THE RELATIVE ACCURACY OF DIRECT AND INDIRECT ESTIMATES OF STATE POVERTY RATES

Allen L. Schirm, Mathematica Policy Research, Inc.
Mathematica Policy Research, Inc., 600 Maryland Avenue, S.W., Suite 550, Washington, DC 20024-2512

Key Words: Small Area, Simulation, Shrinkage

1. INTRODUCTION

Policymakers have become increasingly sensitive in recent years to differences in socioeconomic conditions among regions, states, and localities. They have questioned whether the benefits of our social welfare system are shared equitably, and their concerns have intensified the need for subnational estimates for indicators of well-being and program effectiveness. Such estimates have been used to allocate program funds and to improve program effectiveness by targeting additional resources for expanding participation in those areas where participation falls far short of need.

Although accurate estimates are vital to program success, very little is known about the relative accuracy of alternative estimators used to derive "small area" estimates (U.S. Office of Management and Budget 1993). The leading estimators developed for small area estimation can be classified as direct or indirect. To obtain an estimate for a particular area and a particular time period, a direct estimator uses only data for that area and time period. An indirect estimator "borrows strength," using data from other areas or time periods. In this paper, we assess the relative accuracy of the direct sample estimator and three indirect estimators: (1) a pooled sample estimator, (2) a regression estimator, and (3) a shrinkage estimator that combines direct sample and regression estimates.

In Section 2, we describe the simulation methods used to evaluate the relative accuracy of direct and indirect estimates of state poverty rates. The rationale for our simulation approach and its potential limitations are discussed in Schirm (1993). In Sections 3 and 4, we present and summarize our simulation results. Our principal finding is that shrinkage estimates are substantially more accurate than direct sample, pooled sample, or regression estimates. In the remainder of this section, we discuss direct and indirect estimators.

Aside from its simplicity, the principal advantage of the direct sample estimator is that it is unbiased. Its main disadvantage and the main motivation for considering indirect estimators is that there is often substantial sampling variability in direct estimates for small areas. The Census Bureau publishes Current Population Survey (CPS) sample estimates of state poverty rates with the warning that they "should be used with caution since [they have] relatively large standard errors" (U.S. Bureau of the Census 1993).

An indirect estimator proposed to address this problem of imprecision is the pooled sample estimator. Pooling combines sample data from different time periods. Haveman, Danziger, and Plotnick (1991) derived state poverty rate estimates by combining CPS samples for three consecutive years and dropping overlapping observations from the first and third years. This approach approximately doubles sample sizes and, therefore, reduces standard errors by nearly 30 percent based on conventionally used calculations. The drawback is that a pooled estimator is biased, since a state's pooled poverty rate for a single year is a weighted average of its poverty rates for three years.

An alternative to a time-indirect estimator like the pooled sample estimator is a domain-indirect estimator, which uses data from different domains (areas) rather than different time periods. The regression estimator is domain indirect and commonly used. Developed by Ericksen (1974), it combines sample data with symptomatic information, using multivariate regression to "smooth" direct sample estimates, that is, reduce their sampling variability. The basic regression model for estimating state poverty rates is:

$$Y_s = XB + U,$$

where Y_s is a vector of direct sample estimates, X is a matrix containing data for each state on a set of "symptomatic indicators" typically obtained from census or administrative records data with little or no sampling variability, B is a vector of unknown coefficients, and U is an error term. The regression estimator is:

$$Y_r = X\hat{B},$$

where $\hat{B}$ is the least squares estimate of B. Thus, the regression estimates of state poverty rates are the predicted values from the regression model. In the regression, the state observations are often weighted by a measure of the reliability of the direct sample estimates. Because of regression toward the mean, the regression estimator is biased.

Except in estimating the regression coefficients, the regression estimator makes no use of the direct sample estimates. Likewise, the direct sample estimator ignores the systematic relationships among

state poverty rates. In contrast to these estimators, shrinkage estimators seek to use all available information or, at least, the information that is most relevant and practical to use.

Also known as compromise or composite estimators, shrinkage estimators calculate optimally weighted averages of estimates obtained using other methods, such as direct sample and regression estimates. A shrinkage estimator draws on the relative strengths of the alternative estimates to obtain a better estimate. The strength of the direct sample estimate is unbiasedness, and the strength of the model estimate is low sampling variability. A shrinkage estimator optimally combines alternative estimates to minimize an overall measure of error, like expected mean squared error, that reflects both bias and sampling variability.

The simplest form of a shrinkage estimator is:

$$Y_c = aY_1 + (1 - a)Y_2,$$

where Y_c is the shrinkage estimator that combines the alternative estimators Y_1 and Y_2, a is the vector of weights on the elements of Y_1, $(1 - a)$ is the vector of weights on the elements of Y_2, and $0 \leq a \leq 1$. To optimally combine alternative estimates, a shrinkage estimator weights the estimates according to their relative reliability. Thus, all else equal, a shrinkage estimator would place a large weight on the direct sample estimate for a large state and a small weight on the direct sample estimate for a small state. Fay and Herriott (1979) developed a shrinkage estimator that combined sample and regression estimates of per capita income for small places (population less than 1,000) receiving funds under the General Revenue Sharing Program.

2. THE SIMULATION PROCEDURE

In this section, we describe our simulation procedure. The procedure has four basic steps: (1) specify a population, (2) draw multiple samples from the population, (3) calculate direct and indirect estimates, and (4) compare the relative accuracy of the alternative estimates. After discussing these four steps, we describe the additions to each step required to obtained pooled sample estimates.

2.1 Step 1: Specify a Population

We use the March 1990 CPS sample as the population, ignoring the weights on observations. This gives a total population size of approximately 158,000 individuals and state populations ranging from under 1,300 to over 14,000 across the 51 states (the 50 states and DC). Except for the poverty income thresholds used, we specify the poverty status of each individual in the population using the same definition employed by the Census Bureau in deriving poverty estimates from the CPS. We use the simplified guidelines based on family size and state of residence that are used for determining eligibility for several federal programs as the poverty guidelines for our simulations, averaging guidelines for the first and last six months of 1989 to obtain calendar year 1989 guidelines.

2.2 Step 2: Draw Multiple Samples

In the second step of our simulation procedure, we draw multiple samples from the population. The purpose in drawing multiple samples is to determine how sampling variability contributes to the inaccuracy of direct and indirect estimates.

Step 2a: Calculate the Sample Size for State i, i = 1, 2, ..., 51. Replicating the complex CPS sample design in our simulations is well beyond the scope of this study. Nevertheless, we draw samples to ensure that the standard errors of the direct sample estimates in our simulations will generally equal or be very close to the standard errors--calculated to reflect the complex CPS sample design--for weighted CPS poverty rate estimates.

To simplify the simulation procedure, we use stratified simple random sampling, stratifying only by state. Within strata, we sample without replacement. Our expression for calculating the sample size for state i, which is derived in Schirm (1993), is:

$$n_i = \frac{T_i [s_i^2 + p_i (1 - p_i)]}{T_i s_i^2 + p_i (1 - p_i)}.$$

For the simulations, we set s_i equal to the standard error of the weighted CPS poverty rate estimate for state i. T_i is the population size, and p_i is the poverty rate (expressed as a proportion) in the population specified in Step 1. This p_i is the "true" poverty rate for state i in our simulations. It is easy to show that the estimated standard error for a direct sample estimate for state i in our simulations will generally equal or be very close to s_i. State sample sizes range from about 220 to over 2,200.

Step 2b: Draw, Without Replacement, a Simple Random Sample of Size n_i for State i, i = 1, 2, ..., 51. Individuals in the population are stratified by state, and independent samples are drawn in each state. The 51 state samples constitute a single national sample (henceforth, a "sample").

Step 2c: Draw 1,000 Samples. We repeat Step 2b 1,000 times, drawing 1,000 independent samples. Each of the 1,000 repetitions of our simulation procedure beginning with the drawing of a sample (Step 2b) and ending with the calculation of direct and indirect estimates (Step 3) is an "iteration."

2.3 Step 3: Calculate Direct and Indirect Estimates

Step 3a: Calculate Direct Sample Estimates. For state i, the direct sample estimate of the proportion poor is the number of individuals in the sample who are poor divided by the sample size, n_i. We calculate standard errors using the well-known formula for the standard error of a proportion estimated from a simple random sample drawn without replacement.

Step 3b: Select the Best-Fitting Regression Model. Our regression model regresses the 51 direct sample estimates of state poverty rates on symptomatic indicators measuring state characteristics that are likely associated with interstate differences in poverty rates. We need to specify the symptomatic indicators included in the "best-fitting" regression model in a particular iteration and seek a model accounting for much of the interstate variation in poverty rates with a small number of symptomatic indicators.

We allow for up to five symptomatic indicators: (1) the proportion of the state population receiving Supplemental Security Income, (2) state per capita total personal income, (3) the state crime rate, (4) a dummy variable equal to one for the New England states, and (5) a dummy variable equal to one if at least 1 percent of the state's total personal income is derived from the oil and gas extraction industry.[1] Our model-fitting procedure selects the model that maximizes:

$$\bar{R}^2 = 1 - \left[\frac{51 - 1}{51 - k - 1}\right](1 - R^2),$$

where k is the number of symptomatic indicators in the regression model (ranging from one to five), and R^2 is the usual coefficient of multiple determination. Whereas adding a symptomatic indicator always increases R^2, $\bar{R}^2$ will decrease if the improvement in fit, as measured by R^2, is small. We repeat our model-fitting procedure for each iteration.[2]

Step 3c: Calculate Shrinkage Estimates. We use an Empirical Bayes shrinkage estimator. This estimator was used by Ericksen and Kadane (1985) to estimate population undercounts in the 1980 census for 66 areas covering the entire U.S. and by Schirm, Swearingen, and Hendricks (1992) to estimate state poverty rates and Food Stamp Program participation rates. It was originally developed by DuMouchel and Harris (1983).

Our shrinkage estimator is:

$$Y_c = \left[D + \frac{1}{u^2}P\right]^{-1} DY_s,$$

where Y_c is a (51 × 1) vector of shrinkage estimates, and Y_s is a (51 × 1) vector of direct sample estimates. D is a (51 × 51) diagonal matrix with diagonal element (i,i) equal to one divided by the variance (standard error squared) of the direct sample estimate for state i. $P = I - X(X'X)^{-1}X'$, where I is a (51 × 51) identity matrix and X is a (51 × K) matrix containing data for each state on a set of $k = K - 1$ symptomatic indicators. (The other column of X consists of all ones and allows for an intercept in the regression model.) u^2, a scalar reflecting the lack of fit of the regression model, is estimated by maximizing the likelihood function:

$$L = |W|^{1/2}\,|X'WX|^{-1/2}\exp\left[-\frac{1}{2}Y_s'SY_s\right],$$

where $W = (D^{-1} + u^2I)^{-1}$ and $S = W - WX(X'WX)^{-1}X'W$. The variance-covariance matrix of our shrinkage estimator is:

$$V_c = \left[D + \frac{1}{u^2}P\right]^{-1}.$$

2.4 Step 4: Compare the Relative Accuracy of Direct and Indirect Estimates

In this paper and in Schirm (1993), we compare the relative accuracy of the alternative estimates according to a wide variety of accuracy criteria, including root mean squared errors (RMSEs) and mean absolute errors (MAEs). For all assessments of accuracy, the true poverty rates remain the same across iterations and are the poverty rates in the population specified in Step 1.

2.5 Pooled Sample Estimation

To obtain pooled sample estimates, we must add to the first three steps of our simulation procedure. In Step 1, we must define "populations" from which to draw samples. To simulate the most often used procedure of pooling three consecutive annual samples, we use the nonoverlapping observations from the March 1989 and March 1991 CPS samples. In Step 2, we draw a sample of $n_i/2$ individuals from the March 1989 CPS observations and a sample of $n_i/2$ individuals from the March 1991 CPS observations for state i. These n_i additional individuals are pooled with the n_i individuals selected from the March 1990 CPS to double the sample size. In Step 3, the pooled sample estimate of the proportion poor is the number of poor individuals in the pooled sample divided by the sample size, $2n_i$. We estimate the standard error for the pooled sample estimate by multiplying the standard error for the direct sample estimate by $\sqrt{0.5}$.

3. SIMULATION RESULTS

In this section, we compare the accuracy of point estimates from the direct sample, pooled sample, regression, and shrinkage estimators. Then, we compare how well the four estimators estimate a key feature of the distribution of state poverty rates. Finally, we compare how well the four estimators estimate error, that is, how well estimated standard errors and confidence intervals reflect the uncertainty in the poverty rate estimates.

3.1 Accuracy of Point Estimates

Altogether, from each of the four estimators, we obtain 51,000 estimates--51 state estimates for each of the 1,000 iterations. It is not meaningful to compare the errors in the four estimates for a single state in a single iteration. The estimates and, hence, the estimation errors may be unusual due to unusually large or small sampling errors. To control for the influence of sampling variability and discover what errors are typical, we need to aggregate estimation errors. Schirm (1993) takes three approaches to aggregating errors: (1) aggregating errors across iterations for each state, (2) aggregating errors across states for each iteration, and (3) aggregating errors across all iterations and states. Here, we report results from mainly the second approach since all three imply the same conclusions.

In Table 1, we compare the accuracy of the direct sample, pooled sample, and regression estimators relative to the shrinkage estimator on the basis of RMSEs calculated for each iteration. In the direct sample estimation column, iterations for which the shrinkage estimator has a lower RMSE than the direct sample estimator are counted in the top panel of the table, while iterations for which the shrinkage estimator has a higher RMSE are counted in the bottom panel. Thus, shrinkage increases accuracy for iterations in the top panel and decreases accuracy for iterations in the bottom panel. In both panels, we display the distribution of iterations according to the percent change in the RMSE due to shrinkage. The relative accuracy of the shrinkage estimator falls as we move down in each column.

According to Table 1, shrinkage increases accuracy (reduces the RMSE) 97 percent of the time and decreases accuracy only 3 percent of the time relative to the direct sample estimator. The median reduction in the RMSE is 21 percent. For only 1 percent of iterations does shrinkage increase the RMSE by more than 10 percent. Compared with the pooled sample and regression estimators, shrinkage increases accuracy 90 and 100 percent of the time, respectively. The median reductions in the RMSE are 17 and 34 percent. Thus, relative to the other three estimators, the shrinkage estimator usually increases accuracy substantially. It rarely decreases accuracy and almost never decreases accuracy by much.[3,4,5]

3.2 Distributional Accuracy

Although Schirm (1993) considers several criteria measuring how accurately the alternative estimates represent characteristics of the distribution of state poverty rates, we explore here only whether the estimators rank states accurately in a tail of the poverty rate distribution. We could imagine a federal program providing states with the highest poverty rates some kind of economic assistance. How well do the alternative estimators identify, say, the "top 10" states--the 10 states with the highest poverty rates?

In Table 2, we find that the shrinkage estimator is substantially more likely to identify 9 or 10 of the top 10 states than are the other estimators. In about three-quarters of the iterations, the shrinkage estimator correctly identifies at least 9 of the 10 states with the highest poverty rates. The direct and pooled sample estimators attain that standard less than half the time (in 40 and 47 percent of the iterations, respectively). Although the regression estimator correctly identifies all of the top 10 states in just one iteration, it does get 9 right well over half the time.

3.3 Accuracy in Estimating Error

It is usual statistical practice to provide some expression of the uncertainty associated with point estimates. In this section, we assess the accuracy of estimated standard errors and confidence intervals as expressions of our uncertainty and the error in point estimates. We ask, specifically, whether 95-percent confidence intervals provide 95 percent coverage.

According to Table 3, coverage is very close to 95 percent for the direct sample and shrinkage estimators. For both estimators, over 93 percent of the 51,000 confidence intervals--one for each of the 51 states in each of the 1,000 iterations--contain the true poverty rate.[6] However, for the pooled sample estimator, coverage is below 85 percent, falling substantially short of the nominal (95 percent) level. Coverage for the regression estimator is only a little over 50 percent.

4. SUMMARY

According to the several alternative measures of accuracy considered here or in Schirm (1993), we find that shrinkage estimates are substantially more accurate than direct sample, pooled sample, or

regression estimates. For example, calculating RMSEs and MAEs for each iteration of our simulation procedure, we find that there is at least a 90 percent chance that shrinkage will improve accuracy. The median reductions in the RMSE or MAE are large--about 15 to 20 percent relative to the direct and pooled sample estimators and over 30 percent relative to the regression estimator. Shrinkage rarely decreases accuracy, and even when it does, the loss in accuracy is usually small.

In evaluating the accuracy of estimated standard errors and confidence intervals as expressions of our uncertainty, we find that for the pooled sample and regression estimators, standard errors and confidence intervals are misleading. The standard errors are too small, and the confidence intervals are too narrow, underestimating our uncertainty and giving a false sense of accuracy. In contrast, standard errors and confidence intervals for the direct sample and shrinkage estimators generally reflect accurately the uncertainty in estimated poverty rates.

NOTES

[1]Schirm, Swearingen, and Hendricks (1992) examined these and other symptomatic indicators. Data sources are listed in Schirm (1993).

[2]The regression estimator examined here weights state observations based on the precision of the direct sample estimates and is:

$$Y_r = X(X'DX)^{-1}X'DY_s.$$

To replicate a commonly used approach, we calculate the variance-covariance matrix of the regression estimator according to:

$$V_r = \left[\frac{(Y_s - Y_r)'D(Y_s - Y_r)}{51 - K}\right] X(X'DX)^{-1}X'.$$

X, D, Y_s, and K are defined in Step 3c.

[3]Results for MAEs differ little from the results for RMSEs. Compared with the direct sample, pooled sample, and regression estimators, shrinkage increases accuracy 97, 90, and 100 percent of the time according to MAEs, and the median reductions in the MAE are 20, 17, and 36 percent, respectively.

[4]Because states are different sizes, aggregating errors across states raises the issue of whether to differentially weight state errors. Schirm (1993) finds that estimates of relative accuracy are not sensitive to the weighting scheme used.

[5]Based on RMSEs for states (rather than iterations), the shrinkage estimator increases accuracy for 43, 33, and 33 states compared with the direct sample, pooled sample, and regression estimators, respectively. In the median state, shrinkage reduces the RMSE by 20, 14, and 27 percent. Aggregating errors across both iterations and states, we find that shrinkage reduces RMSEs (and MAEs) by 15 to 20 percent compared with the direct and pooled sample estimators and by 30 to 45 percent compared with the regression estimator.

[6]Shrinkage confidence intervals, however, are substantially narrower than direct sample confidence intervals, implying less uncertainty. The median reduction in width from shrinkage is 22 percent.

ACKNOWLEDGMENTS

I thank John Czajka and Bruce Klein for helpful comments. Cara Olsen provided exceptionally skillful assistance in conducting the simulations.

REFERENCES

DuMouchel, W., and Harris, J. (1983), "Bayes Methods for Combining the Results of Cancer Studies in Humans and Other Species," *Journal of the American Statistical Association*, 78, 293-315.

Ericksen, E. (1974), "A Regression Method for Estimating Population Changes of Local Areas," *Journal of the American Statistical Association*, 69, 867-875.

Ericksen, E., and Kadane, J. (1985), "Estimating the Population in a Census Year: 1980 and Beyond," *Journal of the American Statistical Association*, 80, 98-131.

Fay, R., and Herriott, R. (1979), "Estimates of Incomes for Small-Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269-277.

Haveman, J., Danziger, S., and Plotnick, R. (1991), "State Poverty Rates for Whites, Blacks, and Hispanics in the late 1980s," *FOCUS*, 13, 1-7.

Schirm, A. (1993), *The Relative Accuracy of Sample and Shrinkage Estimates of State Poverty Rates*, Washington, DC: Mathematica Policy Research.

Schirm, A., Swearingen, G., and Hendricks, C. (1992), *Development and Evaluation of Alternative State Estimates of Poverty, Food Stamp Program Eligibility, and Food Stamp Program Participation*, Washington, DC: Mathematica Policy Research.

U.S. Bureau of the Census (1993), *Poverty in the United States: 1992*, Current Population Reports, Ser. P60-185.

U.S. Office of Management and Budget (1993), *Indirect Estimators in Federal Programs*, Statistical Policy Working Paper No. 21.

Table 1. Percentage of Iterations for which the Shrinkage Estimator Has Lower Root Mean Squared Error than the Direct Sample, Pooled Sample, and Regression Estimators

Effect of Shrinkage	Percentage of Iterations: Direct Sample	Pooled Sample	Regression
Shrinkage Increases Accuracy (Lowers RMSE)			
Percent Decrease in RMSE:			
> 30	10	8	71
20 - 30	43	28	28
10 - 20	35	36	1
0 - 10	10	18	0
Shrinkage Decreases Accuracy (Raises RMSE)			
Percent Increase in RMSE:			
0 - 10	2	8	0
> 10	1	2	0

Table 2. Accuracy in Identifying the Ten States with the Highest Poverty Rates

Number of Top Ten States Correctly Identified	Percentage of Iterations: Direct Sample	Pooled Sample	Regression	Shrinkage
6	2	0	0	0
7	14	7	3	1
8	44	46	35	24
9	36	46	62	58
10	4	1	0	16

Table 3. 95-Percent Confidence Interval Coverage

Coverage Criterion	Direct Sample	Pooled Sample	Regression	Shrinkage
Percentage of All 95-Percent Confidence Intervals Including the True Value	94.4	84.3	52.8	93.2