SMALL AREA ESTIMATION: EMPIRICAL EVALUATION OF SEVERAL ESTIMATORS FOR PRIMARY SAMPLING UNITS

Dwight B. Brock, National Center for Health Statistics
Dwight K. French, Energy Information Administration
Barry W. Peyton, Center for Disease Control

I. INTRODUCTION

For a number of years the National Center for Health Statistics, the Bureau of the Census and other agencies have studied the problem of providing estimates of social and health characteristics for small geographic areas. Most of these efforts have been applied to surveys designed to make only national or regional estimates, primarily because of the expense involved in designing and conducting surveys specific to small areas. In the case of disaggregating national data, the estimators used are generally faulty because the samples on which they are based are small, unrepresentative (of the small area) or nonexistent. A number of approaches have been taken to construct and evaluate estimates for these situations. The first, and most widely reported, technique is the "synthetic" estimator (NCHS, 1968; Levy, 1971; Gonzalez, 1973; Levy and French, 1977; NCHS, 1977; and Gonzalez and Hoza, 1978). The difficulty with the synthetic estimators is that they all have an unknown and generally inestimable bias, though they have been useful in providing early evidence of the feasibility of disaggregating national survey data.

Other authors (Gonzalez and Waksberg, 1973; Schaible, Brock and Schnack, 1977a, among others) have argued that in some instances the sample sizes in the small areas of interest are sufficiently large to make estimates for those areas based directly on the observations available for the areas. Their results showed that in some situations the direct estimators outperformed the synthetic whereas in others the synthetic estimators were better. It has been suggested that the question of when to use which type of estimator need not be answered, if one is willing to use a combination of estimators (see, for example Royall, 1973).

Investigations of the use of combinations of estimators for small areas have been reported by Schaible, Brock and Schnack (1977b) and Fay and Herriot (1979). The former used a particular composite estimator consisting of a weighted sum of synthetic and direct estimators, whereas the latter discussed the use of a James-Stein type of estimator for the production of estimates of per capita income. Looney and Brock (1979) described a modification of a James-Stein type estimator for small area estimation, and Royall (1980) has recently made a comparison of James-Stein and composite estimators under linear prediction theory models.

Two other papers in small area estimation deserve special mention. The first is an expository paper by Purcell and Kish (1979) which gives a comprehensive bibliography on the major studies of small area estimators published in the past 10 to 15 years. The second paper, Schaible (1978), describes the details of the minimum-mean-square-error (MMSE) composite estimator used by Schaible, Brock and Schnack (1977b) and in the NCHS publication, State Estimates of Disability and Utilization of Medical Services, 1974-76 (1978). In his paper Schaible developed weights to be applied to the synthetic and direct components of the composite estimator which minimizes its mean square error. Furthermore, he gave conditions under which the optimum weight could be approximated, and described empirically the robustness of the composite estimator against poor estimates of the optimum weight when making estimates for States.

The purpose of our paper is to describe further evaluation of the MMSE composite estimator, using Schaible's results as a point of departure. The choice of component estimators is discussed, a method of estimating the optimum weight for the composite estimator from sample data is described, and the procedures are illustrated and compared using data from the 1969-71 Health Interview Survey (HIS) and the 1970 Census. The small areas under consideration in this paper are HIS Primary Sampling Units (PSU's), which consist of counties, groups of counties and Standard Metropolitan Statistical Areas (SMSA's).

II. ESTIMATORS AND NOTATION

The notation of Schaible (1978) will be followed in this paper as far as possible. The composite estimator studied here takes the same form as that described by Schaible. Let Y'_d and Y''_d be separate estimators for $\bar{Y}_d$, the population value of the characteristic Y for small area d. The composite estimator then takes the form

$$\hat{\bar{Y}}_d = C_d Y'_d + (1-C_d) Y''_d , \qquad (1)$$

where C_d is the weight used in combining the two component estimators.

A variety of choices of component estimators Y'_d and Y''_d was available for use in the composite estimator. In addition to the simplest forms of direct and synthetic estimators, numerous estimators were formed by making modifications and adjustments to these basic estimators. Also, the HIS regional estimate was used to represent each PSU in a given region. Finally, two forms of regression estimators were constructed from auxiliary census information available for each PSU. One type was simply a preducted value based on the best fit obtained through ordinary stepwise regression. The other was a regression adjusted synthetic estimator constructed along the lines of that described by Levy (1971). In all, 19 different estimators were considered as possible candidates for components in the composite estimator. To facilitate the comparison of this large number of estimators in the determination of the best components of the composite estimator, a procedure similar to that followed by Schaible, Brock, Casady and Schnack (1979) was used. The results of this part of the investigation are summarized in Section III.

Before proceeding to those results, however, let us return to the specification of the basic estimators considered as components. Let $y_{d\alpha i}$ denote the observation of interest on the ith individual in αth demographic cell in small area d. Here $i=1, \ldots, n_{d\alpha}$; $d=1, \ldots, D$; and $\alpha=1, \ldots, K$. The simple direct estimator (SD) for small area d is defined as

$$Y_d' = \sum_{\alpha=1}^{K} \sum_{i=1}^{n_{d\alpha}} y_{d\alpha i}/n_d , \qquad (2)$$

where n_d is the full sample size (over all α-cells) in area d. Now let $N_{d\alpha}$ represent the population size in area d and class α. The large area (national or regional) sample mean for the αth class is

$$\overline{Y}_\alpha = \sum_{d=1}^{D} \sum_{i=1}^{n_{d\alpha}} y_{d\alpha i}/n_\alpha , \qquad (3)$$

where D is the number of small areas in the large area and n_α is the large area sample size in class α. Then, the simple synthetic estimator for small area d is defined to be

$$Y_d'' = \sum_{\alpha=1}^{K} \frac{N_{d\alpha}}{N_d} \overline{Y}_\alpha , \qquad (4)$$

where N_d is the total population of interest in area d. The α-cells used to construct the basic synthetic estimator examined in this paper are the 12 cells created by cross-classifying the following variables.

1. Color: white; all other
2. Sex: male; female
3. Age: 15-44 years; 45-64 years; 65 years and over, for labor force variables; 25-44 years; 45-64 years; 65 years and over, for all other variables.

The basic regression estimator was constructed as follows. First the simple direct estimates for all the small areas were regressed on a set of appropriate independent variables available from the census for the small areas (see the Appendix for details). Stepwise methods were used to obtain the best possible fit from the variables that were available, using BMDP2R (Brown, 1977). Then the regression estimator was defined to be

$$Y_d''' = b_0 + b_1 X_{1d} + b_2 X_{2d} + \ldots + b_p X_{pd} , \qquad (5)$$

where the b_i are the estimated coefficients based on the fit of all the data, and X_{id} are the observed values of the independent variables for area d. Finally, the regression adjusted synthetic estimator was constructed by including Y_d'' as defined in equation (4) as an additional independent variable in equation (5).

III. DATA SOURCES AND THE CHOICE OF COMPONENT ESTIMATORS

To determine which small area estimators were most appropriate for use in the composite estimator, the individual estimates were compared empirically to find out which performed best as individual small area estimators. The procedure was to estimate known population values for each primary sampling unit, thus allowing the calculation of actual errors to compare the estimators' performance. The known population values were taken from the Census County and City Data Book (CCDB) (1972), based on the 1970 Census. Comparable variables from the combined 1969-71 National Health Interview Survey were used to calculate the estimates. The variables reported in this paper are: (1) percent of the population over 25 who completed high school, (2) percent of the civilian labor force unemployed, (3) percent of the civilian labor force in manufacturing, and (4) percent of the civilian labor force in sales and clerical work. Hereafter these variables will be referred to as "high school", "unemployment", "manufacturing", and "sales and clerical."

The measuring device used to make these first-level comparisons is the average squared error (ASE), the average over the appropriate set of PSU's of the squared difference between each estimate and its corresponding PSU census value. Average squared errors were computed for each variable/estimator combination for each of three sets of PSU's (or domains): (1) the 21 largest SMSA's*, (2) the set of HIS self-representing PSU's, and (3) the set of all HIS PSU's. The results of these calculations are shown in Table 1.

Examination of the figures in Table 1 shows that between the two direct estimators, the simple direct (eq. 2) generally outperformed the post-stratified. Among the synthetic/regression estimators the simple regression estimator provided the most reliable estimates overall. Thus, the simple direct and simple regression estimators have been chosen as components for the composite estimator. It is noteworthy and appealing from a practical standpoint that these estimators are also the easiest to compute!

IV. ESTIMATION OF WEIGHTS FOR THE COMPOSITE ESTIMATOR

Schaible (1978) showed the optimum weight for the composite estimator (1) as

$$C_d^* = \frac{\text{MSE } Y_{2d}' - E(Y_{1d}' - \overline{Y}_d)(Y_{2d}' - \overline{Y}_d)}{\text{MSE } Y_{1d}' + \text{MSE } Y_{2d}' - 2\,E(Y_{1d}' - \overline{Y}_d)\,(Y_{2d}' - \overline{Y}_d)} \qquad (6)$$

*The Boston Mass. SMSA was excluded from these computations because its boundaries do not follow county lines, thus making direct comparisons with the CCDB extremely difficult.

where Y'_{1d} and Y'_{2d} are the component estimators for the composite estimator. He further stated that if the component estimators are independent and if either is unbiased, then C^*_d can be readily estimated. In our case if Y'_{1d} is the regression estimator (5) and Y'_{2d} is the simple direct estimator (2), these requirements are approximately satisfied, and estimates of C^*_d can be obtained from sample data. The HIS design calls for one PSU per stratum with independent selection among the strata. Thus the simple direct estimator for a PSU is essentially unbiased. Furthermore, it can be argued that the simple direct and regression estimators are essentially independent because of the way in which the regression coefficients were estimated. For a given small area d, the only common ingredient of Y'_d (2) and Y'''_d (5) is but one of 374 observations used in the estimation of the b's in equation (5). If there is any dependence, it would necessarily be small.

We shall hereafter assume that Y'_d and Y'''_d are independent and proceed to estimate C^*_d. This implies that the cross product expectations in (6) drop out and we must estimate only

$$C'_d = \frac{\text{MSE } Y'_{2d}}{\text{MSE } Y'_{1d} + \text{MSE } Y'_{2d}} \quad . \qquad (7)$$

Since Y'_{2d} (simple direct estimator) is approximately unbiased, its mean square error can be estimated by its variance. Estimates of variance for percentages, available from Wilder (1973), were used for this project. The only remaining quantity to be estimated is MSE Y'_{1d}. This estimation was accomplished using the results of Theorem 6 in Levy and French (1977), namely,

$$\widehat{\text{MSE}}\; Y'_{1d} = (Y'_{2d} - Y'_{1d})^2 - \widehat{\text{Var}}\;(Y'_{2d}), \qquad (8)$$

since Y'_{2d} is an unbiased estimator for the small area statistic $\overline{Y}_d$. Thus the estimated weights for the composite estimator (1) take the form

$$\hat{C}_d = \frac{\widehat{\text{Var}}\;(Y'_{2d})}{(Y'_{2d} - Y'_{1d})^2} \qquad (9)$$

Since we are estimating percentages, we restrict $\hat{C}_d$ to the interval $[0,1]$ so that our estimates do not exceed the range 0% to 100%. Thus if $\hat{C}_d > 1$ (or $\hat{C}_d < 0$) we set it equal to 1 (0) and proceed with the computation of the composite estimates.

V. RESULTS

The results of the computations described in the previous section are summarized in Table 2. Graphical illustration of the relationship between the census values and the three estimators for the variable "high school" are given in the figures following Table 2. Lack of space prevented the presentation of more variables. In addition to the scatter plots, the figures contain the summary statistics of Table 2 and an indication of the amount of actual error in each estimate. The latter is represented by the vertical distance from the plotted point to the 45-degree line bisecting each plot.

For the 21 large SMSA's, the simple direct estimator performed quite well, showing average squared errors which ranged from 1.6 percent for "unemployment" to 6.7 percent for "high school." This was expected since the sample sizes in these areas are quite large, ranging from 2,063 to 30,546. In fact a form of this estimator has been used for large SMSA's in practice (Gentile, 1977). However, for three out of the four variables the composite estimator outperformed the simple direct even with these large sample sizes.

As can be seen in Table 2, the results were not so positive for the set of self-representing PSU's. Overall the average squared errors were higher, undoubtedly because the average sample size per PSU is much smaller for this domain. The performance of the composite estimator in this domain was mixed, being better for "high school" but worse for "manufacturing." For "unemployment" the regression estimator had the lowest ASE, but the composite estimator provided a better linear fit. This is similar to the situation observed by Schaible, Brock and Schnack (1977b) for Texas PSU's.

The performance of these estimators continues to deteriorate as more PSU's are added to the domain. For the group of all PSU's the average squared errors are higher overall than those for the self-representing PSU's, the primary reason being the continued decrease in average PSU sample size. The composite estimator offers no improvement over the component estimators and is in fact slightly worse than the component estimators in two cases. Investigation of these cases indicates that for the PSU's with small sample sizes, the numerator of the weight in equation (9) becomes extremely unstable thus suggesting poor performance of the composite estimator in some instances.

VI SUMMARY

We have seen that it is possible in practice to construct a composite small-area estimator from sample data which can provide improved estimates of population characteristics for some situations. The results of Schaible (1978) indicate that if appropriate weights for the component estimators can be estimated, this improved estimation would result for almost all situations. A problem remaining, then, is to develop appropriate weighting schemes for situations in which component estimators are based on small sample size. It may be possible, for example, to "smooth" the weights by averaging, differencing, or model fitting. There may be other as yet undiscovered approaches to improved estimation of weights. We are continuing our investigations in these areas.

Thus, as the demand for small-area data based on large national systems persists, and as budgets at the Federal, State and local governmental levels continue to tighten, the need for improved small area estimation methods will probably increase. Among other topics under consideration for future work include the provision of measures of error for individual small area statistics, the identification of additional sources of independent variables for improved regression estimators and the use of supplemental telephone surveys in a given small area to increase sample sizes.

References

Brown, M. B., ed. (1977). Biomedical Computer Programs, P-Series, Berkeley, University of California Press.

Fay, R. E. and Herriot, R. A. (1979). "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." J. Amer. Statist. Assoc. vol. 74, pp. 269-277.

Gentile, A. (1977). "Health Characteristics by Geographic Region, Large Metropolitan Area, and Other Places of Residence." Vital and Health Statistics, Series 10-113 (No. PHS-78-1540), National Center for Health Statistics.

Gonzalez, M. E. (1973). "Use and Evaluation of Synthetic Estimates." Proceedings of the American Statistical Association, Social Statistics Section, pp. 33-36.

Gonzalez, M. E. and Hoza, C. (1978). "Small Area Estimation with Application to Unemployment and Housing Estimates." J. Amer. Statist. Assoc., vol. 73, pp. 7-15.

Gonzalez, M. E. and Waksberg, J. (1973). "Estimation of the Error of Synthetic Estimates." Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

Levy, P. S. (1971). "The Use of Mortality Data in Evaluating Synthetic Estimates." Proceedings of the American Statistical Association, Social Statistics Section, pp. 328-331.

Levy, P.S. and French, D. K. (1977). "Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey." Vital and Health Statistics Series 2-75 (No. PHS-78-1349), National Center for Health Statistics.

Looney, S. W. and Brock, D. B. (1979) "Application of James-Stein and Related Techniques to Small Area Estimation Problems of the National Center for Health Statistics." Abstract, Biometrics, vol. 35, p. 891.

National Center for Health Statistics (1968). "Synthetic Estimates of Disability." Public Health Service Publication No. 1759.

National Center for Health Statistics (1977). "State Estimates of Disability and Utilization of Medical Services: United States, 1969-71." DHEW Publication No. HRA-77-1241.

National Center for Health Statistics (1978). "State Estimates of Disability and Utilization of Medical Services: United States, 1974-76." DHEW Publication No. PHS-78-1241.

Purcell, N. J. and Kish, L. (1979). "Estimation for Small Domains." Biometrics. vol 35, pp.365-84

Royall, R. M. (1973). "Discussion of two papers on Recent Development in Estimation for Local Areas." Proceedings of the American Statistical Association, Social Statistics Section, pp. 43-44

Royall, R. M. (1980). "Weighted Least Squares, Bayes, and James-Stein Prediction Techniques in Finite Population Sampling Theory." Draft Report, National Center for Health Statistics.

Schaible, W. L. (1978). "Choosing Weights for Composite Estimators for Small Area Statistics." Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 741-746.

Schaible, W. L., Brock, D. B. and Schnack, G. A. (1977a). "An Empirical Comparison of Two Estimators for Small Areas." Presented at the Second Annual Data Use Conference of the National Center for Health Statistics, Dallas, Texas.

Schaible, W. L., Brock, D. B., and Schnack, G. A. (1977b). "An Empirical Comparison of the Simple Inflation, Synthetic and Composite Estimators for Small Area Statistics." Proceedings of the American Statistical Association, Social Statistics Section, pp. 1017-1021.

Schaible, W. L., Brock, D. B., Casady, R. J. and Schnack, G. A. (1979). "Small Area Estimation: An Empirical Comparison of Conventional and Synthetic Estimators for States." Vital and Health Statistics, Series 2-82 (No. PHS-80-1356), National Center for Health Statistics.

U.S. Bureau of the Census (1972). County and City Data Book. Washington, U.S. Government Printing Office.

Wilder, M. (1973). "Current Estimates from the Health Interview Survey. 1970". Vital and Health Statistics, Series 10-72, (No. HRA-74-1054), National Center for Health Statistics.

Acknowledgements

The authors would like to thank Carmen Hines and Jacqueline Barclift of the Office of Statistical Research for computing estimates and providing graphical presentations for this paper. The authors also thank Kay Barrett and Cindy Decinti for typing the manuscript.

TABLE 1. Average Squared Errors of PSU Estimates for Selected Domains, Attribute Variables and Estimators: Health Interview Survey, 1969-71

DOMAIN AND ESTIMATOR	VARIABLE			
	High School	Unem-ployment	Manu-facturing	Sales & Clerical
21 LARGE SMSA's				
Simple Direct	6.7	1.6	2.7	2.6
Poststratified SD	5.6	1.8	11.1	18.8
Regional Mean	43.4	1.0	27.2	23.2
Synthetic	38.9	1.3	36.7	12.2
Regression	10.4	1.3	20.3	3.2
Regression-adj. Synthetic	25.1	1.2	25.2	6.3
SELF-REPRESENTING PSU's				
Simple Direct	21.1	3.5	8.3	10.6
Poststratified SD	18.8	3.5	20.7	26.8
Regional Mean	43.5	1.5	58.8	18.1
Synthetic	38.9	1.9	65.5	26.0
Regression	20.9	1.8	33.9	6.5
Regression-adj. Synthetic	27.1	2.0	33.1	8.3
ALL PSU's				
Simple Direct	32.2	5.8	17.7	12.6
Poststratified SD	31.3	7.3	29.9	24.5
Regional Mean	84.7	3.0	114.9	34.5
Synthetic	74.9	3.3	114.2	67.5
Regression	36.7	3.3	85.9	5.6
Regression-adj. Synthetic	39.5	3.0	87.0	7.8

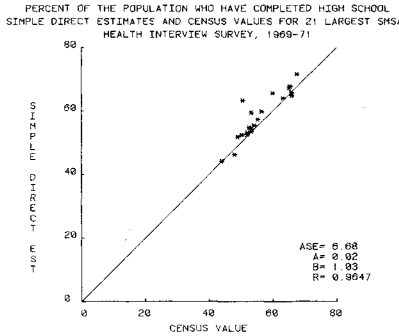
TABLE 2. Summary Statistics* for PSU Estimates for Selected Domains, Attribute Variables and Estimators: Health Interview Survey, 1969-71

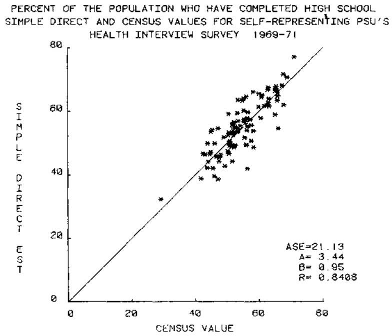
DOMAIN AND ESTIMATOR	VARIABLE															
	High School				Unemployment				Manufacturing				Sales and Clerical			
21 LARGE SMSA's	ASE	A	B	R	ASE	A	B	R	ASE	A	B	R	ASE	A	B	R
Simple Direct	6.68	0.02	1.03	0.96	1.61	2.02	0.54	0.56	2.71	0.72	0.99	0.98	2.59	-0.21	1.02	0.80
Regression	10.42	9.49	0.86	0.92	1.31	2.25	0.49	0.60	20.25	10.91	0.58	0.77	3.18	19.75	0.30	0.54
Composite	5.78	-0.81	1.04	0.97	1.21	2.19	0.50	0.62	6.29	2.28	0.91	0.93	1.71	12.10	0.58	0.80
SELF-REPRESENTING PSU's																
Simple Direct	21.13	3.44	0.95	0.84	3.49	2.04	0.63	0.47	8.30	2.17	0.91	0.96	10.62	-2.60	1.10	0.75
Regression	20.92	18.90	0.66	0.79	1.78	3.80	0.23	0.48	33.91	9.37	0.61	0.81	6.46	6.97	0.72	0.72
Composite	19.55	8.44	0.85	0.83	2.48	3.07	0.37	0.40	11.37	4.02	0.83	0.94	8.05	1.13	0.95	0.75
ALL PSU's																
Simple Direct	32.20	1.79	0.94	0.88	5.84	3.16	0.27	0.24	17.70	2.06	0.91	0.93	12.56	-0.72	1.02	0.83
Regression	36.72	13.61	0.70	0.83	3.27	4.35	0.02	0.09	85.88	16.00	0.32	0.58	5.61	2.71	0.86	0.89
Composite	41.25	5.15	0.86	0.86	6.34	3.08	0.24	0.23	19.76	2.05	0.91	0.93	12.56	-0.20	0.99	0.77

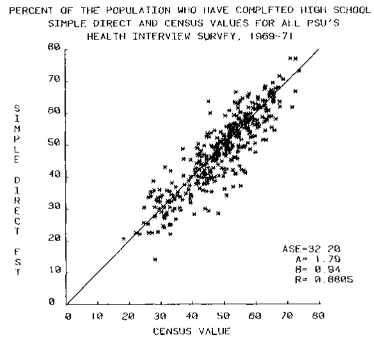
* The following notation applies to the statistics in this table:

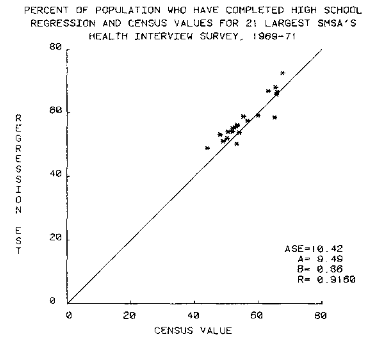
ASE - average squared error

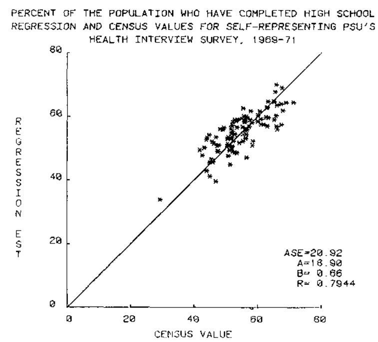
A - intercept of simple regression of estimate on census value
B - slope of simple regression of estimate on census value
R - correlation coefficient of estimate with census value

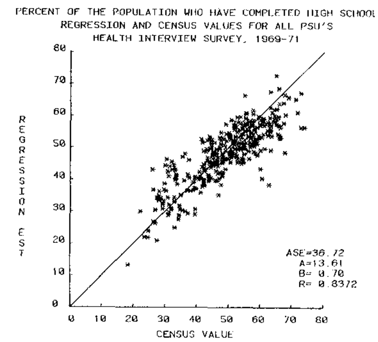


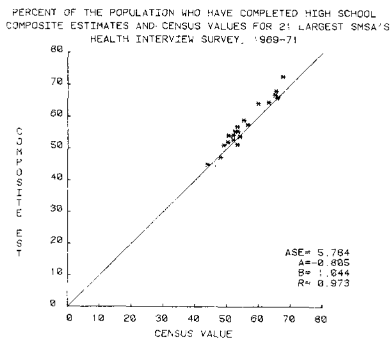


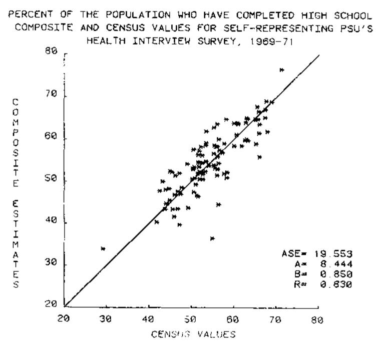


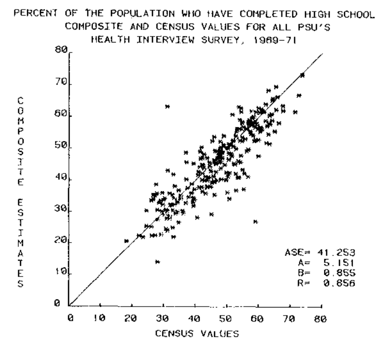












APPENDIX   Independent Variables Used in Constructing Regression Estimators

The following variables, available for each PSU from the County and City Data Book (1972), were used to estimate the regression coefficients in equation (5):

1. Population Density
2. Change in Population, 1960-1970
3. Percent Net Migration
4. Percent Urban
5. Percent of Persons Living in One-person Household
6. Percent of the Population White
7. Median Age in the PSU
8. Median Per Capita Income in the PSU
9. Percent of the Population Below Poverty Level
10. Per Capita Government Expenditure

The variables actually used for each domain/variable combination are given in the table below:

Domain	Dependent Variable	Independent Variable
21 Large SMSA's	High School	3, 5, 9
	Unemployment	4, 5, 9
	Manufacturing	3
	Sales & Clerical	3
Self-representing PSU's	High School	1, 5, 7, 8
	Unemployment	9, 10
	Manufacturing	2, 5, 7, 9
	Sales & Clerical	4, 6, 8
All PSU's	High School	1, 2, 3, 5, 7, 8, 9
	Unemployment	6, 10
	Manufacturing	1, 2, 3, 5, 8, 9, 10
	Sales & Clerical	3, 4, 6, 8