RECENT DEVELOPMENTS IN ESTIMATION FOR LOCAL AREAS

Eugene P. Ericksen, Institute for Survey Research, Temple University

1. Introduction

The regression-sample data method of postcensal estimation is a procedure by which one can combine sample survey data with symptomatic information to obtain local estimates of the criterion variable being measured by the survey data. This method has been tested extensively using population growth as the criterion (Ericksen, 1973a, 1973b), first, for the period beginning in 1960 and ending in 1964-67, and then more extensively for 1960 through 1970. The steps of the procedure in the latter test were as follows:

a. Sample estimates of population growth were obtained for the primary sampling units selected into the national sample of the Current Population Survey. These 1970 estimates of current population were divided by the corresponding 1960 Census populations giving sample estimates of 1960-70 population growth.

b. Symptomatic indicators, in this case 1970/1960 ratios of births, deaths, and school enrollment, were compiled for the sample psus and a multiple regression equation was computed using the sample estimates of population growth as the dependent variable. A second equation was then computed using the series of ratio-correlation estimates calculated at the Population Division of the Bureau of the Census as a fourth symptomatic indicator.

c. Values of the symptomatic indicators for counties were substituted into the regression equations and estimates were made of the 1960-70 population growth. This step was carried out for 2,586 counties in 42 states.

Corresponding estimates for these counties were made at the Population Division using four standard demographic techniques which have traditionally been used to estimate population growth. Of these techniques, the ratio-correlation technique was the most accurate. Little was gained from averaging estimates of two or more standard techniques. The regression estimates produced by our combination of sample data and symptomatic information were more accurate than those of any single or combination of standard techniques. This was particularly true when the series of ratio-correlation estimates were added as a fourth symptomatic indicator. There were moderate reductions in the mean error, but the greatest gain was in the reduction of the number of large errors, which was over 20 per cent. These results are presented in Table 1.

Some of the prominent features of our new method of postcensal estimation are the following:

a. Estimates of population growth have been shown to be more accurate. Part of the reason for this gain is that it is not necessary to make any assumptions concerning the nature of relationships beyond those of least squares linear regression. One of the difficulties of the ratio-correlation technique, for example, is the assumption of the continuance of past relationships.

b. Other series of estimates can be incorporated as symptomatic indicators. By including the series of ratio-correlation estimates as a symptomatic indicator, we have a way of correcting for the bias of ratio-correlation which arises from assuming the continuance of past relationships.

c. There is a procedure by which the mean squared error of the regression estimates can be calculated. Given this facility for measuring error, we can systematically test various combinations of symptomatic indicators to determine the composition of the optimal set. Because of the presence of the within-psu sampling error, this does not necessarily include all available symptomatic indicators.

2. Current Activity at the Bureau of the Census

A project is currently under way at the Bureau of the Census to compute yearly estimates of population growth since 1970 by our regression-sample data method. A determined effort is being made to obtain symptomatic data for all counties in each of the 50 states and the District of Columbia. It now appears that births, deaths, and school enrollment will be available, but with some time lag, for counties in all but a small handful of states. Additional data on automobile registrations will be available for some states and it is also expected that data on income tax exemptions will be available for all states (Zitter and Word, 1973). Substantial gains are anticipated from the use of tax records, even outside our regression-sample data format. In view of the changing relationships among variables, and the possibility that other symptomatic indicators will become available, the following instructions are pertinent to potential users of the regression-sample data technique:

a. Applications of and experimentation with the ratio-correlation technique have shown conclusively that relationships among a given set of variables can be expected to change over time. We have shown that the series of regression-sample data estimates were relatively accurate when computed over a given ten-year period for particular sets of three and four symptomatic variables. However, the accuracy of the regression-sample estimates relative to those of other techniques could change for a shorter estimation period beginning in 1970. It is also possible that the most accurate regression-sample data estimates would be computed with a different set of symptomatic indicators in this period. We can test this possibility by inspecting the mean squared error of the regression estimates and the correlations of the various indicators with the sample estimates of population growth.

b. In the absence of correlations between the sampling error and the value of the symptomatic indicators, the estimated regression equation using the sample estimates of population growth as the dependent variable is an unbiased estimate of the regression equation which would be obtained using Census tabulations of population growth if they were available. However, the presence of the within-psu sampling error will lower the observed values of the correlation coefficients. Low observed values of the correlation coefficients do not necessarily mean that the errors of the regression estimates will be large.

c. There are some unsolved problems regarding the inference from a sample of psus to a universe of counties. The mean squared error of the regression estimates refers to the accuracy of estimates for psus, when the units of interest may be counties. To the extent that counties are different from psus, reductions in the mean squared error for psus may not improve the accuracy of estimates for counties. A second unresolved problem has to do with specification errors arising from the distribution of the within psu sampling errors. If the size and direction of these errors vary systematically with values of the symptomatic indicators, the assumptions of linear regression may not be met. We have found this to be a minor problem in our application that resulted in larger errors for units with extreme growth rates, but the problem could be more important in other applications.

3. The Mean Squared Error

We have shown elsewhere (Ericksen, 1973a, 1973b) that the mean squared error of the regression sample data estimates can be expressed by the formula:

$$\frac{E(Y - \hat{Y})'(Y - \hat{Y})}{n} = \frac{(n - p - 1)\sigma_u^2}{n} + \frac{(p + 1)\sigma_v^2}{n} \qquad (3.1)$$

where σ_u^2 = the between-psu variance unexplained by the indicators,

σ_v^2 = the within-psu variance,

n = the number of psus in the sample, and

p = the number of symptomatic indicators.

When n is large relative to p, the mean squared error is determined by (1) σ_u^2, which decreases when new symptomatic indicators are added, and (2) the within-psu component of error which increases when indicators are added. If there were no within-psu component of error, optimal results would be obtained by maximizing p, i.e., by utilizing all available symptomatic information. We have found in our applications, however, that the presence of within-psu sampling variability has often meant that the optimal set of symptomatic indicators did not include all that were available.

In the test of 2,586 counties, there were seven symptomatic indicators available: births, deaths, school enrollment, and the four standard estimates. As shown in Table 2, where the ratios of the 1970 to the 1960 Decennial Census populations were the dependent variable, gains in the accuracy of regression estimates for psus were obtained by increasing the number of symptomatic indicators from four to seven. However, in the more realistic application, when the within-psu component of error was present, the increase from four to seven indicators actually brought about an increase in the errors. The mean error of the 2,586 county estimates increased from 4.2 per cent to 4.7 per cent. A similar result was obtained when six variables, with 51 observations (one for each state and the District of Columbia) were available.

The fact that the optimal set of indicators included four variables was due to the nature of the structural relationships and the size of the within-psu variance. We have evidence that these change over time, as shown in Table 3. In particular, for the Current Population Survey sample, the within-psu variance increased. This is because the CPS sample was based on the 1960 Census, and that patterns of subsequent growth were uneven, leading to variation in the size of sample segments within psus. This trend leads us to expect that more symptomatic indicators should be used in shorter time periods. On the other hand, the relationships among the variables appear to become stronger as time passes. In spite of the increasing within-psu variability which dampens the observed correlations, these observed correlations grew larger from 1963 through 1967. In shorter periods, changes in population size, as well as in the symptomatic indicators, appear to be due more to random fluctuations. As time passes, changes in the variables are larger, and the relationships among these changes more systematic. This leads to the contrary expectation that the optimal set of indicators would be smaller for a shorter time period. To determine the optimal set of indicators, we must estimate the mean squared error in each estimating situation.

Because the true values of the criterion variable are unobserved, the mean squared error cannot be estimated directly. To obtain equation (3.1), we must first compute the mean of the squared differences between the regression estimates and the sample estimates for the sample psus and then subtract an allowance for the within-psu sampling error. The mean squared difference between the regression and sample estimates can be expressed by the formula:

$$\frac{E(Y_o - Y)'(Y_o - Y)}{n} = \frac{(n - p - 1)\ (\sigma_u^2 + \sigma_v^2)}{n} \qquad (3.2)$$

To obtain (3.1) we need to subtract the term $(n - 2p - 2)\sigma_v^2/n$. In order to obtain a good estimate of the mean squared error, we clearly need to have a good estimate of σ_v^2.

When we reported earlier results (Ericksen, 1973b), we did not feel that a good estimate of σ_v^2 was available. We had computed half-samples defined by the eight rotation groups of the CPS (U.S. Bureau of the Census, 1963) and had overestimated the mean error of the sample estimates for psus. This is because sample segments within the CPS sample had not been placed equally into rotation groups within individual psus. However, when the half-samples were formed on the basis of sample segments without regard to rotation group, a better estimate was obtained. The derivation of equation (3.2) depends on the values of (1) the sampling error and (2) the structural errors of regression, along with the sampling errors being unrelated to the symptomatic indicators. Our technique for estimating the mean squared error is particularly sensitive to these specification errors, as the following illustration shows.

The practical question we faced in the 1970 test was whether or not improvements in accuracy over that given by the ratio-correlation technique would be obtained by adding births, deaths, and school enrollment as symptomatic indicators in a regression equation. We found that the ratio-correlation estimates accounted for 92.7 per cent of the variance of the actual 1970/1960 ratios of population of the sample psus. Adding the three symptomatic indicators, the coefficient of determination, R^2, was increased to .951, a clear increase in the explained and reduction in the unexplained variance. However, the increase in the explained variance of the sample estimates of 1960-70 population growth obtained by adding the three symptomatic indicators to ratio-correlation was much smaller, from 41.7 per cent to 42.8 per cent. This was due to the presence of the within-psu error which is not reduced by adding symptomatic information. The observed variance of the distribution of sample estimates before regression was .0438. Using the series of ratio-correlation estimates as a single symptomatic indicator, the mean squared difference of the regression and sample estimates as expressed by equation (3.2) was .0255. This was reduced to .0250 when the number of symptomatic indicators was increased from one to four. Our estimate of the within-psu variance is σ_v^2 = .0253. Subtracting the allowance for this component of error, our final estimates of the mean squared error are .0004 where the ratio-correlation estimate is a single indicator and .0001 with four indicators. This is a very small difference considering the size of the within-psu variance and the mean squared difference between the regression and sample estimates. A small fluctuation could have seriously altered the observed results. When the number of symptomatic indicators was increased to seven, the coefficient of determination was R^2 = .432, the mean of the squared differences was .0249, and the final estimate of the mean squared error, .0002. These differences are so small that one may be on safer, although less scientific, grounds simply to observe that the increase in R^2 from .417 to .428 is large enough to produce a good reduction in error while guessing that the further increase to R^2 = .432 is not, given the increase in the number of symptomatic indicators.

Some of the difficulties in estimating σ_v^2 arise from the fact that the within-psu sampling error is positively correlated to the growth rate, and hence to the values of the symptomatic indicators. The correlation between the actual error of the CPS estimate and the estimated within-psu variance is +.45. This affects the estimate both of σ_v^2 and the way we obtain an estimate of equation (3.1) from equation (3.2). A second source of error is the correlation between the within-psu error and the growth rate, which is + .06. This introduces curvilinearity, since the sample estimates of the fastest growing areas tend to be too large and those of the slowest growing too small, thus biasing the estimation of regression coefficients. One result of this was that estimates of areas with extreme growth rates had larger errors. This particular problem is covered in the literature on econometrics where the usual solution is to apply a transformation. We have attempted several such solutions, but have yet to find a transformation which allows us to reduce the errors of the extreme cases without increasing the errors of the majority of cases which have moderate values.

One obvious procedure for reducing the mean squared error of the regression estimates is to reduce the within-psu variance. This could be done by improving the within-psu sample design, or, as we will attempt to do, by introducing more sample data. In our program at the Census Bureau, we plan to eventually request tabulations from other government surveys such as HIS and NCS. This will reduce σ_v^2 where the psus in the various surveys are the same and reduce the ratio $(p + 1)/n$ and therefore the within-psu component of error in equation (3.1) in cases where the psus are different.

This necessarily reduces the errors of the primary sampling units, but the effects on county estimates are uncertain. To illustrate this point, when the regression equation with three symptomatic indicators, births, deaths, and school enrollment, was recomputed using the 1970/1960 Census population ratios as the dependent variable, i.e., setting σ_v^2 equal to zero, the mean error of the psu estimates was 2.8 per cent. This compares to the mean error of 3.2 per cent when the CPS estimates were the dependent variable. The difference between 2.8 and

3.2 per cent was due to the within-psu error. However, when the two equations were used to make county estimates, the mean error was 4.4 per cent in both cases. The Census ratio equation, computed without the within-psu error, had done a better job of making psu estimates, but the transition from psus to counties had become more difficult. When the distribution of errors was broken down by size of the 1970 county population, it was found that use of the Decennial Census ratios in place of the CPS estimates had reduced the mean error for all categories of counties with population greater than 25,000, but that the mean error had increased among counties smaller than 25,000. Counties in this last category were the majority of all counties but were least similar to the CPS sample psus which usually consisted of combinations of counties picked with probabilities proportional to the size of the total population.

4. New Strategies and Plans

Given the limited gains obtained from reducing the within-psu component of error, and our lack of success in finding suitable transformations to reduce errors, the most promising approach to reducing our errors appears to be the introduction of new symptomatic information. One variable which has been shown to reduce errors is automobile registrations. Data were available in the 1970 test for 2,223 counties in 32 states. A five-variable regression equation, also including births, deaths, school enrollment, and the ratio-correlation estimate was computed and county estimates made. The mean error of these estimates was 3.8 per cent and 122 errors were greater than 10 per cent. The corresponding figures for this set of counties for the four variable regression equation omitting automobile registrations were 4.1 per cent with 148 large errors and, for the standard series of ratio-correlation estimates, 4.5 per cent with 220 large errors.

Another promising, but as yet untested, variable is the number of exemptions on income tax returns. Changes in address of persons listed on income tax forms are to be used to estimate net migration and when added to recorded natural increase, could give extremely accurate estimates of population growth. It is quite possible that these estimates would be sufficiently accurate in themselves so that little gain would be obtained by computing regression-sample data estimates. But it is more likely that some bias will be introduced because of the characteristics of persons not listed on tax forms or whose likelihood of being listed on a form varies at point of origin and destination. In such a case, this bias could be corrected by using the tax estimate as a symptomatic indicator in a regression equation possibly including other symptomatic indicators with sample data as the dependent variable.

Finally, we have made plans to attempt to estimate other variables such as racial composition, unemployment, and median family income. Although births and deaths are available by race in many counties, the chief barrier faced here is the lack of symptomatic information. Data on wages and work force appear to be available in metropolitan areas, but we are still searching for symptomatic data available on a national basis. If such data can be found, we can combine our symptomatic information and sample data with estimates which can be generated by other means. One such series would be the synthetic estimates being discussed in this session.

BIBLIOGRAPHY

Ericksen, Eugene P. "A Method for Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas," Demography, 10, 137-160, 1973a.

Ericksen, Eugene P. "A Regression Method for Estimating Population Changes of Local Areas," 1973b, manuscript submitted for publication.

U.S. Bureau of the Census, Technical Report No. 7, "The Current Population Survey--A Report on Methodology," 1963. U.S. Government Printing Office, Washington, D.C.

Zitter, Meyer, and David Word. "Use of Administrative Records for Small-Area Population Estimates," 1973, presented at the Annual Meeting of the Population Association of America.

ACKNOWLEDGEMENT

The research upon which this paper is based was carried out in cooperation with the Bureau of the Census, in particular with Benjamin J. Tepping of the Research Center for Measurement Methods and with the Population Division, who supplied all the symptomatic information used in computing the regression equations as well as special tabulations concerning the distributions of errors of the ratio-correlation and other standard estimates. Special thanks are due to Lori Kessler and Harris Miller without whose assistance at Temple University this research could not have been carried out. The findings, recommendations, and conclusions in this paper are the sole responsibility of the author and are not necessarily endorsed by the U.S. Government. The data in this paper are the result of tax-supported research and, as such, are not copyrightable. The data may be freely reprinted with the customary crediting of the source.

Table 1: Relative Accuracy of Standard and Regression-Sample Data Estimates of Population Growth, 1960 to 1970, for 2,586 Counties in the United States

Procedure	Mean Error[1]	Number of Counties With Error 10 Per Cent or Greater
Vital Rates	7.4	673
Component Method II	7.2	645
Composite	5.9	407
Ratio-Correlation	4.6	264
Component Method II, Composite, Ratio-Correlation, Averaged[2]	4.7	249
Regression-Sample Data, 3 Symptomatic Indicators[3]	4.4	220
Regression-Sample Data, 4 Symptomatic Indicators[3]	4.2	194

[1] All estimates were multiplied by an appropriate constant in order to sum to a separately estimated 42-state total.

[2] This was the most accurate combination of the four standard techniques.

[3] The 3-variable equation used was $\hat{Y} = .158 + .218X_1 + .142X_2 + .520X_3$.

The 4-variable equation used was $\hat{Y} = .058 - .097X_1 + .045X_2 + .214X_3 + .745X_4$.

Source of Standard Estimates and Estimate of 42-State Total: U.S. Bureau of the Census, Current Population Reports, Series P-26, No. 21, "Federal-State Cooperative Program for Local Population Estimates: Test Results - April 1, 1970," Washington, D.C.: Government Printing Office, 1973.

Table 2: Mean Errors Obtained With Various Sets of Symptomatic Indicators

Units of Estimates are 444 Primary Sampling Units.

Number of Symptomatic Indicators[2]	Mean Percentage Error, Psus[1]	
	Dependent Variable 1970 Census/1960 Census	Dependent Variable 1970 CPS/1960 Census
3	2.83	3.20
4	2.60	2.92
7	2.11	3.24

Units of Estimates are 50 States and District of Columbia.

Number of Symptomatic Indicators[3]	Mean Percentage Error, States[1]	
	Dependent Variable 1970 Census/1960 Census	Dependent Variable 1970 CPS/1960 Census
3	1.22	1.64
4	1.08	2.16
6	1.07	3.91

[1] Mean percentage error, comparing regression estimate with Census tabulation. "Dependent variable" is that used to compute regression equation.

[2] Set of three indicators included births, deaths, and school enrollment. The fourth indicator was the ratio-correlation estimate,and indicators five through seven were the composite, component method II, and vital rates estimates.

[3] Set of three indicators included births, school enrollment, and work force. The fourth indicator was deaths,and indicators five and six were automobile registrations and income tax returns.

Table 3: Values of Estimated Within-Psu Variance of Population Growth and Coefficients of Determination, 1963 through 1967

Year	Within-Psu Variance[1]	Coefficient of Determination (R^2)[2]
1963	.0253	.016
1964	.0378	.021
1965	.0383	.085
1966	.0458	.117
1967	.0473	.264

[1] Computed as squared difference between random half-samples defined by rotation group.

[2] Three symptomatic indicators were births, deaths, and school enrollment in each case.