METHODS USED FOR SMALL AREA POVERTY AND INCOME ESTIMATION

Robin Fisher, US Bureau of the Census
US Bureau of the Census, Washington, DC 20233

Key Words: Poverty, Small Areas, CPS

Abstract

Existing postcensal estimates of poverty and income at the county level are considered inadequate for various reasons: The Census is rapidly dated and the March CPS is not sufficiently reliable, especially for those counties which are not sampled by CPS. The goal of The Small Area Income and Poverty Estimates (SAIPE) project is to form these estimates. We modeled the number of poor in various age categories and median household income as a function of various variables taken from administrative records. We recognize two sources of "error" -- sampling error and model error -- and apply a shrinkage estimator to obtain estimates of number of poor or median income by county. Finally, a ratio adjustment is used to make estimates consistent with the SAIPE state estimates. We describe the methods used to obtain these estimates and their standard errors and present some empirical evaluations of the models.

0. Introduction

Existing postcensal estimates of poverty and income at the county level are considered inadequate for various reasons, including the fact that the Census is rapidly dated and the March CPS is not sufficiently reliable, especially for those counties which are not sampled by CPS. The goal of The Small Area Income and Poverty Estimates (SAIPE) project is to form these estimates. We modeled the number of poor in various age categories and median household income as a function of various variables taken from administrative records. Of particular interest to this paper are number of poor, especially those aged 5 to 17 years. We recognize two sources of "error", sampling error and model error, and apply a shrinkage estimator to obtain estimates of number of poor or income by county. To form the necessary estimates of the variance components, we use a Best Linear Unbiased predictor, estimated with a modification of the MINQUE(0) estimator. Finally, a ratio adjustment is used to make estimates consistent with the SAIPE state estimates, derived independently.

We describe the methods used to obtain these estimates and their standard errors and present some empirical evaluations of the models. The model was used to "predict" numbers of poor in 1990. We then compared the results to those from the 1990 Census to evaluate our model. The results of the comparisons are presented.

Section 1 describes some aspects of the data. Section 2 describes the small area estimation model we used. Section 3 describes the methods we used to form the county level estimates. Section 4 discusses the estimation of the variance components, section 5 describes the calculation of standard errors and confidence intervals, section 6 the raking and the necessary adjustment to the variance, and section 7 describes the model evaluations. We conclude in section 8.

1. Data

We form estimates of poverty for every county in the US. The approach is to use CPS estimates of the number of poor at the county level as the response variable in a regression equation with administrative records data as predictors. The CPS has sample in only about a third of the counties in the US. We form parameter estimates on the basis of those counties, then apply the estimated model to the remaining counties. Once we have the regression predictions, we form the Empirical Bayes (EB) estimator by taking a weighted average of the CPS direct estimate and the regression prediction. This way we can make use of the information in the individual CPS county estimates.

The data we get directly from the CPS need some modification. First, the primary sampling units (PSUs) in the CPS design include collections of counties or minor civil divisions; these PSUs are chosen from one or more in the strata. The weights for observations in CPS include a factor for the inverse of the probability of selection of the PSU. When we form county-level aggregates, then, we need to multiply by the probability of selection of the appropriate PSU so these aggregates are approximately unbiased at the county level. For details of the CPS design, see (Bureau of the Census, 1996).

In an effort to reduce the variance of the response in our regression model, we took a three-year weighted average of observations in the county, weighted by the number of housing units with children 5 to 17 years old in the poverty universe. This also had the effect of increasing the number of counties with any sample cases at all.

Specifically, if C_{ij} is the count of interviewed housing units in which at least one person age 5-17 is found in county *i* in year *j*, U_{ij} is the estimated number of related persons age 5-17 in the poverty universe in county *i* in year *j*, obtained using the CPS sampling weights adjusted to represent counties, and D_{ij} is the similarly estimated number of related persons age 5-17 in families in poverty in county *i* in year *j*, then the value of the poverty rate with which county *i* is characterized in the regressions is

$$S_i = \frac{\sum_j C_{ij}(D_{ij}/U_{ij})}{\sum_j C_{ij}} ,$$

The number of persons in the poverty universe in county *i* is

$$T_i = \frac{\sum_j C_{ij} U_{ij}}{\sum_j C_{ij}} ,$$

and the number of related persons age 5-17 in families in poverty with which county *i* is characterized is

$$P_i = S_i T_i .$$

2. The Model

We modeled the log of the 3-year average of CPS number of poor as a linear function of the logs of variables derived from administrative records data: food stamps, number of poor from tax forms, number of exemptions, population, and the last census number of poor. Put another way, we assume the regression model with two sources of "error", one associated with the deviation of the "true" county log number of poor from the mean regression curve and one associated with sampling in CPS. We will refer to the former as the county random effect or model error and the latter as sampling error.

We use a best linear unbiased predictor (BLUP) with a few modifications. In the usual BLUP case, we can estimate the variance of the error components provided some assumptions about the covariance structures are satisfied. These assumptions are satisfied here, but we thought the Census could provide some information about the fit of the model itself, that is, the magnitude of the random effects variance. We therefore modeled the Census using the same methods, assuming a common variance on the random effects. Then we estimated the random effects variance from the Census of the model.

A. CPS Model

We assume the vector of CPS estimates of log number of poor persons for the counties has sampling properties

$$\boldsymbol{Y}_c|\boldsymbol{\mu}_c \sim N[\boldsymbol{\mu}_c, \boldsymbol{V}_{c\epsilon}]$$

which should be read as a normal distribution with mean vector μ_c and covariance matrix $\mathbf{V}_{c\epsilon}$. We assume $\mathbf{V}_{c\epsilon}$ is diagonal. The mean vector has the distribution

$$\boldsymbol{\mu}_c \sim N[\boldsymbol{X}_c\beta_c, \boldsymbol{V}_u].$$

The covariance matrix $\mathbf{V}_u$ has the form $v_u\mathbf{I}$, for some scalar v_u. Here we assume that the random effects variance is constant across counties.

We can express this as a linear model:

$$\boldsymbol{Y}_c = \boldsymbol{X}_c\beta_c + \boldsymbol{u}_c + \boldsymbol{\epsilon}_c,$$

where $\mathbf{u}_c \sim N[\mathbf{0}, \mathbf{V}_u]$ and $\epsilon_c \sim N[\mathbf{0}, \mathbf{V}_{c\epsilon}]$. The $\mathbf{X}_c\beta_c$ term contains the explanatory variables including information from the administrative records. The second term is the random effect. The last term represents the sampling error.

The assumption of normality of the log number of poor, equivalent to an assumption of lognormality of the number of poor at the county level, is necessary only insofar as it justifies our calculation of variances, described below. It is also helpful for tests of hypotheses, but those are not our primary concern.

B. Census Model

The terms in the CPS model above are identifiable when the covariance structures $\mathbf{V}_u$ and $\mathbf{V}_{c\epsilon}$ are different, but we thought that the Census might have information about the random affects variance; indeed we thought the random effects variances might be the same, so we assume the same model holds for the Census and that the random effect variance $\mathbf{V}_u$ is the same for Census and CPS. Then we estimated the random effects variance from the Census data. The hope was that the Census, with its much higher precision, would yield better estimates of $\mathbf{V}_u$. The hope seems to have been borne out; see the *results* section.

We assume a Census model similar to the CPS model. We assume the vector of Census estimates of log number of poor persons for the counties has sampling properties

$$\boldsymbol{Y_d}|\boldsymbol{\mu_d} \sim N[\boldsymbol{\mu_d}, \boldsymbol{V_{d\epsilon}}]$$

Again, $\mathbf{V}_{d\epsilon}$ is diagonal. The mean vector has the distribution

$$\boldsymbol{\mu_d} \sim N[\boldsymbol{X_d}\beta_d, \boldsymbol{V_u}].$$

The covariance matrix $\mathbf{V}_u$ is common with that in the CPS model.

We can express this as a linear model just as we did in the CPS model:

$$\boldsymbol{Y_d} = \boldsymbol{X_d}\beta_d + \boldsymbol{u_d} + \boldsymbol{\epsilon_d},$$

where $\mathbf{u}_d \sim N[\mathbf{0}, \mathbf{V}_u]$ and $\epsilon_d \sim N[\mathbf{0}, \mathbf{V}_{d\epsilon}]$. The $\mathbf{X}_d\beta_d$ term contains the explanatory variables to describe the Census responses. The second term is the random effects term with variance common to the one in the CPS model. The last term represents the sampling error and is estimated with Generalized Variance Functions (GVFs) in the Census model. See (Bureau of the Census, 1993) Note the census part of the model is similar to the Fay-Herriot(1979) model.

3. Estimation

The shrinkage estimator has the form

$$\tilde{Y}_i = \hat{Y}_i + (1 - a_i)(Y_i - \hat{Y}_i).$$

where $\hat{Y}_i$ is the estimated mean, $\mathbf{X}\hat{\beta}$ and Y_i is the direct estimate, when it is available. The variable a_i is a weight which depends on the relative sizes or the variances of u and ϵ. An expression for the weights is given below. The variable a_i is chosen to minimize the expected squared difference between the estimator and the true county log number of poor μ_i.

The rule defined by a value a_i that minimizes the expected difference between $\hat{Y}_i$ and μ_i is the best linear unbiased predictor and the corresponding value for a_i is

$$a_i = \frac{V_{c\epsilon i}}{V_{c\epsilon i} + V_u},$$

$\tilde{Y}_i$ is estimated by performing a weighted least squares procedure with weights equal to

$$\textit{weight for county } i = \frac{1}{(\hat{V}_u + \hat{V}_{\epsilon i})}.$$

The estimation of $\hat{V}_u$ and $\hat{V}_{\epsilon i}$ is discussed in the next section.

4. Variance Component Estimation

It remains to estimate the variances V_u and $V_{c\epsilon}$. First we estimate the random effects variance, V_u. This information comes from the Census, which we assume has the same random effects variance as CPS. We apply the same basic form of this model to the Census as we apply to the CPS.

The GVF variances we had for the Census were for the number of poor. We needed to transform these to variances for the log number poor. Recall we assume the number poor in county i is lognormal. That is, if W_i is the number of poor in county i and
$Y_i = \ln(W_i)$,

$$Y_i|\mu_i \sim N(\mu_i, V_{\epsilon i})$$

and

$$W_i|\mu_i \sim LN(\xi_i, \eta^2{}_i).$$

We estimate $\eta^2{}_i$ with Census GVFs and solve for $V_{\epsilon i}$.

If we fit either the CPS or the Census regression with ordinary least squares, we have

$$E(SSE) = E(\boldsymbol{Y}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{Y}) = E(\boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y})$$

$$= \sum V_{ui} tr(\boldsymbol{M}\, diag(\boldsymbol{e_i})) + \sum V_{\epsilon i} tr(\boldsymbol{M}\, diag(\boldsymbol{e_i}))$$

where $\mathbf{e}_i$ is the i^{th} column of the identity matrix and $\mathbf{M}=\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Now let m_{ii} be the i^{th} diagonal of the projection matrix $\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ or one minus the leverage of the i^{th} observation. In our application, we have $V_{ui}=V_u$ for all i, so

$$E(SSE) = V_u df_e + \sum V_{\epsilon i} m_{ii},$$

and

$$E(MSE) = V_u + \frac{1}{df_e}\sum V_{\epsilon i} m_{ii}.$$

If the $V_{\epsilon i}$ have unbiased estimators $\hat{V}_{\epsilon i}$, an unbiased estimator for V_u is

$$\hat{V}_u = MSE - \frac{1}{df_e}\sum \hat{V}_{\epsilon i} m_{ii}.$$

See (Fay and Herriot, 1979) or Christiansen (1987). This is just the MINQUE(0) (Christiansen, (1987) estimator with $V_{\epsilon i}$ given.

If we fit OLS to the CPS model,

$$E(MSE) = V_u + \frac{1}{df_e}\sum V_{c\epsilon i} m_{ii}.$$

We make the assumption that $V_{c\epsilon i} = \sigma^2 d_i$, where d_i is known. Here, $d_i = 1/n_i$, Where n_i is the CPS sample size in county i. Note this is approximately equivalent to the assumption that $cv^2(W_i) = \sigma^2/n_i$. Some other models for d_i are examined in section 7. The equation is

$$E(MSE) = V_u + \frac{\sigma^2}{df_e}\sum_i \frac{m_{ii}}{n_i}.$$

Solving for σ^2 yields

$$\hat{\sigma}^2 = \frac{df_e E(MSE) - df_e V_u}{\sum_i \frac{m_{ii}}{n_i}}.$$

Now we can write our estimator for $V_{\epsilon i}$ in CPS for log number of poor :

$$= \frac{SSE - df_e V_u}{n_i \sum_j \frac{m_{jj}}{n_j}}.$$

$$\hat{V}_{c\epsilon i} = \hat{\sigma}^2 / n_i$$

This estimator is an unbiased quadratic estimator for $V_{c\epsilon i}$. *(Reference?)*

5. Standard Errors and Confidence Intervals

The expected squared error for the EB estimator, if we ignore the variance of the a_i's, is

$$R_i = a^2{}_i V(\hat{Y}_i) + a_i V_u.$$

See Henderson(1975). In many situations, the variance of the a_i's can be pretty important in that the variance contributed by the estimation of the EB weights may not be negligible compared to the other variances. One example is the Fay-Herriot model, which is very similar to the one above, except the sampling variance is known and the random effects variance is estimated. (Fay and Herriot, 1979) In that case,- the estimator above tends to nonnegligibly underestimate the total variance. In our case, the CPS sampling variance is estimated from the data and, at this writing, the census sampling variance is assumed known. In this case, the underestimation of variance, as measured with the method described by Prasad & Rao(1990), seems to be negligible.

It's not obvious from the equation above, but when the a_i's are close to 1, the variances of the final estimates are close to that of the $\hat{Y}_i$'s. In our application, then, the variances of the estimates in the log scale are somewhat uniform, so the CV's of the estimated number of poor are somewhat uniform.

5.1 Transformations of Estimates back to 'Number of Poor'.

It remains to convert our estimates of log number of poor to estimates for number of poor. To convert the estimates themselves, we note the mean of a lognormal distribution is $\exp(\mu + \sigma^2/2)$, where μ and σ^2 are the mean and variance, respectively, of the corresponding normal distribution. We form the same transformation with $\sigma^2 = \hat{V}_u$.

The variance estimates also need to be formed for the number of poor. We do this simply by recognizing that $\text{var}(\log(w)) \approx cv^2(w)$, which is an approximation that works well when the right hand side of that equation is small.

5.2 Confidence Intervals

We follow the common practice of forming our confidence intervals thus:

$$(\tilde{Y}_i - z_{\alpha/2} R_i^{1/2} \ , \ \tilde{Y}_i + z_{\alpha/2} R_i^{1/2}),$$

where $z_{\alpha/2}$ is the 1-α/2 quantile of the standard normal distribution. See Morris(1983). This gives us symmetric confidence intervals, which are not completely appropriate for the lognormally distributed number of poor.

6. Raking

One of the requirements of our estimators is that they sum to the State estimates (Fay 97). They are not constrained to this in the estimation, so ratio adjustment step was included to ensure that they do. County j in state i is multiplied by the ratio of the state estimate of poor to the sum of the county estimates of poor. That is,

$$\hat{W}_{ij} = \frac{Y_{Tj} \tilde{Y}_{ij}}{\sum_i \tilde{Y}_{ij}},$$

where $\hat{W}_{ij}$ is the so-called raked estimator of poor in state j, county i, $\tilde{Y}_{Tj}$ is the estimate of poor in state j, and $\tilde{Y}_{ij}$ is the estimate for county i in state j. The variance on this estimator is different from the unraked estimator. The new variance is approximated with a Taylor expansion about the expectations of the three factors in the above equation. Unfortunately, this depends on the covariance of the county- and state-based estimates, which has not been estimated.

7. Some Evaluation Notes

It would never do to consider a model with no alternatives and not evaluate any of the assumptions. It is also nice to consider some alternatives with respect to the estimation procedure itself. In this section we examine some of these topics.

7.1 Other Models

We have a list of alternative models, some suggested by our colleagues at the National Academy of Science. In this paper we consider the estimates made for 1990, for the purpose of comparison to the Census, and for 1994, and make comparisons to two other models. The first is based on the assumption that the county shares within each state are the same as at the previous Census; the Census counts, then, are simply raked to the state-based estimates of Fay. We will refer to this estimate as U1. The second estimator is based on the assumption that the county ratios, poor/population, are the same as at the previous census; the ratios from the previous Census are multiplied by the current population estimate and raked the resulting numbers to the state estimate. We refer to this model as U2.

Another model modification proposed by the panel (NRC, 1997) models the log of the rate rather than the log number poor but keeps the set of dependent variables described above. We present it here because it looks competitive with the SAIPE model where a number of other proposals have turned out not to be as interesting. We refer to this model as D.

A minimum requirement for our estimates is that they perform better than the Census. Models U1 and U2 seem like very straightforward improvements over the Census, and we would like our model to do better than they do.

7.1.1. Numerical Results

We have several criteria for the evaluation of the models and estimates, including the traditional regression diagnostics and comparisons to the Census in 1990. We content ourselves with Table 1, which shows the mean relative difference and mean absolute relative difference between the estimates and the Census number of poor for children 5 to 17 years old for each of the three models. The relative difference is

$$\frac{model\ \ estimate\ \ of\ \ number\ \ of\ \ poor}{census\ \ estimate\ \ of\ \ number\ \ of\ \ poor} - 1.$$

Note the numbers in Table 1 are reported for the raked estimates, since the models U1 and U2 are by definition raked.

Table 1.
Measure of Comparison to the Census
for some Models.

	Mean Relative Difference	Mean Absolute Relative Difference
U1	17.5%	29.3
U2	15.65%	27.0%
SAIPE Model	2.9%	15.7%
D	1.7%	17.1%

Clearly the SAIPE model and the D model are better than U1 or U2, at least as we measure it here. We chose the SAIPE model over the D model partly for its improved performance with the mean absolute relative difference. There was also the consideration that the estimated rates in the D model would need to be converted into a number of poor, that being our parameter of interest, and in so doing, we would need to multiply by an estimate of a population. We were not able to evaluate the quality of the population estimates, so we were not sure what contribution they would make to either the variance or bias of the final estimates.

8. Conclusion

We formulated a model to estimate the number of poor at the county level by forming the regression of CPS direct estimates on administrative records data. We suggested some criteria by which we could judge the model and showed that the model performs better by these criteria than some of the more obvious alternatives. We also examined some variations one some of our methods and assumptions and we saw that we did not do too badly compared to them, although we saw how we might make some improvements. In particular, we may use a slightly different model for the variances and we may replace the constrained MINQUE(0) estimator with the MLE.

We did not consider other models in this paper. Several have been suggested, particularly some which model rates as a function of administrative records data. That has been left for another paper.

9. References

Bureau of the Census (1996), "CPS Annual Demographic Survey -- March Supplement" http://www.bls.census.gov/cps/ads/adsmain.htm.

Bureau of the Census (1993), *1990 Census of Population, Social and Economic Characteristics for Various States (1990 CP-2-various)*, Appendix C, US Government Printing Office, Washington, DC.

Christiansen, Ronald (1987), *Plane Answers to Complex Questions*, Springer-Verlag, New York.

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

Henderson, C. R., (1975), "Best Linear Unbiased Estimation and Prediction Under A Selection Model," *Biometrics*, 31, 4423-447

Prasad, N. G., and Rao, J. N. K., (1990), "The Estimation of the Mean Squared Error of Small-Area Estimators", *Journal of the American Statistical Association*, 85, 163-171

Acknowledgments

The author is grateful to Richard Griffiths and to Carol King for their vaulable comments and to Paul Siegel for his valuable contibutions.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author(s) and do not necessarily reflect those of the Census Bureau.