For example, many investigators of infant mortality differentials analyze linked birth and infant death certificate data. The notable strengths of these data are that they are population data, they offer large numbers of births (and, thus, the power to study a rare event such as infant death), and they offer the most reliable information available on important factors such as birthweight. However, such data do not have reliable socioeconomic information, leading several researchers to employ aggregate census-based variables to proxy for individual characteristics (Wise et al. 1985; Gould and LeRoy 1988; Gould et al. 1989; Collins and David 1990; 1992; Geronimus in press).

Despite increased use of this approach, little empirical attention has been paid to the question of whether aggregate information is a valid proxy for micro-level socioeconomic characteristics.

An extensive literature exists on the relationship between coefficients estimated first on micro level and then on aggregate data (Robinson 1950; Theil 1954; Grunfeld and Griliches 1960; Duncan, Cuzzort and Duncan 1961; Hannan 1991; Hanushek, Rivkin and Taylor 1985), but little work has been done to analyze the effect of using aggregate proxies for micro quantities when analyzing micro data. In this study, we evaluate the use of aggregate measures from the census as proxies for individual socioeconomic information by comparing estimates based on such proxies to estimates using survey measures in two specific health-related examples. We build on recent work on the biases inherent in using proxies when estimating statistical relationships (e.g., Bound et al. 1994; Lee and Sepanski 1995) to consider the specific examples within the context of a statistical framework that enables us to understand the nature and source of biases involved.

The recent creation of a data set linking census information to microdata from the Panel Study of Income Dynamics (the PSID-geocode file) together with a special release of the National Maternal and Infant Health Survey (NMIHS) that includes geographic identifiers provided us with a unique opportunity to perform this analysis. We analyze these data to gauge the validity of using aggregate census information to proxy for individual characteristics and also to estimate how well such proxies serve to control for the confounding effect of socioeconomic factors and sociodemographic covariates on two important health outcomes: general health status and infant mortality.

2. STATISTICAL FRAMEWORK

We are interested in understanding the relationship between coefficients obtained using micro data versus aggregate proxies for an individual's socioeconomic characteristics when estimating health outcome equations. We do not intend to imply that micro-level measures offer the "gold standard" for measuring the full construct of interest or that any specific socioeconomic indicator (e.g., income or education) offers

the conceptually pure or complete measure of socioeconomic group. For heuristic purposes we construct the differences between the two sets of estimates (those based on aggregate versus micro-level variables) as reflecting the biases inherent in using aggregate data as a proxy for the micro-level data an investigator would have used if it had been available. However, the algebra we present is valid, independently of this interpretation. To simplify our exposition of this statistical framework we discuss a single socioeconomic indicator, but it is straightforward to generalize the expressions we present to ones that would include multiple socioeconomic indicators (Bound et al. 1994).

We are interested in estimating the relationship:

$$y_{ij} = \beta_1 x_{ij} + \beta_2 z_{ij} + \epsilon_{ij} \quad (1)$$

where i indexes individuals and j indexes identifiable groups (for example, j might index occupations or geographic locations). We do not observe z_{ij} directly, but instead observe a proxy for z_{ij}, z_j^* that varies across, but not within, groups. In the examples we will analyze, y_{ij} will represent a health outcome for individual i in location j, x_{ij} some characteristic of the individual (such as race), z_{ij} an individual socioeconomic characteristic (such as family income), and z_j^* an aggregate socioeconomic characteristic, such as the median family income for those living in the j_{th} location. In some cases, but not all, it may be appropriate to think of z_j^* as representing the within-group mean of z_{ij}.

We make the standard assumptions that $cov(x_{ij}, \epsilon_{ij}) = 0$ and $cov(z_{ij}, \epsilon_{ij}) = 0$. Thus, OLS estimates of equation 1 will give consistent estimates of β_1 and β_2. In our case, these orthogonality assumptions are partly definitional. We are interested in how well estimates using z_j^* as a proxy for z_{ij} compare to the estimates we would have obtained had we used z_{ij} itself. For the moment we also assume that $cov(z_j^*, \epsilon_{ij}) = 0$ (that the proxy z_j^* does not itself belong in equation 1).

Imagine the regression of z_{ij} on x_{ij} and then the regression of z_{ij} on x_{ij} and z_j^*:

$$z_{ij} = \gamma_1 x_{ij} + \nu_{ij} \quad (2)$$

$$z_{ij} = \gamma_1' x_{ij} + \gamma_2' z_j^* + \nu_{ij}' \quad (3)$$

where, by construction, $cov(x_{ij}, \nu_{ij}) = cov(x_{ij}, \nu'_{ij}) = cov(z^*_j, \nu'_{ij}) = 0$. Were z_{ij} perfectly correlated with z^*_j, γ'_1 would equal 0. More generally, we would expect that the inclusion of z^*_j in equation 3 would attenuate, but not eliminate the correlation between z_{ij} and x_{ij}. Thus, for example, in our context, we would expect that even within narrowly defined geographic areas, whites will tend to have higher incomes than blacks.

If we regress y_{ij} on x_{ij}, the estimated coefficient of x_{ij} is biased:

$$\text{plim} \quad \hat{\beta}_1 = \beta_1 + \beta_2\gamma_1 \ .$$

If we regress y_{ij} on x_{ij} and z^*_j, the estimated coefficients on x_{ij} and z^*_j will also give biased estimates of β_1 and β_2. In this case, it is straightforward to show that

$$\begin{aligned} \text{plim} \quad \hat{\beta}_1 &= \beta_1 + \beta_2\gamma'_1 \\ \text{plim} \quad \hat{\beta}_2 &= \beta_2\gamma'_2 \ . \end{aligned}$$

As long as γ'_1 is smaller in magnitude than γ_1, including z^*_j in the equation lessens the bias of $\hat{\beta}_1$. However, including z^*_j will not eliminate the bias on $\hat{\beta}_1$ as long as $\gamma'_1 \neq 0$. The quantity $1-\gamma'_1/\gamma_1$ gives the proportional reduction in the bias on $\hat{\beta}_1$ due to including z^*_j as a proxy for z_{ij}. Also, $1-\gamma'_2$ gives the proportional downward bias on $\hat{\beta}_2$.

So far, we have assumed that $cov(z^*_j, \epsilon_{ij}) = 0$. In many cases this assumption may not be accurate. If a typical micro-level measure, such as family income, does not itself completely capture the theoretical construct of a socioeconomic group, an aggregate variable may act as a proxy for the omitted components. For example, if over and above being poor, living in a poor neighborhood matters—then, in a statistical sense, the geocode variables "belong" in equation 1. This implies that when we enter the aggregate variable into the outcome equation, the aggregate may pick up a coefficient in its own right in addition to acting as a proxy for individual-level information.

If the assumption that $cov(z^*_j, \epsilon_{ij}) = 0$ is not valid, there is an additional source of bias on the estimated coefficients. Imagine the regression of ϵ_{ij} on x_{ij} and z^*_j:

$$\epsilon_{ij} = \delta_1 x_{ij} + \delta_2 z^*_j + \mu_i \ . \qquad (4)$$

If $cov(z_j^*, \epsilon_i) \neq 0$, then $\delta_2 \neq 0$. Additionally, if z_j^* and x_{ij} are correlated, $\delta_1 \neq 0$. In this case

$$\text{plim} \quad \hat{\beta}_1 = \beta_1 + \beta_2 \gamma_1' + \delta_1 \tag{5}$$

$$\text{plim} \quad \hat{\beta}_2 = \beta_2 \gamma_2' + \delta_2 \ . \tag{6}$$

If the partial correlations between the outcome and the individual-level variable and aggregate proxy have the same sign (i.e., δ_2 and β_2 are of the same sign), as would seem natural in our examples (that is, both wealthier individuals and residents of wealthier neighborhoods will be more likely to experience positive health outcomes than both low-income individuals or those who reside in poor neighborhoods), the omitted variable bias raises the magnitude of the estimated effect, β_2. On the other hand, the sign of δ_1 depends on both the sign of the partial correlation between y_{ij} and z_j^* (i.e., the sign of δ_2) and the sign of the correlation between x_{ij} and z_j^*. In our example, we expect the probability that a respondent is white to be positively associated with the median income in a census tract. This, together with the fact that we expect δ_2 to be positive, implies that we expect δ_1 to be negative, offsetting some of the bias resulting from the fact that the aggregate variable is only an imperfect proxy for the micro-level one.

To summarize, there are two sources of bias when using an aggregate measure to proxy for a micro variable. The first, an errors-in-variables bias, arises since the aggregate variable is only imperfectly correlated with the micro variable it is representing. The second, an aggregation bias, arises from the fact that the aggregate variable may itself be correlated with the residual in the micro-level equation. In such situations, purely micro-level and purely aggregate-level analyses produce different results (Firebaugh 1978). While it is not a mathematical necessity, it is likely that these two biases work in opposite directions. The fact that the aggregate proxy is only imperfectly correlated with the micro variable tends to exert a downward bias on the coefficient on the aggregate proxy and to imply that the aggregate variable imperfectly controls for potential confounding between the micro-level socioeconomic factor and other micro-level covariates. On the other hand, because the aggregate variable is often a proxy for a broader

construct than the micro-level variable (Hammond 1973), it picks up a larger coefficient. This second source of bias is likely to affect not only the coefficient on the proxy, but also the coefficients on other variables.

It is worth pointing out that equations 5 and 6 hold true not only in the population, but also in particular samples. Suppose we have data on both micro and aggregate proxies for z, z^m, and z^a and data on y and x. Let β_1^m and β_2^m represent OLS estimates of β_1 and β_2 based on the micro data y, x, and z^m, while β_1^a and β_2^a represent estimates based on the aggregate proxy y, x, and z^a. Then

$$\beta_1^a = \beta_1^m + \beta_2^m \hat{\gamma}_1' + \hat{\delta}_1$$
$$\beta_2^a = \beta_2^m \hat{\gamma}_2' + \hat{\delta}_2 \ ,$$

where $\hat{\gamma}_1'$, $\hat{\gamma}_2'$, $\hat{\delta}_1$ and $\hat{\delta}_2$ represent the empirical analogues to corresponding population parameters. With data that include both micro and aggregate variables, we can estimate empirical analogues to equations 1-6. Such estimates allow us to gauge the magnitude of the two biases in specific cases. Because the magnitude and even the sign of the net bias depends on a large number of factors, one would not expect results from a specific case to be directly generalizeable to another.

3. DATA

We have conducted two separate analyses to derive empirical estimates of the biases inherent in using aggregate proxies for micro-level socioeconomic characteristics in health outcome equations. To gauge whether results in a specific case are generalizable to another outcome or alternative data set, we study two outcomes (infant mortality and general health status), employing two micro-level data sets, the 1985 wave of the PSID and the NMIHS.

3.1 The Panel Study of Income Dynamics

The PSID is an ongoing longitudinal study of the determinants of family income (Hill 1992; Institute for Social Research 1988). Data from a representative sample of persons have been collected annually since 1968. In 1985, over 60 percent of the original set of sample households remained in the study. Weights have been constructed to account

for differential attrition as well as for the initial oversampling of some groups and the expansion over time of the proportion of younger families in the sample. Validation studies have documented that analyses of the PSID yield nationally representative results for blacks and whites when the sample weights are applied (Duncan & Hill 1989; Becketti et al. 1988).

We have focused on the self-reported health status among adults and have used responses to a question asking respondents to rate their health on a 5-point scale from excellent to poor. While measures based on these questions are subjective, such measures consistently have been found to be highly correlated with clinical measures (Maddox and Douglas 1973; LaRue et al. 1979; Ferraro 1980; Mossey and Shapiro 1982) and to predict subsequent death, health care utilization, and labor market behavior (Manning, Newhouse and Ware 1982)—often better than clinical measures do. For convenience we have assumed that the ordered categorical responses reflect an underlying latent continuous variable that is distributed normally, with larger values representing better health.

We restricted the study sample to the men and women from the original 1968 PSID families who identify themselves as black or white and who are between the ages of 18 and 64. Using information derived from the PSID-geocode file on the zip code of the respondent's current residence, we matched census information on median family income and mean educational attainment of the adult population to individual records. At the zip code level of aggregation, 95 percent of the 1985 PSID respondents were matched.

The PSID-geocode file also contains census tract information for respondents. Census tracts typically represent smaller and more homogeneous populations than do zip code areas. Nationally, the typical zip code contains roughly 25,000 inhabitants, while the typical census tract contains 5,000. However, rural areas are not always tracted. As a result, we were able to match only 72 percent of the PSID sample to 1985 census tract information. Regardless, results based on census tract characteristics for respondents are similar to results based on zip

code characteristics. We report results only for zip code areas.

3.2 The National Maternal and Infant Health Survey

Understanding the extent to which the association between race and infant death is confounded by socioeconomic background is a central question for investigators of the black/white disparity in infant mortality. As noted above, data commonly used to study infant mortality differentials are derived from vital statistics registrations, which include little socioeconomic information.

We analyzed a special release of the 1988 National Maternal and Infant Health Survey (NMIHS), which, unlike the public access version, includes geographic identifiers at the zip code level (U.S. Department of Health and Human Services, 1991). The NMIHS was conducted by the National Center for Health Statistics (NCHS) to study factors related to poor pregnancy outcomes. Information was gathered from vital records (birth and death certificates) as well as a mother's questionnaire. We used data from the live birth and infant death files and restricted the sample to records with complete data on birthweight, mother's age, parity, zip code of residence, household income, and household composition. We linked data on individual respondents to aggregate data (drawn from the 1980 census) on median family income and mean educational attainment of the adult population in their zip code areas. Eighty-seven percent of the NMIHS respondents had valid 1980 zip codes and were therefore matchable to aggregate census data.

The NMIHS oversampled infant deaths, but we chose not to reweight the data. Thus, there is a substantially higher fraction of black women in our sample than we would expect to find in a random sample of the population of mothers. The oversampling of infant deaths should not bias estimated logit coefficients (Farewell 1979). However, we used the within-strata weights provided on the NMIHS to correct for the explicit oversampling of blacks on the survey as well as nonresponse and other factors.

Study variables for both the PSID and NMIHS are shown in Table 1. Census STF files contain detailed tabulations of the nation's pop-

ulation and housing characteristics, offering a range of pieces of information from which to construct socioeconomic proxies. We found that different possible aggregate measures are highly correlated, including measures based on family income, education, occupation, unemployment, AFDC receipt, and poverty. We report results here using median family income in a person's zip code area as a proxy for actual family income, primarily because it is the measure used most commonly in the relevant health literature. We also report results based on the mean educational attainment in the person's zip code area as a proxy for actual educational attainment.

4. FINDINGS

4.1 The Panel Study of Income Dynamics

Table 2 shows weighted means, standard deviations, and simple first order correlations between the individual and aggregate variables used in our analysis of the PSID data. Means for comparable micro and aggregate variables are usually of similar magnitude. However, there is substantially less variation in the aggregate variables than in the micro-level variables, reflecting the fact that aggregate variables represent averages across individuals. Correlations between aggregate and micro variables are all in the expected directions. The correlation between the two aggregate variables is quite high, which suggests that the two quantities may represent different indicators of the same underlying phenomena.

Ultimately, we are interested in the validity of the inferences one might make using the aggregate variables as proxies for individual characteristics. Table 3 presents regressions of our health measure, first on demographic variables alone (race, age, and gender), then on demographic variables and family income or education measured at the individual level, and finally, on demographic variables and the log median family income or the average educational attainment in the person's zip code. We also report estimates from models where income and education measures are included together. When either income or education is entered separately, the coefficient on the aggregate variable is larger than the one on the corresponding individual variable. At the

individual level, family income and educational attainment have independent explanatory power when both are included in the regression. In contrast, neither the aggregate income variable nor the aggregate education variable has much explanatory power after the other has been included in the regression. Such findings bolster the plausibility of the suggestion that various aggregate variables may be measures of the same underlying construct—a construct that may be more global than those measured by various micro-level variables, which, while correlated, are more narrow, focused, or distinct. When looking at the effect of various socioeconomic proxies on estimated race coefficients, we see that using aggregate measures in our equations has less impact than using micro-level measures. The aggregate measures appear less adequate at controlling for confounding between family income or educational attainment and race than micro-level measures.

To understand the sources of the discrepancies between the results based on micro proxies and those based on aggregate proxies, we estimated empirical analogues to equations 2, 3, and 4. These are presented in Table 4. These results show that even after controlling for the median family income or the average educational attainment in a person's zip code, a strong association remains between race and individual family income or education (compare columns 1 and 2 or 4 and 5). As a result, we would expect to find that the aggregate variables are less adequate than the correesponding micro variables as controls for the confounding effect of income on the association between race and health.

The third column of Table 4 gives results from regressing the residual (from the regression of health on family income measured at the individual level) on the demographic variables and the aggregate income proxy. The residual is positively correlated with the aggregate variable, but it is weakly and negatively correlated with the race dummy. The fact that the aggregate variable is correlated with the residual from the equation of interest tends to raise the estimated coefficient on the socioeconomic indicator, and to lower the magnitudes of the coefficients on the sociodemographic variables when an aggregate variable is used to substitute for a micro-level one.

The fourth, fifth, and sixth columns of Table 4 show results from replicating the family income analyses using education variables instead. Here we see that controlling for the mean education in a person's zip code area reduces the association between race and education by roughly 50 percent. At the same time, regressing the residual (from the regression of health on race and education) on race and aggregate education shows the residual to be positively correlated with the education variable, but negatively correlated with the race variable. Again, we see the errors-in-variable and the aggregation biases working in opposite directions. By looking back at Table 3, we can see that the net result—whether we are talking about education or income—is that the aggregate variable picks up a larger coefficient than the corresponding micro variable, but it does less well at controlling for the confounding effects of socioeconomic factors on the association between race and overall health.

To understand the bias inherent in the simultaneous inclusion of both aggregate proxies, we have to regress the individual income and education variables on the aggregate income and education variables. Results from these regressions are presented in the seventh and eighth columns of Table 4. Note that in these regressions the association between race and family income, on the one hand, and between race and educational attainment, on the other, is virtually as strong when controlling for both aggregate variables as when controlling for only one. The last column in the table shows the results from the regression of the residual (from the regression of health on family income and education) on both aggregate variables. The residual continues to be correlated with the aggregate variables. Again, the implication from the results reported in columns 7, 8, and 9 is that the errors-in-variables bias works in the opposite direction from the aggregation bias.

4.2 The National Maternal and Infant Health Survey

Means, standard deviations, and correlations for the NMIHS data are reported in Table 5. Because the NMIHS sample is composed of mothers, it is a younger population than the PSID sample. The oversampling of infant deaths in the sample raises the fraction of the moth-

ers who are black. However, the correlations between the aggregate and micro variables in the NMIHS sample are similar to those in the PSID sample.

Table 6 shows the estimated logistic regression coefficients for the effect of various factors on infant survival. Here, the coefficient on the aggregate income variable (0.195) is more than twice the size of that on the individual income variable (0.081). For education, coefficients based on aggregate and micro-level measures are similar (0.064 and 0.052), although the micro-level coefficient is larger in this case. At the same time, while both micro-level variables appear to control the confounding effects of socioeconomic group and maternal age better than the aggregate proxies, the aggregate education variable controls the confounding of socioeconomic group and race better than the respective micro variables do.

Changes in logistic regression coefficients across the various specifications cannot be decomposed in the way the linear regression coefficients can. Still, analogues to equations 2 and 3 provides valuable information about the differences between the various specifications. Table 7 reports results from the regression of family income and educational attainment on demographic and socioeconomic proxy variables. Including the aggregate proxies has a greater effect on the race coefficient than on the maternal age coefficient. In fact, there appears to be little association between race and educational attainment after controlling for age, parity, and mean educational attainment in the zip code area. This explains why, in this sample, the aggregate education variable seems to do as well as the individual education variable at controlling for the confounding of socioeconomic group and race. On the other hand, the coefficient on average educational attainment itself is substantially less than 1, implying a downward bias on the aggregate education variable that is only partially addressed by the fact that the aggregate education variable is associated with infant survival, even after controlling for the actual education of the mother.

The differences between our results for the PSID and NMIHS with respect to the education variables are noteworthy. Changing samples

and outcome variables changed not only the magnitude, but also the direction of the net "bias" involved when mean education was used as a proxy for individual education! The difference between the two sets of results appears to be explained by the fact that the NMIHS sample is much younger than the PSID sample. Black/white differences in educational attainment are substantially narrower for more recently born cohorts. Thus, the raw racial difference in educational attainment is much smaller in the NMIHS (0.22 years) than in the PSID (1.14 years). As a result, educational differences account for very little of the black/white differences in outcomes for younger cohorts. Moreover, less variance in educational attainment among more recent cohorts serves to lower the coefficient on the aggregate education variable when individual education is regressed on demographic characteristics and the aggregate education variable in the NMIHS sample. In any case, the differences between the two sets of results indicate the variation across samples in the magnitude and even the direction of the discrepancies that result from using aggregate variables as proxies for micro-level characteristics.

5. DISCUSSION

Results from our analyses of PSID and NMIHS data are not neatly summarized. They highlight the pitfalls of employing aggregate census-based variables as proxies for micro-level socioeconomic characteristics. Aggregate proxies tend to exaggerate the effect of individual education or family income on health outcomes, while not adequately controlling for the confounding between micro-level socioeconomic characteristics and other variables of interest, most notably race. However, we have seen one instance where this is not true. When we estimated the effect of education on infant survival we found the aggregate variable picking up a slightly smaller coefficient than the corresponding micro variable, and performing better than the micro variable at controlling for the confounding effect of socioeconomic group on the association between race and infant survival. The discrepancy between the PSID and NMIHS results on education highlights the fact that the difference between coefficients based on micro and aggregate socioeconomic proxies depends on both the sample and the variables used.

Our results suggest the difficulties in interpreting socioeconomic proxies derived from aggregate census data as if they were micro-level variables. Inferences must be drawn cautiously. Researchers are on firmest ground if they interpret such results in a literal way. For example, the findings in our examples could be interpreted descriptively as reflecting health differentials between residents of more- and less-advantaged areas. However, when this methodology is employed, the explanations for these differences must remain speculative.

By suggesting that interpretations be literal, we are not suggesting that, in this context, coefficients on aggregate variables be interpreted as "contextual" or "neighborhood" effects. Our analysis speaks to the circumstance where the micro-level variable is unavailable and only an aggregate proxy is included in estimated models. As we have shown, in such a case it would be problematic to interpret the estimate as a contextual effect because, in the absence of a micro-level measure, the aggregate measure picks up individual as well as contextual effects.

Our findings are inconsistent with the conclusion Krieger (1992) drew from her analysis of a select HMO sample in Northern California—that aggregate census-based proxies are good substitutes for micro-level measures of individual socioeconomic characteristics. By obtaining similar point estimates of the relative risks of health outcomes by socioeconomic group, whether she used aggregate census-based measures or the crude individual-level measures available in her data, Krieger demonstrated the relative usefulness of the aggregate census-based approach in her study sample. However, our statistical framework and empirical findings lead us to question whether such findings imply the general validity of using aggregate census-based measures to proxy for individual characteristics. Because Krieger's empirical work was not interpreted in light of a statistical framework, the nature and source of biases could not be discussed. Now that we are able to place her results in the context of such a framework, we would argue that finding consistent results between micro and aggregate variables is the exception, not the rule.

One might expect that aggregate variables based on more narrowly defined geographic areas—such as census tracts or block groups—would

be better proxies for micro-level characteristics than ones measured at the zip code level. However, we were able to replicate the PSID analysis using census tract data and to obtain qualitatively similar results to the zip code level analysis. Krieger (1992) compared census tract to block group level results in her HMO analysis, finding little difference between the two sets of results. We do not have empirical evidence to rule out the possibility that block group data for a less select sample might better represent individual characteristics than census tracts or zip codes. Pragmatically, census block-level data are rarely available to researchers and do not exist for rural areas. Our analysis is consistent with the kind of data available to most investigators—particularly investigators who have employed the census-based strategy.

Although we have focused on the use of aggregate socioeconomic proxies for estimating health outcome equations, the framework we have developed should be generally applicable to other situations where researchers have merged aggregate data into micro data samples. In such situations, researchers generally are not in a position to use micro data to validate inferences based on aggregate data. The most conservative lesson we have drawn from these results is that investigators are ill-advised to interpret results of regressions based on aggregate variables as if they were based on micro-level variables. However, the framework we have presented can be used to guide thinking about the likely direction of bias in such estimates. In situations where important variation exists in the relevant independent variables within aggregate units, the use of aggregate proxies will tend to yield underestimates of the effect of the micro variable, while inadequately controlling for confounding effects. However, in cases where the aggregate variable might represent a broader construct than the micro-level construct, estimates based on the aggregate data are likely to exaggerate the effect of the micro-level counterpart on outcomes of interest.

Since Robinson's classic paper outlining the "ecological fallacy" (Robinson 1950), researchers have been wary of interpreting estimates based on aggregate data. The results reported here suggest that this should be a concern not only in the case where the unit of analysis is an aggregate unit, but also in the case where the unit of analysis is a micro

unit and aggregate variables are used to proxy micro-level constructs.

References

Antonovsky, A. (1967), "Social Class, Life Expectancy and Overall Mortality," *Milbank Memorial Fund Quarterly*, 45, 31–73.

Becketti S., Gould, W., Lillard, L., and Welch, F. (1988), "The Panel Study of Income Dynamics after Fourteen Years: An Evaluation," *Journal of Labor Economics*, 6, 472–492.

Bound, J., Brown, C., Duncan, G., and Rodgers, W. (1994), "Evidence in the Validity of Cross-sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, 12, 345–368.

Card, D. and Krueger, A. (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100, (1), 1–40.

Coleman, J. S., et al. (1966), *Equality of Educational Opportunity*. Washington, DC: Government Printing Office.

Collins, J. W. and David, R. J. (1992), "Differences in Neonatal Mortality by Race, Income, and Prenatal Care," *Ethnicity Dis*, 2, 18–26.

_____. (1990), "The Differential Effect of Traditional Risk Factors on Infant Birthweight among Blacks and Whites in Chicago," *American Journal of Public Health*, 80, 679–681.

Duncan, G. J. and Hill, M. (1989), "Assessing the Quality of Household Panel Survey Data: The Case of the PSID," *Journal of Business and Economic Statistics*, 7, 441–451.

Duncan, O. D., Cuzzort, R. P., and Duncan, B. (1961), *Statistical Geography: Problems in Analyzing Areal Data*. Glencoe: Free Press.

Farewell, V. T. (1979), "Some Results on the Estimation of Logistic Models Based on Retrospective Data," *Biometrica*, 66, (1), 27–32.

Ferraro, K. F. (1980), "Self-ratings of Health Among the Old and Old-Old," *Journal of Health and Social Behavior*, 21, 377–383.

Firebaugh, G. (1978), "A Rule for Inferring Individual-level Relationships from Aggregate Data," *American Sociological Review*, 43 (August), 557–572.

Geronimus, A. T. (in press), "Black/White Differences in the Relationship of Maternal Age to Birthweight: A Population-based Test of the Weathering Hypothesis," *Social Science and Medicine.*

Gould, J. B., Davey, B., and LeRoy, S. (1989), "Socioeconomic Differentials and Neonatal Mortality: Racial Comparison of California Singletons," *Pediatrics*, 83, 181–186.

Gould, J.B. and LeRoy, S. (1988), "Socioeconomic Status and Low Birth Weight: A Racial Comparison," *Pediatrics*, 82, 896–904.

Grunfeld, Y. and Griliches, Z. (1960), "Is Aggregation Necessarily Bad?" *Review of Economics and Statistics*, pp. 1–3, February.

Hammond, J. L. (1973), "Two Sources of Error in Ecological Correlations," *American Sociological Review*, 38, 764–777.

Hannan, M. T. (1991), *Aggregation and Disaggregation in the Social Sciences*, Revised Edition. Lexington, MA: Lexington Books, 144 pages.

Hanushek, E. A., Rivkin, S. Q., and Taylor, L. L. (1995), "Aggregation and the Estimated Effects of School Resources," Rochester Center for Economic Research Working Paper No. 397.

Hill, M. S. (1992), *The Panel Study of Income Dynamics: A User's Guide.* Newbury Park, CA: Sage Publications.

Institute for Social Research (1988), *A Panel Study of Income Dynamics: Procedures and Tape Codes*, 1985 Interviewing Year (Documentation), Volume I, Wave XVIII, A Supplement. Ann Arbor, MI: Institute for Social Research, University of Michigan.

Kitagawa, E. M. and Hauser, P. M. (1973), *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*, Cambridge, MA, Harvard University Press.

Krieger, N. (1992), "Overcoming the Absence of Socioeconomic Data in, Medical Records: Validation and Application of a Census-based Methodology," *American Journal of Public Health*, 92, 703–710.

LaRue, A., L. Bank, L. Jarvic, and M. Hewtland (1979), "Health in Old Age: How Physicians' Ratings and Self-ratings Compare," *Journal of Gerontology*, 34 (September), 687–691.

Lee, L. and Sepanski, J. H. (1995), "Consistent Estimation of Linear and Nonlinear Errors-in-Variables Models with Validation Information," *Journal of the American Statistical Association*, 90 (No, 429, March), 130–140.

Lucas, R. E. B. (1977), "Hedonic Wage Equations and Psychic Wages in the Return to Schooling," *American Economic Review*, 67 (No. 4, September), 549–558.

Maddox, G. and E. Douglas (1973), "Self-Assessment of Health: A Longitudinal Study of Elderly Subjects," *Journal of health and Social Behavior*, 14 (1), 87–93.

Manning, Jr., W. G., Newhouse, J. P., and Ware, Jr., J. E. (1982), "The Status of Health in Demand Estimation; or, Beyond Excellent, Good, Fair, and Poor," in *Economic Aspects of Health*, edited by Victor R., Fuchs, National Bureau of Economic Research. Published by the University of Chicago Press.

Mossey, J. M. and E. Shapiro (1982), "Self-Rated Health: A Predictor of Mortality among the Elderly," *American Journal of Public Health*, 72, 800–808.

Olson, C. A. (1981), "An Analysis of Wage Differentials Received by Workers on Dangerous Jobs," *Journal of Human Resources*, 16 (No. 2, Spring), 167–185.

Pappas, G., Queen, S., Hadden, W., and Fisher, G. (1993), "The Increasing Disparity in Mortality between Socioeconomic Groups in the United States, 1960 and 1986," *New England Journal of Medicine*, 329, 103–109.

Preston, S. H. and Haines, M. (1991), *Fatal Years: Child Mortality in Late Nineteenth-Century America.* Princeton, NJ: Princeton University Press.

Robinson, W. S. (1950), "Ecological Correlations and the Behavior of Individuals," *American Sociological Review*, 15, 351–357.

Solon, G. (1992), "Intergenerational Income Mobility in the United States," *American Economic Review*, 82, 393–408.

Thaler, R. and Rosen, S. (1976), "The Value of Saving a Life: Evidence from the Labor Market," in *Household Production and Consumption*, edited by Nestor Terleckyj, NBER Studies in Income and Wealth No. 90. New York: Columbia Press, pp. 265–298.

Theil, H. (1954), *Linear Aggregation of Economic Relations.*Amsterdam: North-Holland Publishing Company.

U.S. Department of Health and Human Services, Public Health Service, National Center for Health Statistics (August 1991), *Public Use Data Tape Documentation: 1988 National Maternal and Infant Health Survey.* Hyattsville, MD: Centers for Disease Control, National Center for Health Statistics.

Viscusi, W. K. (1978), "Wealth Effects and Earnings Premiums for Job Hazards," *Review of Economics and Statistics*, 60 (No.3, August), 408–416.

Williams, D. R. (1990), "Socioeconomic Differentials in Health: A Review and Redirection," *Social Psychology Quarterly*, 53, 81–99.

Wise, P. H., Kotelchuck, M., Wilson, M. L., and Mills, M. (1985), "Racial and Socioeconomic Disparities in Childhood Mortality in Boston," *New England Journal of Medicine*, 313, 360–366.

Table 1. PSID and Census Variable Descriptions

Individual	Description
Health:	Self-Rated Health on a 5-point scale from Excellent to Poor. Recoded under the assumption that the 5 categories reflect an underlying standard normal latent variable, with positive values reflecting good health.
Survival:	Infant survival through first year. 1 = survived; 0 = died.
White:	Race of respondent. 1 = white; 0 = Black.
Age:	Age of respondent.
Income:	Natural Logarithm of family income defined in terms of 1980 dollars.
Education:	Educational Attainment in years of respondent.
Aggregate	
Income:	Log of Median Family Income in the respondents zip code area.
Education:	Mean educational attainment for those aged 25+ in the respondents zip code area.

Table 2. Simple Statistics, PSID Sample

Variable	Mean	Standard Deviation	Pearson Correlation Coefficients: Individual: Health	White	Income	Education	Aggregate: Income	Education
Individual								
Health	0.00	0.94	1.00					
White	0.89	0.32	0.15	1.00				
Income	10.19	0.83	0.21	0.27	1.00			
Education	12.88	2.59	0.37	0.13	0.33	1.00		
Aggregate								
Income	9.91	0.28	0.20	0.30	0.39	0.31	1.00	
Education	11.84	1.23	0.22	0.20	0.32	0.39	0.77	1.00

Notes: Weighted, using PSID sample weights. Sample size = 6,454.

Table 3. The Effects of Race and SES on General Health

Explanatory Variable	None	SES Proxy					
		Individual			Aggregate		
White	0.468	0.279	0.339	0.244	0.318	0.352	0.332
	(0.035)	(0.035)	(0.035)	(0.034)	(0.036)	(0.035)	(0.036)
Income		0.279		0.174	0.558		0.154
		(0.014)		(0.014)	(0.040)		(0.062)
Education			0.113	0.094		0.146	0.120
			(0.004)	(0.004)		(0.009)	(0.014)
R^2	0.102	0.156	0.193	0.211	0.127	0.137	0.137

Notes: All equations include controls for age and gender. Standard errors are in parenthesis. Weighted using PSID sample weights. Sample size = 6,454.

Table 4. Decomposing the Bias Inherent in Using an Aggregate Proxy, PSID Sample

Explanatory Variable	SES Proxy: Income			SES Proxy: Education			SES Proxy: Income & Education		
	Individual Income	Income	Residual	Individual Education	Education	Residual	Income	Education	Residual
White	0.676	0.412	−0.086	1.140	0.522	−0.051	0.416	0.499	−0.021
	(0.031)	(0.031)	(0.043)	(0.100)	(0.094)	(0.041)	(0.031)	(0.097)	(0.042)
Aggregate Proxy									
Income		0.983	0.318				0.825	0.179	−0.043
		(0.034)	(0.048)				(0.052)	(0.164)	(0.071)
Education					0.779	0.068	0.047	0.749	0.054
					(0.024)	(0.011)	(0.012)	(0.037)	(0.016)
R^2	0.116	0.218	0.007	0.058	0.189	0.006	0.212	0.189	0.003

Notes: All equations include controls for age and gender. Standard errors are in parenthesis. Weighted using PSID sample weights. Sample size = 6,454.

Table 5. Simple Statistics, NMIHS Sample

Variable	Mean	Standard Deviation	Pearson Correlation Coefficients						
				Individual				Aggregate	
			Survival	White	Age	Income	Education	Income	Education
Individual									
Survival	0.33	0.47	1.00						
White	0.77	0.42	0.14	1.00					
Age	25.96	5.68	0.09	0.19	1.00				
Income	9.55	1.38	0.12	0.36	0.40	1.00			
Education	12.79	2.23	0.11	0.17	0.43	0.41	1.00		
Aggregate									
Income	9.85	0.29	0.09	0.35	0.26	0.35	0.30	1.00	
Education	11.63	1.19	0.08	0.24	0.24	0.32	0.35	0.76	1.00

Notes: Weighted to correct for differential nonresponse, but not to correct for the oversampling of infant deaths. Sample size = 10,787.

Table 6. The Effects of Race, Age and SES on Infant Survival

Explanatory Variable	None	SES Proxy					
		Individual			Aggregate		
White	0.563	0.493	0.552	0.501	0.523	0.537	0.527
	(0.049)	(0.051)	(0.049)	(0.051)	(0.051)	(0.050)	(0.051)
Age	0.040	0.028	0.023	0.019	0.033	0.033	0.033
	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
Income		0.081		0.061	0.195		0.081
		(0.017)		(0.018)	(0.078)		(0.112)
Education			0.064	0.055		0.052	0.038
			(0.011)	(0.011)		(0.019)	(0.027)
χ^2	322.1	344.5	355.7	367.6	328.4	329.9	330.4

Notes: All equations include controls for parity. Standard errors in parenthesis. Weighted to correct for differential nonresponse, but not to correct for the oversampling of infant deaths. Sample size = 10,787.

Table 7. Decomposing the Bias Inherent in Using an Aggregate Proxy, NMIHS Sample

Explanatory Variable	Dependent Variable: Individual Income			Dependent Variable: Individual Education		
White	0.873	0.708	0.086	0.219	0.006	0.183
	(0.028)	(0.029)	(0.002)	(0.045)	(0.045)	(0.004)
Age	0.098	0.088	0.721	0.204	0.182	0.026
	(0.002)	(0.002)	(0.029)	(0.004)	(0.004)	(0.046)
Aggregate Proxy						
Income		0.804	0.452			−0.172
		(0.042)	(0.060)			(0.097)
Education			0.116		0.419	0.449
			(0.014)		(0.016)	(0.023)
R^2	0.265	0.290	0.294	0.260	0.305	0.305

Notes: All equations include controls for parity. Standard errors in parenthesis. Weighted to correct for differential nonresponse, but not to correct for the oversampling of infant deaths. Sample size = 10,787.