Pilot Study on Combining Direct Estimates of Income and Poverty from the American Community Survey with Predictions from a Model

Rick Griffin, U.S. Census Bureau

Abstract[1]

Small area estimation using models that borrow strength from relationships between variables across geographic areas has become increasingly popular. Typically, these approaches combine direct estimates with model estimates. The American Community Survey (ACS) produces direct five-year estimates at the census tract level for income and poverty. An improvement in the accuracy of these estimates as measured by the estimated sampling error is desired. This pilot study will compare the mean squared error of three potential model-based estimation methods with direct five-year estimates of income and poverty at the tract level. The goal is to make a preliminary assessment on the potential gain in accuracy from using these model-based estimates. For maximum improvement, these models require administrative data correlated with income or poverty. If one of these methods could produce significant improvement in accuracy of the estimates, we want to investigate the development of an application for data users to use publically available administrative data to produce model-based estimates to combine with the published ACS direct estimates in order to improve accuracy.

Key Words: Borrow Strength; Multivariate Regression; Measurement Error: Empirical Bayes; Small Area Estimation;

1. Introduction

The U.S. Census Bureau is investigating model-based improvement of American Community Survey (ACS) poverty and income estimates. The first goal is to develop a model-based estimation process that creates improvement in mean squared error for ACS five year estimates of poverty and income. This paper is limited to estimation methodologies using empirical Bayes approaches. All these approaches result in a small area estimate that is a weighted average of the direct estimate and the model estimate, which borrows strength from data on the relationship between dependent variables and independent variables across all small areas. These weights are functions of the estimate of the model error and the estimate of the direct estimate's sampling error. The weights are functions of sampling error and model error estimated from the data. Three general approaches are considered: (1) the classical Fay-Herriot (1979) empirical Bayes approach; (2) a multivariate regression extension of the classical Fay-Herriot model; and (3) a model adapted to handle measurement error in independent predictor variables.

It has been suggested that perhaps ACS estimates correlated with poverty and income could be used as predictor variables. However, there are consequences when the independent variables are estimates with non-trivial sampling variances. Fay (1987) and Datta, Fay and Ghosh (1991) describe multivariate Bayes analysis in small area estimation that uses these correlated estimates

[1] *This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.*

as additional dependent variables. They consider applications where the independent variable Z to be used in estimating Y comes from the same survey that is used to estimate Y. The treatment of Z as part of the independent variables X in standard linear regression may give misleading estimates depending on the nature of the sampling covariances between Y and Z. Viewing the problem as multivariate linear regression for the combined vector (Y,Z) may lead to a more correct formulation of the problem. Estimation of model error can be done several ways. Examples are maximum likelihood, restricted maximum likelihood, and method of moments. A simple unbiased method of moments estimator suggested by Prasad and Rao (1993) is used for this simulation study. The other methods require iteration to convergence but likely have smaller variance.

Another approach using empirical Bayes methods to deal with measurement error in independent variables is suggested by Ybarra and Lohr (2008). They present an empirical Bayes small area estimator for which the classical Fay-Herriot model is expanded allowing for measurement error (in our application, sampling error) while still treating the predictor variable as an independent variable. Their paper assumes that the estimated independent predictor variable is uncorrelated with the target estimated dependent variable. However, their formulas are expanded to account for such correlation in an unpublished Ph.D. thesis (Ybarra 2003). Here we will use the formulas allowing for this correlation.

This paper describes the methodology and provides results of a simulation study comparing the mean squared errors of small area estimates using direct estimates, the classical Fay-Herriot empirical Bayes approach, the multivariate regression empirical Bayes approach, and the empirical Bayes method incorporating measurement error in independent variables. This initial work will use models with one independent variable and a constant term for each dependent variable. The direct estimate uses the survey weights for sample in the small area only and is assumed unbiased. For the simulation study, the true value of the small area statistics to be estimated is known and the empirical mean and variance over a large number of sample draws will be used. Thus, the mean squared errors can be estimated.

2. Overview of Simulation Study Plan

The plan is to use data from ACS five year tract-level estimates (2006-2010) data available on www.census.gov for Erie County, Pennsylvania (70 tracts). Two poverty and income statistics will be used. The ACS estimates will be treated as the true values. For some of the Empirical Bayes based estimators, independent administrative record variables are required for each tract. These will be generated assuming a simple regression model between the true value (assumed known) and the independent administrative record variable.

Using the published ACS margins of error, the sampling variances of these statistics will be calculated. Assuming normality, 1000 unbiased ACS estimates will be generated for each statistic for each tract. Note that the assumption of normality is a theoretical limitation since income and poverty estimates are likely to have skewed distributions such as the Pareto distribution for income. Thus, we will have the true values for each statistic, the needed administrative record variables, and 1000 independent unbiased direct estimates for each tract. Using this information, the classical Fay-Herriot empirical Bayes estimator, the multivariate empirical Bayes estimator, and the Fay-Herriot empirical Bayes estimator incorporating measurement error in independent variables will be calculated for each of the two statistics for each tract for each of the 1000 simulated sets of ACS direct estimates. Using these data the empirical mean squared error (MSE) will be calculated for each of the four estimators for each tract.

2.1. Calculation of True Values

The two statistics we will estimate are (1) the estimated number of families with income less than the poverty rate in the last 12 months and (2) the average family income. ACS 5 year (2006-2010) data is used.

$\hat{Y}_j$ = estimated number of families with income in the past 12 months below poverty level for tract j

$\hat{Z}_j$ = estimated average family income in the past 12 months (in inflation adjusted dollars) for tract j

$\theta_{jT1} = \hat{Y}_j$ = value treated as the true number of families with income in the past 12 months below poverty level for tract j (1)

$\theta_{jT2} = \hat{Z}_j$ = value treated as the true average family income for tract j (2)

2.2. Generating the Independent Administrative Record Variables

The approach will be the same for generating the administrative record variable associated with each of the two true statistics for each tract. Thus, the 1 and 2 subscripts are omitted in this section.

θ_{jT} = the true statistic for tract j

A_j = the independent administrative record variable associated with this true statistic for tract j

Assume the following simple linear model holds with known parameters α.

$A_j = \alpha_0 + \alpha_1 \theta_{jT} + v_j$ with v_j independent $N(0, \sigma_v^2)$

Assume $\sigma_v^2 = K\bar{\theta}_T$. ; where $\bar{\theta}_T = \dfrac{\sum_{j=1}^{70} \theta_{jT}}{70}$. Note that σ_v^2 is different for each of the two true statistics.

Start with a given value of K, for example K = .5, generate v_j from $N(0, \sigma_v^2)$ independently for each tract j.

For the parameters α, initially use $\alpha_0 = (.1)\overline{\theta_T}$ and $\alpha_1 = 1.1$.

Thus $A_j = .1\bar{\theta}_T + 1.1\theta_{jT} + v_j$

Once the A_j values are calculated run, a simple ordinary least squares regression using the model $\theta_j = \beta_0 + \beta_1 A_j + \varepsilon_j$. Fit the model and find the R^2 value. The plan is to experiment with K values to find R^2 values we want to use for estimation. Several R^2 values will be used to determine the sensitivity of results to the quality of the administrative records.

2.3. Generating 1000 direct estimates for each tract

For each tract j, $\psi_{j1} = Var(\hat{\theta}_{j1}), \psi_{j2} = Var(\hat{\theta}_{j2})$. These variance terms are assumed known for each tract based on published ACS margins of error (divide the ACS 90% error value by 1.645 to get the standard error).

For k = 1,….,1000, generate $\hat{\theta}_{jk1}$ from $N(\theta_{jT1}, \psi_{j1})$ and $\hat{\theta}_{2jk}$ from $N(\theta_{jT2}, \psi_{j2})$. Here we are treating the production ACS estimate as the true value and generating 1000 simulated ACS estimates with the calculated variances.

2.4. Classical Fay-Herriot Model and Estimation (one independent variable)

The methodology is the same for the estimated household poverty rate and average household income so the 1 and 2 subscripts are omitted.

For each tract j and for each k = 1,….,1000, assume that the unbiased small area estimate $\hat{\theta}_{jk}$ is related to auxiliary data $\alpha_j = (1, A_j)^T$ through a linear model.

$\hat{\theta}_{jk} = \theta_{jT} + e_{jk}$ and $\theta_{jT} = \alpha_j^T \beta + v_j, j = 1,...,m$ (m is the number of tracts)
where $\beta = (\beta_0, \beta_1)^T$ is the 2x1 vector of regression coefficients, e_{jk} are independent $N(0, \psi_j)$, v_j are independent $N(0, \sigma_v^2)$ and e_{jk} and v_j are independent.

An estimate of the model variance, calculated independently for each k, from Prasad and Rao (1990) is as follows

$$\hat{\sigma}_{kv}^2 = \frac{1}{m-2}[\sum_{j=1}^{m}(\hat{\theta}_{jk} - \hat{\beta}_{0K,OLS} - \hat{\beta}_{1K,OLS} A_j)^2 - \sum_{j=1}^{m}\psi_j(1 - \alpha_j^T (A^T A)^{-1} \alpha_j) \quad (3)$$

where OLS indicates ordinary least squares estimation is used (no sampling or model error terms needed) and

$$A = \begin{pmatrix} 1 & A_1 \\ . & . \\ . & . \\ . & . \\ 1 & A_m \end{pmatrix}.$$

The Empirical Bayes estimates are as follows

$$\hat{\beta}_k = \left(\sum_{j=1}^{m} \frac{\alpha_j \alpha_j^T}{\psi_j + \hat{\sigma}_{kv}^2} \right)^{-1} \sum_{j=1}^{m} \frac{\alpha_j \hat{\theta}_{jk}}{\psi_j + \hat{\sigma}_{kv}^2}$$

$$\hat{\gamma}_{jk} = \frac{\hat{\sigma}_{kv}^2}{\psi_j + \hat{\sigma}_{kv}^2}$$

$$\hat{\theta}_{jk,FHClassical} = \hat{\gamma}_{jk} \hat{\theta}_{jk} + (1 - \hat{\gamma}_{jk}) \alpha_j^T \hat{\beta}_k$$

The empirical mean squared error is estimated by:

$$\hat{MSE}(\hat{\theta}_{j,FHClassical}) = \frac{1}{1000} \sum_{k=1}^{1000} (\hat{\theta}_{jk,FHClassical} - \theta_{jT})^2$$

2.5. Multivariate Model and Estimation (two independent variables)

For each j and k = 1,….,1000, the basic data are the two component vectors $\hat{\theta}_{jk} = (\hat{\theta}_{jk1}, \hat{\theta}_{jk2})^T$ j = 1,…,m. $\hat{\theta}_{jk1}$ is the estimate of interest and $\hat{\theta}_{jk2}$ is believed to be strongly correlated with it. Note that either one could be considered the estimate of interest.

Let $\theta_j = (\theta_{jT1}, \theta_{jT2})^T$.

$\hat{\theta}_{jk}$ are independent $N(\theta_j, \psi_j)$, where $\psi_j = \begin{pmatrix} \psi_{j1} & \psi_{j12} \\ \psi_{j12} & \psi_{j2} \end{pmatrix}$

$$A_j = \begin{pmatrix} 1 & A_{ij} & 0 & 0 \\ 0 & 0 & 1 & A_{j2} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

θ_j are independent $N\left(\begin{pmatrix} \beta_1 + \beta_2 A_{j1} \\ \beta_3 + \beta_4 A_{j2} \end{pmatrix}, \begin{pmatrix} \sigma_{v1}^2 & \sigma_{v12} \\ \sigma_{v12} & \sigma_{v2}^2 \end{pmatrix} \right)$

The diagonal terms of the covariance matrix for the vector θ_j are each estimated using equation (3). Note that multivariate regression is the same as univariate regression if all errors are given the same weight as is done for ordinary least squares. Thus, equation (3) can be used twice, once for $\hat{\sigma}_{vk1}^2$ and once for $\hat{\sigma}_{vk2}^2$. Then use

$$\hat{\sigma}_{vk12} = \frac{1}{m-2} [\sum_{j=1}^{m} (\hat{\theta}_{jk1} - \hat{\beta}_{k1,OLS} - \hat{\beta}_{k2,OLS} A_{j1})(\hat{\theta}_{jk2} - \hat{\beta}_{k3,OLS} - \hat{\beta}_{k4,OLS} A_{j2})$$

Let $\hat{D}_k = \begin{pmatrix} \hat{\sigma}^2_{vk1} & \hat{\sigma}_{vk12} \\ \hat{\sigma}_{vk12} & \hat{\sigma}^2_{vk2} \end{pmatrix}$

Then the Empirical Bayes multivariate estimator is given by

$$\hat{\beta}_k = \left(\sum_{j=1}^{m} A_j^T (\psi_j + \hat{D}_k)^{-1} A_j \right)^{-1} \sum_{j=1}^{m} A_j^T (\hat{\psi}_j + \hat{D}_k)^{-1} \hat{\theta}_{jk}$$

$$\hat{\theta}_{jk,multi} = \hat{D}_k (\psi_j + \hat{D}_k)^{-1} \hat{\theta}_{jk} + \psi_{ij} (\psi_j + \hat{D}_k)^{-1} A_j \hat{\beta}_k = (\hat{\theta}_{jk1,multi}, \hat{\theta}_{jk2,multi})^T$$

For s = 1,2 $\hat{MSE}(\hat{\theta}_{jks,multi}) = \frac{1}{1000} \sum_{k=1}^{1000} (\hat{\theta}_{jks,multi} - \theta_{jTs})^2$.

2.6. Measurement Error Model and Estimation (one independent variable)

Either $\hat{\theta}_{jk1}$ or $\hat{\theta}_{jk2}$ can be the independent variable with measurement error (i.e., sampling error). Here $\hat{\theta}_{jk2}$ is the independent variable.

$$X_j = \begin{pmatrix} 1 \\ \theta_{jT2} \end{pmatrix} \quad \hat{X}_{jk} = \begin{pmatrix} 1 \\ \hat{\theta}_{jk2} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\hat{\theta}_{jk1} = X_j^T \beta + (X_j - \hat{X}_{jk})^T \beta + v_j + e_j \quad r_j = (X_j - \hat{X}_{jk})^T \beta + v_j$$

e_j are independent sampling errors $N(0, \psi_j)$, v_j are independent model errors $N(0, \sigma_v^2)$ and e_i and v_i are independent.

$$C_j = MSE(\hat{X}_{jk}) = \begin{pmatrix} 0 & 0 \\ 0 & \psi_{j2} \end{pmatrix} \quad C_{j\hat{x}\hat{\theta}} = \begin{pmatrix} 0 \\ \psi_{j12} \end{pmatrix}$$

Assume that the sample variance and covariance terms are known although in practice they need to be estimated.

$$MSE(r_j) = \sigma_v^2 + \beta_1^2 \psi_{j2}$$

$$\mathrm{cov}(r_j, e_j) = -\beta_1 \psi_{j12}$$

$$MSE(r_j + e_j) = \sigma_v^2 + \beta_1^2 \psi_{j2} + \psi_{j1} - 2\beta_1 \psi_{j12}$$

If β and σ_v^2 were known, the minimum mean squared error estimator amongst all linear combinations of $\hat{\theta}_{jk}$ and $\hat{X}_j^T \beta$ is given by

$\hat{\theta}_{jk,measerror} = \gamma_j \hat{\theta}_{jk} + (1-\gamma_j)\hat{X}^T_{jk}\beta$, where $\gamma_j = \dfrac{MSE(r_j) - Cov(r_j, e_j)}{MSE(r_j + e_j)}$.

Since β and σ_v^2

Since β and σ_v^2 are unknown, first let $\hat{\beta}_k^{(1)} = \left(\sum_{j=1}^{m} \hat{X}_{jk}\hat{X}^T_{jk} - C_j \right)^{-1} \left(\sum_{j=1}^{m} \hat{X}_{jk}\hat{\theta}_{jk1} - C_{j\hat{X}\hat{\theta}} \right)$

Note: The matrix to be inverted may not exist. If that happens for a given k, C_j and $C_{j\hat{X}\hat{\theta}}$ will be set to 0. The frequency of this event will be tabulated. **Note that this event did not occur for these simulations.**

Then $\hat{\sigma}^2_{kv} = m^{-1}\sum_{j=1}^{m}(\hat{\theta}_{jk} - \hat{X}^T_{jk}\hat{\beta}_k^{(1)})^2 - \hat{\psi}_j - \hat{\beta}_{k1}^{(1)}\psi_{j2} + 2\hat{\beta}_{k1}^{(1)}\psi_{j12}$

Then compute $\hat{\beta}^{(2)}$ using the weights $w_j = \dfrac{1}{\hat{\sigma}^2_{kv} + (\hat{\beta}^{(1)}_{k1})^2\psi_{j2} - 2\hat{\beta}^{(1)}_{k1}\psi_{j12} + \psi_{j1}}$.

$$\hat{\beta}_k^{(2)} = \left(\sum_{j=1}^{m} w_j(\hat{X}_{jk}\hat{X}^T_{jk} - C_j) \right)^{-1} \left(\sum_{j=1}^{m} w_j(\hat{X}_{jk}\hat{\theta}_{jk} - C_{j\hat{X}\hat{\theta}}) \right)$$

Then $\hat{\theta}_{jk,measerror} = \hat{\gamma}_{jk}\hat{\theta}_{jk} + (1-\hat{\gamma}_{jk})\hat{X}^T_{jk}\hat{\beta}_k^{(2)}$ where

$$\hat{\gamma}_{jk} = \frac{\hat{\sigma}^2_{kv} + (\hat{\beta}^{(2)}_{k1})^2\psi_{j2} - \hat{\beta}^{(2)}_{k1}\psi_{j12}}{\hat{\sigma}^2_{kv} + (\hat{\beta}^{(2)}_{k1})^2\psi_{j2} - 2\hat{\beta}^{(2)}_{k1}\psi_{j12} + \psi_{j1}}$$

$$\hat{MSE}(\hat{\theta}_{j,measerror}) = \frac{1}{1000}\sum_{k=1}^{1000}(\hat{\theta}_{jk,measerror} - \theta_{j1T})^2$$

3. Results and Summary of Simulations

For tract j and model-based estimator t (classical Fay-Herriot, multivariate regression, or measurement error), let $Z_j = \dfrac{\hat{MSE}(\hat{\theta}_{j,t})}{Var_{sampling}(\hat{\theta}_{j,t})}$

and $\overline{Z} = \dfrac{\sum_{j=1}^{70} Z_j}{70}$. This statistic is the average ratio over the 70 tracts of the mean squared error of the model based estimator and the mean squared error of the direct estimator (equal to the sampling variance since we assume the direct estimator is unbiased). An average ratio less than 1

indicates an improvement using the model based estimator. Var($\hat{MSE}(\hat{\theta}_{j,t})$) is very small for each tract j due to having 1000 simulations. Thus the variance of $\bar{Z}$ is negligible.

3.1 Classical Fay-Herriot Estimator

For each target estimate (number of families below poverty level and average family income), two sets of administrative variables were obtained with varying predictive value for the true population value (as measured by R^2). Table 1 provides the results. The p value for the F test of the null hypothesis that the regression coefficient for the single independent value is 0 is shown. These are very small (all much less than .001) and shown for comparison purposes only. Larger R^2 values have smaller p values as expected.

Table 1 Average Ratio (Z) of Fay/Herriot MSE to Direct MSE

	R^2 for AD (p value for Ho:β=0)	Average Z
Average Family Income	.46 (1.3e-10)	0.9271
	.74 (2.2e-16)	0.8581
Families Below Poverty Level	.50 (6.8e-12)	0.7458
	.80 (2.2e-16)	0.5940

Increasing the R^2 value increases the improvement from model-based estimation. For average family income, increasing R^2 from .46 to .74 increased a gain of about 7% in MSE to a gain of about 14%. For the number of families below the poverty level, increasing R^2 from .50 to .80 improved a gain of about 25% in MSE to a gain of about 41%. Note that the average coefficient of variation (CV) for the direct poverty estimate was .414, while for the direct income estimate the average CV was .094. There is greater potential for improvement in MSE for direct estimates that are less reliable.

3.2 Multivariate Regression Estimator

Two simulations using average family income and number of families in poverty were done. The first simulation used administrative data (AD) with R^2 values of .46 and .50 for income and poverty respectively. The second simulation used AD with R^2 values of .72 and .77 for income and poverty respectively. We wanted to see if either of these improved on the classical Fay/Herriot estimator with comparably accurate AD. In addition, another ACS estimate highly correlated with both average family income and number of families below the poverty line f(median owner occupied housing unit value) was simulated as a second dependent variable for

both average family income and number of families below the poverty line. The average CV for the tract level direct estimate of owner occupied housing unit value was .061. We wanted to see if using an estimate with less sampling error as the second dependent variable would improve estimation. Table 2 provides the results.

Table 2 Average Ratio (Z) of Multivariate Regression MSE to Direct MSE

Dependent Variables	R^2 for AD (p value for Ho:β=0)	Average Z
1. Avg. family income	.46 (1.3e-10)	.9333
2. #Families in Poverty	.50 (6.8e-12)	.6484
1. Avg. family income	.72 (2.2e-16)	.8537
2. #Families in Poverty	.77 (2.2e-16)	.6193
1. Avg. family income	.74 (2.2e-16)	.8592
2. Median Owner Occ. HU Value	.73 (2.2e-16)	.9146
1. #Families in Poverty	.50 (6.8e-12)	.7565
2. Median Owner Occ. HU Value	.73 (2.2e-16)	.9136

Comparing Table 2 with Table 1, there is little difference between the results for average family income. The improvements remain at about 7% and 14% depending on the R^2 value for the AD. For the number of families in poverty with R^2 of .50 for the AD using average family income with R^2 of .46 for the AD as the second dependent variable, multivariate regression improved a gain of about 25% in MSE to a gain of about 35%. However, for the number of families in poverty with R^2 of .50 for the AD using median owner occupied housing unit value with R^2 of .73 for the AD as the second dependent variable, multivariate regression produced a gain of about 25% the same as the classical Fay/Herriot model. There was also little change for poverty with a R^2 of .80 (41% gain in MSE for Fay/Herriot) and the multivariate using poverty with R^2 of .77 for AD and income with R^2 of .74 for AD (38% gain in MSE).

3.3 Measurement Error Model Estimator

For average family income the measurement error model estimator was simulated three times with number of families in poverty, median owner occupied housing unit value, and number of

female head of household families with children less than 18 years of age and no husband as the independent variable with measurement (sampling) error. Three simulations were also done for number of families in poverty using average family income, median owner occupied housing unit value, and number of female head of household families with children less than 18 years of age and no husband as the independent variables. Table 3 provides the results.

Table 3 Average Ratio (Z) of Measurement Error MSE to Direct MSE

Dependent	Independent	Average Z
Avg. Family Income	#Families in Poverty	.9939
#Families in Poverty	Avg. Family Income	.8207
Avg. Family Income	Median Owner Occ. HU Value	.9312
#Families in Poverty	Median Owner Occ. HU Value	.8300
Avg. Family Income	Female HH,child<18;no husband	.9745
#Families in Poverty	Female HH,child<18;no husband	.8736

The best gain in MSE for average family income was about a 7% gain using median owner occupied housing unit value as the independent variable. This is the same gain as from the Fay/Herriot model with R^2 of .46 for AD but not as good as the 14% gain with R^2 of .74 for AD. For number of families in poverty, gains of about 18% and 17% were obtained using average family income and median owner occupied housing unit value as the independent variables respectively. These gains in MSE are not as good is the gain of 25% from the Fay/Herriot model with R^2 of .50 for AD and the 41% gain with R^2 of .80 for AD. However, the measurement error model uses only ACS data, requiring no administrative data.

3.4 Plots

All the model based estimates simulated are weighted averages of the direct estimate for the tract and a weighted regression estimator that borrows strength from other tracts. For more reliable direct estimates, the model based shrinkage estimator relies more heavily on the direct estimates. This can be illustrated by plotting the average ratio of the model based estimate to the direct estimate over the 1000 simulations as a function of the direct estimate coefficient of variation for each tract. This is illustrated for selected simulations in Plots 1 through 6 shown at the end of the paper. A least squares regression line (red) and locally weighted polynomial regression line (blue) are shown on each plot.

4. Summary

The simulations of the classical Fay Herriot Model produced a reduction in average mean square error for average family income of about 7 percent and 14% for two simulations with varying correlation from the administrative data. For number of families below the poverty line, two simulations of the Fay/Herriot model produced a reduction in MSE of 25% and 41%. Using two dependent variables for a multivariate regression model did not show any further reduction in estimation of average family income over the Fay/Herriot model for three simulations. For number of families below the poverty line, two of three simulations produced about the same reduction in MSE using multivariate regression as using the Fay/Herriot model. The multivariate poverty estimate using median owner occupied housing unit value as a second dependent variable produced about a 35% gain in MSE compared with a 25% gain for the Fay/Herriot model. The measurement error model produced a maximum gain in MSE of 7% for average family income and 18% for number of families below the poverty line. The measurement error model uses only ACS data, requiring no administrative data.

Greater percentage gains in MSE for number of families in poverty than for average family income are likely obtained because the sampling error average coefficient of variation is higher for the poverty estimate (41%) than the CV for the income estimate (9%). Plots demonstrate that model based shrinkage estimates are closer to the direct estimates when the direct estimates have smaller CVs.

References

Datta, G.S., Fay R.E., and Ghosh, M. (1991), Hierarchical and Empirical Multivariate Bayes Analysis in Small Area Estimation," *Proceedings of the 1991 Annual Research Conference*, Bureau of the Census.

Fay, R. E. (1987), "Application of Multivariate Regression to Small Domain Estimation," *Small Area Statistics*, Wiley and Sons, 91-102.

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

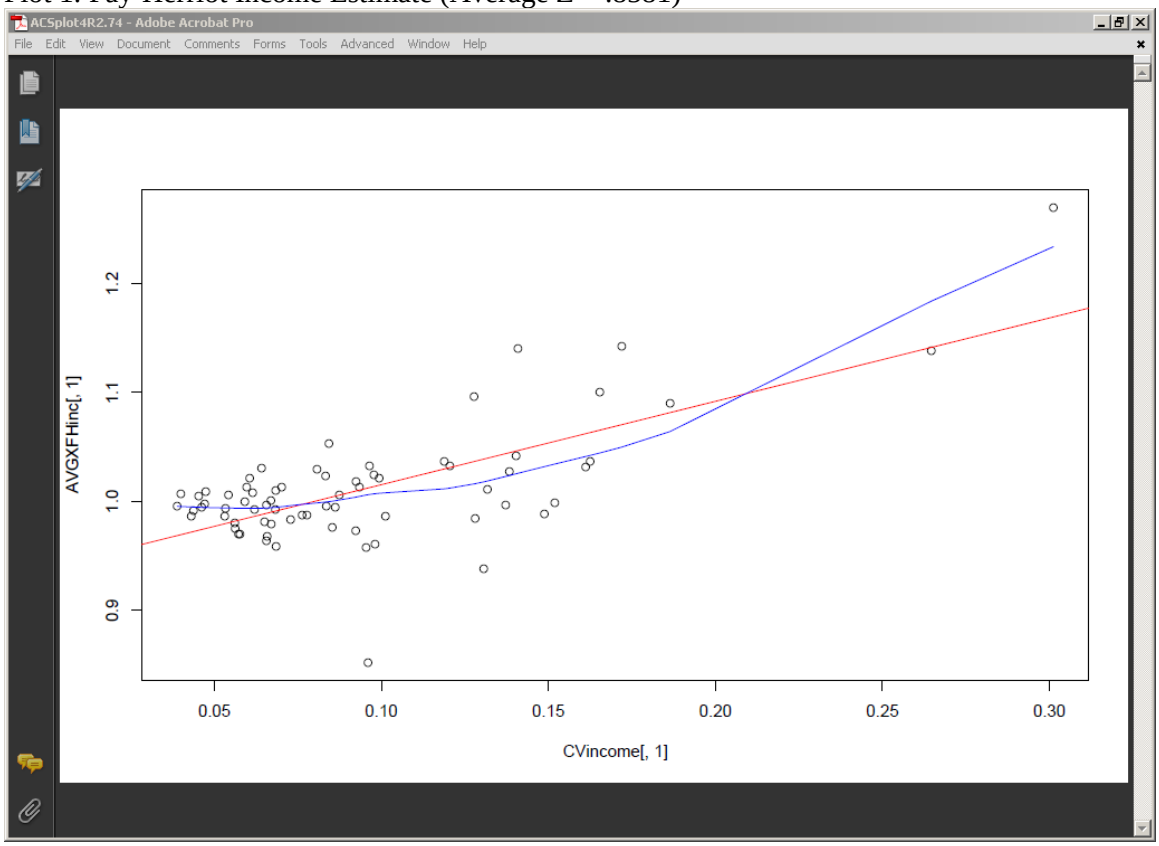
Prasad, N.G.N. and Rao, J.N.K. (1990), "The Estimation of Mean Squared Error of Small Area Estimators," *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K. (2003), "Small Area Estimation", *Wiley and Sons.*

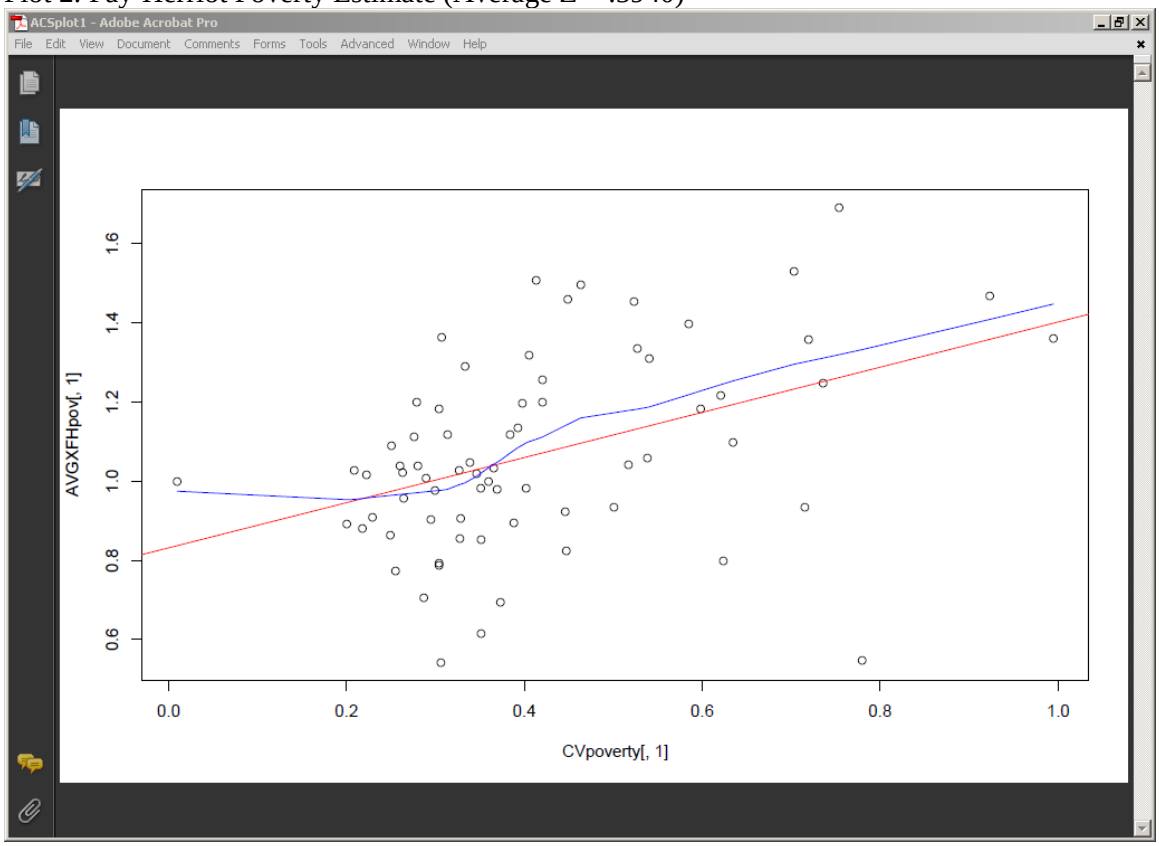
Ybarra, L.M.R., unpublished PH.D thesis, Arizona State University

Ybarra, L.M.R. and Lohr, S. L. (2008), "Small are estimation when auxiliary information is measured with error", *Biometrika*, 95, 4, 919-931

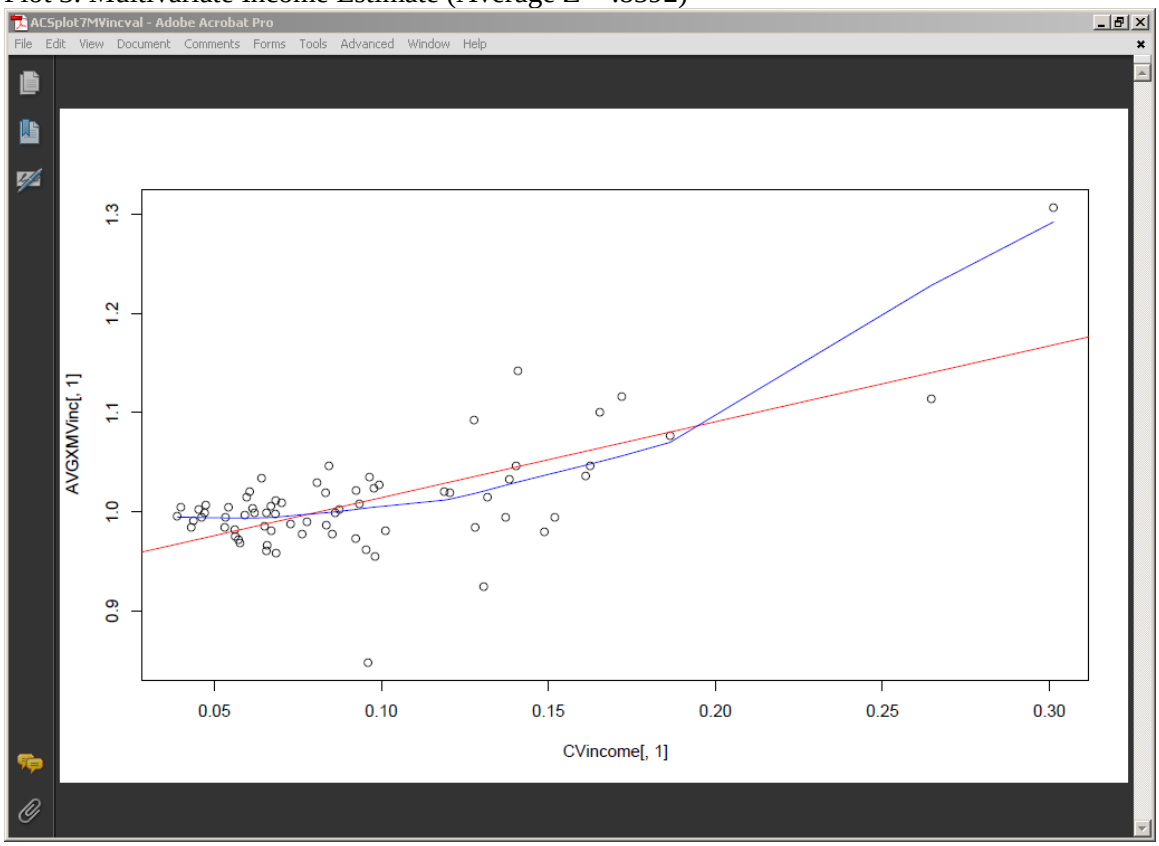
Plot 1: Fay-Herriot Income Estimate (Average Z = .8581)



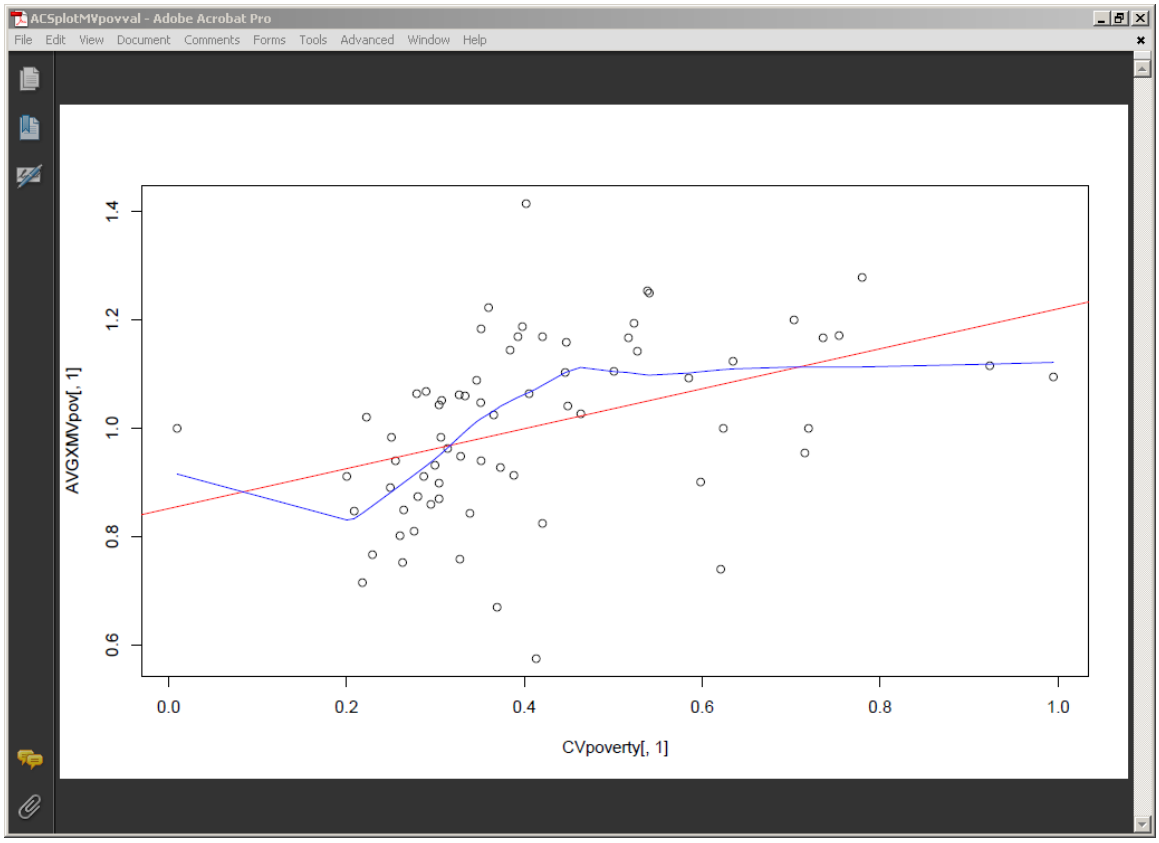
Plot 2: Fay-Herriot Poverty Estimate (Average Z = .5940)



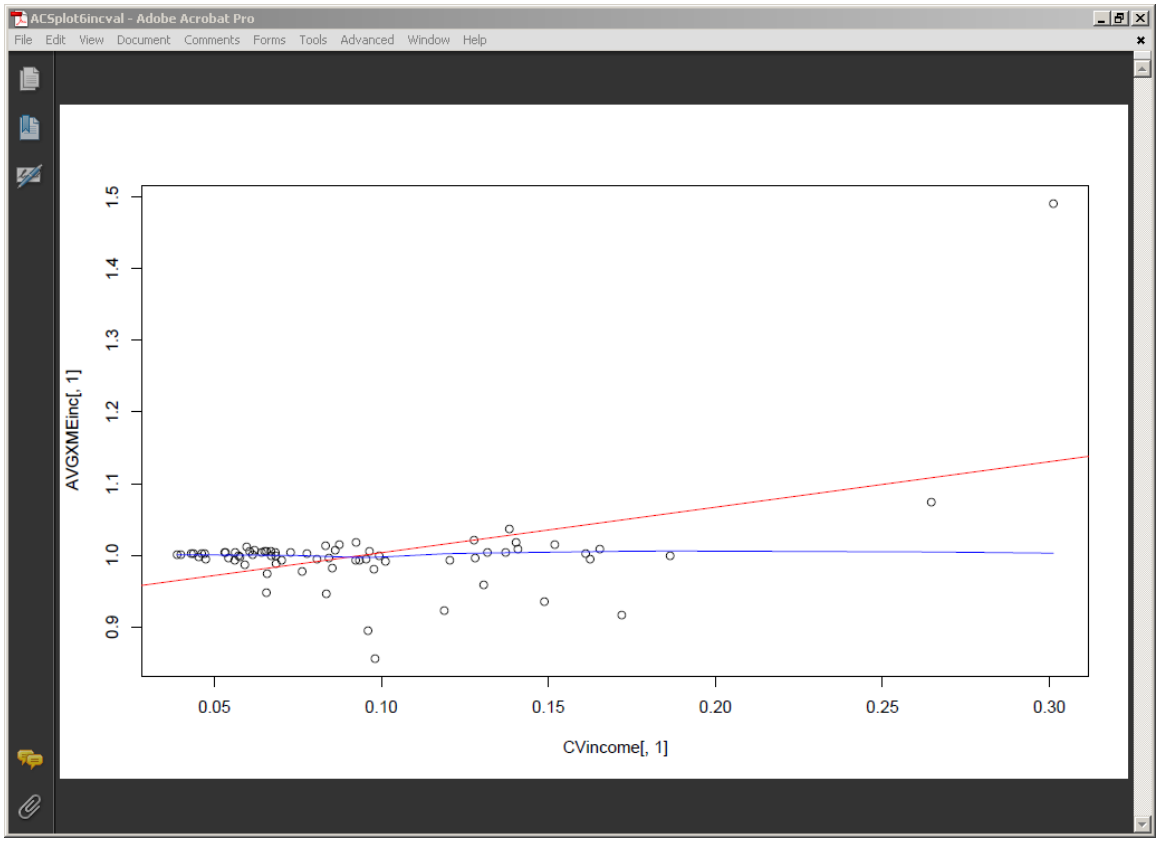
Plot 3: Multivariate Income Estimate (Average Z = .8592)



Plot 4: Multivariate Poverty Estimate (Average Z = .7565)



Plot 5: Measurement Error Income Estimate (Average Z = .9312)



Plot 6: Measurement Error Poverty Estimate (Average Z = .8207)