Using Publically Available Administrative Data to Improve Direct Estimates of Income and Poverty from the American Community Survey

Richard A. Griffin
U.S. Census Bureau

Abstract[1]

The American Community Survey (ACS) produces direct five-year estimates at the census tract level for income and poverty. Small area estimation using models that borrow strength from relationships between variables across geographic areas may improve the accuracy of these estimates. Typically, these approaches combine direct estimates with model estimates that make use of administrative data. A 2012 pilot study used simulated administrative data to demonstrate the potential gain in accuracy from using three model-based estimation methods. The Longitudinal Employee-Household Dynamics (LEHD) Origin-Destination Employment Statistics files constitute a publically available source of administrative data potentially correlated with income or poverty. This paper uses the LEHD files and published ACS data to produce tract-level estimates and estimated mean squared errors using each of these three model-based estimation methods. Results are compared with the sampling variance of the published ACS estimates, which are assumed unbiased.

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 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. A 2012 pilot study (Griffin 2012) demonstrated potential improvement in accuracy using three empirical Bayes approaches and simulated administrative data. These approaches resulted 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. Three general approaches were 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

[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.

Datta, Fay and Ghosh (1991) describe multivariate Bayes analysis in small area estimation that uses these correlated estimates 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 application. 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 applies these three approaches (Classical Fay-Herriot, multivariate regression, and measurement error) using actual publically available administrative data. The Longitudinal Employee-Household Dynamics (LEHD) Origin-Destination Employment Statistics files constitute a publically available source of administrative data potentially correlated with income or poverty. The LEHD files and published ACS data are used to produce tract-level estimates and estimated mean squared errors using each of these three model-based estimation methods. Results are compared with the sampling variance of the published ACS estimates, which are assumed unbiased.

2. Overview of Methodology

Data are used from ACS five year tract-level estimates (2006-2010) available at www.census.gov for Erie County, Pennsylvania (m = 70 tracts). Tract level estimates of poverty, income, and property value are examined. These estimates are assumed to be unbiased and the published ACS margins of error are used as measures of root mean square error for these estimates.

Two statistics are estimated using all three model-based estimation methods (Classical Fay-Herriot, multivariate regression, and measurement error): (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. In addition, the estimated median owner occupied housing unit value is presented for the Fay-Herriot and multivariate models.

2.1 Administrative Data

The Longitudinal Employee-Household Dynamics (LEHD) Origin-Destination Employment Statistics files can be downloaded at http://lehd.ces.census.gov/data/. This

website includes detailed documentation on the creation of these files. These files are compiled from administrative records data collected by a large number of states for both jobs and firms, and enhanced with information integrated from other data sources at the Census Bureau. Despite the fine geographic and industry detail, the confidentiality of the underlying micro-data is maintained by the application of new, state-of-the-art protection methods. The LEHD Origin-Destination Employment Statistics (LODES) files were used for this application. Data files are state-based and organized into three types: Origin-Destination (OD), Residence Area Characteristic (RAC), and Workplace Area Characteristics (WAC). Here only the 2010 RAC file for Erie county Pennsylvania for job type "All Jobs" is used for demonstration purposes. Similar procedures could be implemented for any state included on the LODES files.

The RAC file is a census block level file. The 15 digit Census Block code includes 6 digits for the census tract number. Using this tract code, census block totals were aggregated to create a tract-level file. The file has 43 count categories but only the following five characteristics were used for the models in this demonstration. These independent variables are as follows:
A1: Total number of Jobs
A2: Number of Jobs for workers age 29 or younger
A3: Number of Jobs for workers age 30-54
A4: Number of Jobs with earnings $1250/month or less
A5: Number of Jobs with earnings $1251/month to $3333/month

Using ordinary least squares regression with these five independent variables, an intercept term, and a dependent variable equal to the ACS 5 year (2006-2010) direct estimate of either the income, poverty, or value statistic described above, resulted in the following:

Characteristic	R^2 value	F Statistic (5 and 64 DF)	p-value
Income	.775	44.02	< 2.2 e-16
Poverty	.451	10.49	< 2.2 e-7
Value	.812	55.27	< 2.2 e-16

Note that the same independent variables were used for all three regression models and using other available independent variable from the 43 available could potentially improve these fits.

2.2 Classical Fay-Herriot Model

For each tract j, assume that the unbiased small area estimate $\hat{\theta}_j$ is related to auxiliary data $\alpha_j = (1, A_{j1}, A_{j2}, A_{j3}, A_{j4}, A_{j5})^T$ through a linear model.

$\hat{\theta}_j = \theta_j + e_j$ and $\theta_j = \alpha_j \beta + v_j$, $j = 1,...,m$ (m is the number of tracts)
where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T$ is the 6x1 vector of regression coefficients, e_j are independent $N(0, \psi_j)$, v_j are independent $N(0, \sigma_v^2)$, and e_j and v_j are independent.

An estimate of the model variance from Prasad and Rao (1990) is as follows

$$\hat{\sigma}_v^2 = \frac{1}{m-2}[\sum_{j=1}^{m}(\hat{\theta}_j - \hat{\beta}_{0,OLS} - \hat{\beta}_{1,OLS} A_{j1} - \hat{\beta}_{2,OLS} A_{j2} - \hat{\beta}_{3,OLS} A_{j3} - \hat{\beta}_{4,OLS} A_{j4} - \hat{\beta}_{5,OLS} A_{j5})^2$$

$$- \sum_{j=1}^{m} \psi_j (1 - \alpha_j^T (A^T A)^{-1} \alpha_j)] \quad (1)$$

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

$$A = \begin{pmatrix} 1 & A_{11} & A_{12} & A_{13} & A_{14} & A_{15} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & A_{m1} & A_{m2} & A_{m3} & A_{m4} & A_{m5} \end{pmatrix}.$$

The Empirical Bayes estimates are as follows

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

$$\hat{\gamma}_j = \frac{\hat{\sigma}_v^2}{\psi_j + \hat{\sigma}_v^2}$$

$$\hat{\theta}_{j,FHClassical} = \hat{\gamma}_j \hat{\theta}_j + (1 - \hat{\gamma}_j) \alpha_j^T \hat{\beta} \quad (2)$$

The mean squared error is estimated (Rao 2003 section 7.1.5) by

$$mse(\hat{\theta}_j) = g_{1j}(\hat{\sigma}_v^2) + g_{2j}(\hat{\sigma}_v^2) + 2g_{3j}(\hat{\sigma}_v^2) \qquad (3)$$

where,

$$g_{1j}(\hat{\sigma}_v^2) = \frac{\hat{\sigma}_v^2 \psi_j}{\hat{\sigma}_v^2 + \psi_j} = \gamma_j \psi_j$$

$$g_{2j}(\hat{\sigma}_v^2) = (1-\gamma_j)^2 \alpha_j^T [\sum_{j=1}^{m} \frac{\alpha_j \alpha_j^T}{\hat{\sigma}_v^2 + \psi_j}]^{-1} \alpha_j$$

$$g_{3j}(\hat{\sigma}_v^2) = \frac{\psi_j^2}{(\hat{\sigma}_v^2 + \psi_j)^3} [\frac{2}{m^2} \sum_{j=1}^{m} (\hat{\sigma}_v^2 + \psi_j)^2$$

2.5. Multivariate Model Estimation (two independent variables)

For each j, the basic data are the two-component vectors $\hat{\theta}_j = (\hat{\theta}_{j1}, \hat{\theta}_{j2})^T$ j = 1,…,m. $\hat{\theta}_{j1}$ is the estimate of interest and $\hat{\theta}_{j2}$ is believed to be strongly correlated with it. Note that either one could be considered the estimate of interest.

Let $\theta_j = (\theta_{j1}, \theta_{j2})^T$.

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

$$A_j = \begin{pmatrix} 1 & A_{j1} & A_{j2} & A_{j3} & A_{j4} & A_{j5} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & A_{j1} & A_{j2} & A_{j3} & A_{j4} & A_{j5} \end{pmatrix}$$

Note that the same independent variables are used for each of the two components.

$$\beta^T = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 & \beta_9 & \beta_{10} & \beta_{11} & \beta_{12} \end{pmatrix}$$

θ_j are independent .

$$N(\begin{pmatrix} \beta_1 + \beta_2 A_{j1} + \beta_3 A_{j2} + \beta_4 A_{j3} + \beta_5 A_{j4} + \beta_6 A_{j5} \\ \beta_7 + \beta_8 A_{j1} + \beta_9 A_{j2} + \beta_{10} A_{j3} + \beta_{11} A_{j4} + \beta_{12} A_{j5} \end{pmatrix}, \begin{pmatrix} \sigma_{v1}^2 & \sigma_{v12} \\ \sigma_{v12} & \sigma_{v2}^2 \end{pmatrix})$$

The diagonal terms of the covariance matrix for the vector θ_j are each estimated using equation (1). 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 (1) can be used twice, once for $\hat{\sigma}_{v1}^2$ and once for $\hat{\sigma}_{v2}^2$. Then use

$$\hat{\sigma}_{v12} = \frac{1}{m-2}[\sum_{j=1}^{m}(\hat{\theta}_{j1} - \hat{\beta}_{1,OLS} - \hat{\beta}_{2,OLS} A_{j1} - \hat{\beta}_{3,OLS} A_{j2} - \hat{\beta}_{4,OLS} A_{j3} - \hat{\beta}_{5,OLS} A_{j4} - \hat{\beta}_{6,OLS} A_{j5}) X$$
$$(\hat{\theta}_{j2} - \hat{\beta}_{7,OLS} - \hat{\beta}_{8,OLS} A_{j1} - \hat{\beta}_{9,OLS} A_{j2} - \hat{\beta}_{10,OLS} A_{j3} - \hat{\beta}_{11,OLS} A_{j4} - \hat{\beta}_{12,OLS} A_{j5})$$

Let $\hat{D} = \begin{pmatrix} \hat{\sigma}_{v1}^2 & \hat{\sigma}_{v12} \\ \hat{\sigma}_{v12} & \hat{\sigma}_{v2}^2 \end{pmatrix}$.

Then the Empirical Bayes multivariate estimator is given by

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

$$\hat{\theta}_{j,multi} = \hat{D}(\psi_j + \hat{D})^{-1} \hat{\theta}_j + \psi_j (\psi_j + \hat{D})^{-1} A_j \hat{\beta} = (\hat{\theta}_{j1,multi}, \hat{\theta}_{j2,multi})^T \quad (4)$$

The mean squared error is estimated (Rao 2003; 8.1.3) by

$$\hat{MSE}(\hat{\theta}_{j,multi}) = J + (J\hat{D}^{-1} A_j)(Z^{-1} A_j^T \hat{D}^{-1} J) \quad (5)$$

where, $J = (\psi_j^{-1} + \hat{D}^{-1})^{-1}$ and $Z = \sum_{j=1}^{m} A_j^T (\psi_j + \hat{D})^{-1} A_j$

This is a 2x2 matrix for each tract j. The elements 1,1 and 2,2 on the diagonal provide the mean squared error estimates for the two components.

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

Either $\hat{\theta}_{j1}$ or $\hat{\theta}_{j2}$ can be the independent variable with measurement error (i.e., sampling error). Here $\hat{\theta}_{j2}$ is the independent variable with sampling error.

$$X_j = \begin{pmatrix} 1 \\ \theta_{j2} \end{pmatrix} \quad \hat{X}_j = \begin{pmatrix} 1 \\ \hat{\theta}_{j2} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\hat{\theta}_{j1} = \hat{X}_j^T \beta + (X_j - \hat{X}_j)^T \beta + v_j + e_j$$

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

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

Let $r_j = (X_j - \hat{X}_j)^T \beta + v_j$

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

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

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

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

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

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

Since β and σ_v^2 are unknown, first let $\hat{\beta}^{(1)} = \left(\sum_{j=1}^{m} \hat{X}_j \hat{X}_j^T - C_j \right)^{-1} \left(\sum_{j=1}^{m} \hat{X}_j \hat{\theta}_{j1} - C_{j\hat{X}\hat{\theta}} \right)$

Then $\hat{\sigma}_v^2 = m^{-1} \sum_{j=1}^{m} (\hat{\theta}_j - \hat{X}_j^T \hat{\beta}^{(1)})^2 - \hat{\psi}_j - \hat{\beta}_1^{(1)} \psi_{j2} + 2\hat{\beta}_1^{(1)} \psi_{j12}$

Then compute $\hat{\beta}^{(2)}$ using the weights $w_j = \dfrac{1}{\hat{\sigma}_v^2 + (\hat{\beta}_1^{(1)})^2 \psi_{j2} - 2\hat{\beta}_1^{(1)} \psi_{j12} + \psi_{j1}}$.

$$\hat{\beta}^{(2)} = \left(\sum_{j=1}^{m} w_j (\hat{X}_j \hat{X}_j^T - C_j) \right)^{-1} \left(\sum_{j=1}^{m} w_j (\hat{X}_j \hat{\theta}_j - C_{j\hat{X}\hat{\theta}}) \right)$$

Then $\hat{\theta}_{j,measerror} = \hat{\gamma}_j \hat{\theta}_j + (1-\hat{\gamma}_j)\hat{X}_j^T \hat{\beta}^{(2)}$ (6)

where, $\hat{\gamma}_j = \dfrac{\hat{\sigma}_v^2 + (\hat{\beta}_1^{(2)})^2 \psi_{j2} - \hat{\beta}_1^{(2)} \psi_{j12}}{\hat{\sigma}_v^2 + (\hat{\beta}_1^{(2)})^2 \psi_{j2} - 2\hat{\beta}_1^{(2)} \psi_{j12} + \psi_{j1}}$.

The mean squared error is estimated using the jackknife variance estimator suggested by Ybarra and Lohr (2008).

(k) indicates estimation leaving out tract k.

$$M1_{j(k)} = \gamma_j \psi_j - \gamma_{j(k)} \psi_j$$

$$M2_{j(k)} = (\hat{\theta}_{j(k),measerror} - \hat{\theta}_{j,measerror})^2$$

$$E_{j1} = \gamma_j \psi_j + \frac{m-1}{m} \sum_{k=1}^{m} M1_{j(k)}$$

$$E_{j2} = \frac{m-1}{m}\sum_{k=1}^{m} M2_{j(k)}$$

$$\hat{MSE}(\hat{\theta}_{j,measerror}) = E_{j1} + E_{j2} \qquad (7)$$

3. Results of Estimation

First, direct estimates of each of the three statistics for each of the 70 tracts ($\hat{\theta}_j$) and their standard errors ($\sqrt{\psi_j}$) were obtained from the American Community Survey data provided at www.census.gov. The average coefficient of variation (CV) was .094 for the direct income estimate, .414 for the direct poverty estimate, and .061 for the direct value estimate.

3.1 Classical Fay-Herriot Estimator

For each of the three statistics for each of the 70 tracts, $\hat{\theta}_{j,FHclassical}$ (equation (2)) and $mse(\hat{\theta}_{j,FHclassical})$ (equation (3)) were calculated.

Table 1 provides results averaged over the 70 tracts. Column (1) is the average $\hat{\gamma}_j$ value which is the weight given to the direct estimate in $\hat{\theta}_{j,FHclassical}$. Column (2) is the average ratio of the root mean square error (MSE) to the standard error (SE) of the direct estimate. Columns (3), (4), and (5) show the average proportion of the estimated mean squared error $mse(\hat{\theta}_{j,FHclassical})$ that comes g_{1j}, g_{2j}, and $2g_{3j}$ respectively.

Table 1 Average Results over 70 Tracts for Fay-Herriot Estimator

Statistic	Weight for Direct Estimator	Root MSE Fay-Herriot/ SE Direct	Proportion of MSE from		
	(1)	(2)	g_{1j} (3)	g_{2j} (4)	$2g_{3j}$ (5)
Average Family Income	.767	.890	.957	.026	.017
# Families in Poverty	.661	.833	.926	.042	.032
Median Owner Occupied Value	.894	.951	.984	.010	.006

The weight applied to the direct estimator is less when the direct estimate has more sampling variance. The column (1) values are smallest for the poverty estimate and the direct poverty estimate has a much higher CV than the direct income or direct value estimates. The reduction in root MSE using the Fay-Herriot estimator is greater when less weight is applied to the direct estimate. The reduction in average root MSE (the direct estimate is assumed unbiased) is about 17% for poverty, 11% for income, and 5% for value.

For MSE estimation, columns (3), (4), and (5) indicate that over 92% of the total MSE estimate comes from the g_1 term which is the sampling variance of the direct estimator multiplied by its weight.

3.2 Multivariate Regression Estimator

Three multivariate regression (MVR) estimators and their MSE estimates were calculated for each of the 70 tracts using equations (4) and (5). The three multivariate estimators were (1) using the direct income and poverty estimates as dependent variables; (2) using the direct income and value estimates as dependent variables; and (3) using the direct poverty and value estimates as dependent variables. The independent variables were the same for all estimates as described in section 2.5.

Table 2 provides the average ratio of the root mean squared error of the MVR estimator to the standard error of the direct estimate for each of the two components of the three MVR applications.

Table 2 Average Ratio of Root MSE of MVR estimator to SE of Direct Estimator

MVR Estimator	Components	Root MSE MVR/ SE Direct
(1)	Average Family Income	.863
	# Families in Poverty	.804
(2)	Average Family Income	.872
	Median Owner Occupied Value	.897
(3)	# Families in Poverty	.805
	Median Owner Occupied Value	.910

Comparing these ratios to the ratio for the Fay-Herriot estimator in Table 1 demonstrates the improvement over univariate regression by using multivariate regression.

The 11% reduction in root MSE for the income statistic was improved to about a 14% reduction using poverty as an additional dependent variable and to about a 13% reduction using value as an additional dependent variable.

The 17% reduction in root MSE for the poverty statistic was improved to about a 20% reduction using income or value as an additional dependent variable.

The 5% reduction in root MSE for the value statistic was improved to about a 10% reduction using income as an additional dependent variable and to about a 9% reduction using poverty as an additional dependent variable.

3.3 Measurement Error Model Estimator

For average family income the measurement error (ME) model estimator was used three times (3 separate estimation models) with (1) number of families in poverty, (2) median owner occupied housing unit value, and (3) 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 separate ME estimator models were also done for number of families in poverty using (1) average family income, (2) median owner occupied housing unit value, and (3) 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 average ratio of the root mean squared error of the ME estimator to the standard error of the direct estimate for each of the these six ME estimators. Equation (6) was used for the ME estimates and equation (7) for MSE estimation.

Table 3 Average Ratio of Root MSE of ME estimator to SE of Direct Estimator

Statistic Estimated	Independent Variable (with sampling error)	Root MSE ME/ SE Direct
Average Family Income	# Families in Poverty	.992
Average Family Income	Median Owner Occupied Value	1.043
Average Family Income	Female HH, child<18; no husband	.954
# Families in Poverty	Average Family Income	.864
# Families in Poverty	Median Owner Occupied Value	.987
# Families in Poverty	Female HH, child<18; no husband	.991

Comparing these ratios to the ratio for the Fay-Herriot estimator in Table 1 demonstrates that the reductions in root MSE using the measurement error model are not as large as the reductions from the Fay-Herriot model.

The average reduction in MSE for income for Fay-Herriot was about 11% and the best reduction ME estimator, with Female Head of Household with Children under 18 and No Husband as the independent variable, was about 5%. Using value as the independent variable resulted in an average increase in root MSE.

The average reduction in MSE for poverty for Fay-Herriot was about 17% and the best reduction ME estimator, with income as the independent variable, was about 14%.

3.4 Plots

All the results presented in Tables 1, 2, and 3 are averages over the 70 tracts in Erie county, PA. Also of interest is the distribution of these estimates across the 70 tracts.

Four graphs showing these distributions for each of the 15 estimators evaluated were generated.
For a longer version of this paper, these tables were designated as Table 4 through Table 18.
They all looked similar to Table 4 below for the Fay-Herriot Estimate of Average Family Income except for Table 14, also shown below, for the Measurement Error Estimate of Average Family Income (with Value as dependent variable).

Each table has four graphs.

- Upper Left Hand Corner – Plot of the direct estimate as a function of the model-based estimate. A least squares regression line (red) and locally weighted polynomial regression line (blue) are shown in each plot.
- Upper Right Hand Corner – The Density function of the direct estimate (includes mean and standard deviation)
- Lower Left Hand Corner – The Density function of the model-based estimate (includes mean and standard error deviation)
- Lower Right Hand Corner – The Density function of the ratio of the root mean squared error of the model-based estimate to the standard error of the assumed unbiased direct estimate.

These graphs show that the distribution of the model-based estimates over the 70 tracts are similar to the distributions of the direct estimate of the same statistic. The standard deviation among tracts is less, as expected, making for a similar shape with less spread. The density of the ratio of the root mean squared error of the model-based estimate to the standard error of the assumed unbiased direct estimate illustrates that the probability of a ratio greater than 1 is small for all estimators except for the Measurement Error Estimate of Average Family Income (with Value as dependent variable, Table 14).

4. Summary

For the 70 tracts in Erie County, PA, the Fay-Herriot estimator provided an average reduction in the ratio of root mean square error to standard error of the direct estimate of about 11% for Average Family Income, 17% for Number of Families in Poverty, and 5% for Median Owner Occupied Value. The Multivariate Regression estimator increased this reduction to about 13% for the income statistic, 20% for the poverty statistic and 10% for the value statistic. All these estimates used the same set of tract level independent variables (number of jobs by age of worker and number of jobs by monthly earnings) from the publically available Longitudinal Employee-Household Dynamics (LEHD) Origin-Destination Employment Statistics files. The best reductions in this ratio using the Measurement Error estimator were a 5% reduction for the income statistic using the number of female head of household families with children under age 18 and no husband statistic as the independent variable and a 14% reduction for the poverty statistic using the income statistic as the independent variable. The Value statistic was not estimated using the Measurement Error estimator.

As expected, the results demonstrate the direct estimates with higher coefficients of variation are prime candidates for borrowing strength via model-based estimation to reduce mean squared error.

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.

Griffin, R. (2012), "Pilot Study on Combining Direct Estimates of Income and Poverty from the American Community Survey with Predictions from a Model", *Proceedings of the Section on Survey Reseach Methods, American Statistical Association.*

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 Area Estimation when Auxiliary Information is Measured with Error," *Biometrika*, 95, 4, 919-931.

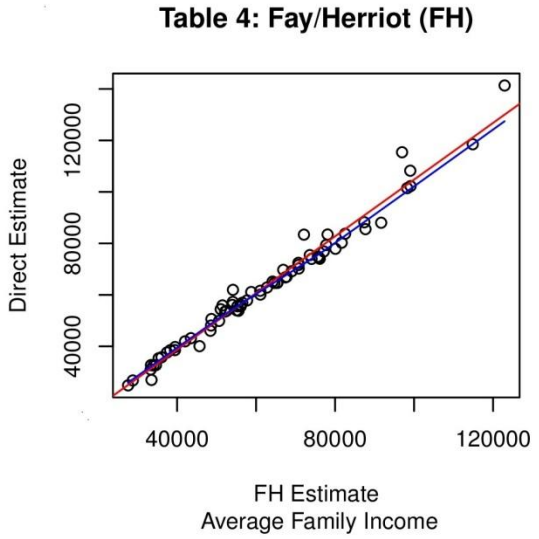
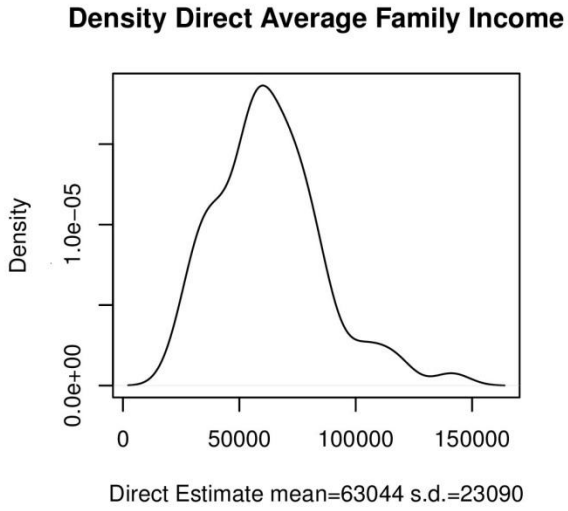
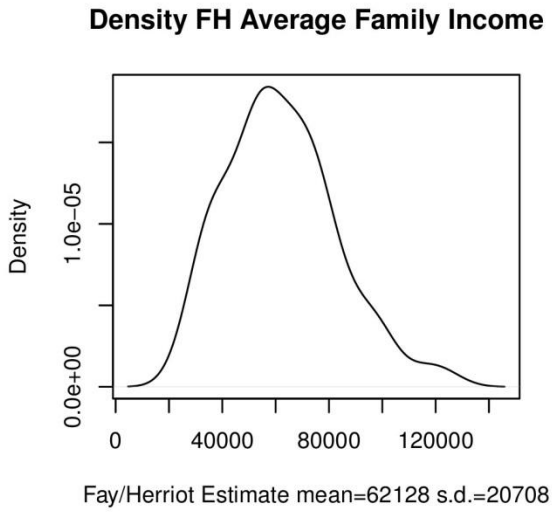


Table 4: Fay/Herriot (FH)
Direct Estimate
40000
80000
120000
40000
80000
120000
FH Estimate
Average Family Income


Density Direct Average Family Income
Density
0.0e+00
1.0e−05
0
50000
100000
150000
Direct Estimate mean=63044 s.d.=23090


Density FH Average Family Income
Density
0.0e+00
1.0e−05
0
40000
80000
120000
Fay/Herriot Estimate mean=62128 s.d.=20708

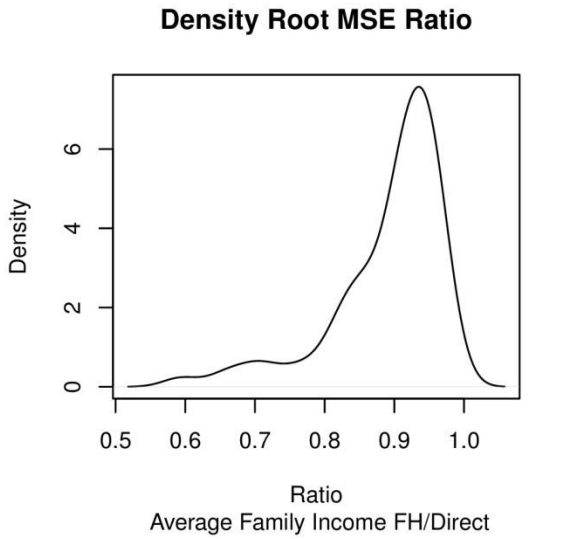
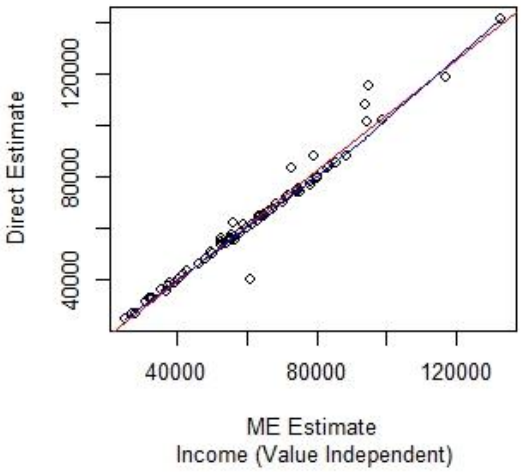
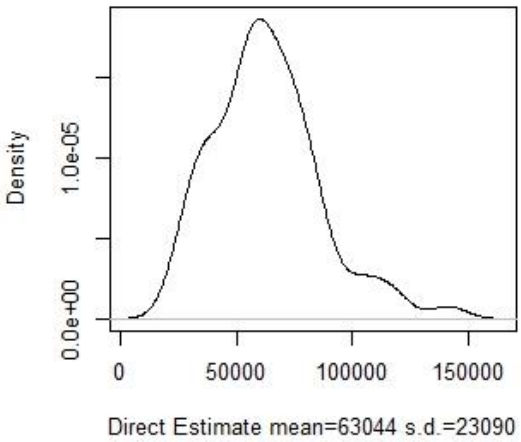

Density Root MSE Ratio
Density
0
2
4
6
0.5
0.6
0.7
0.8
0.9
1.0
Ratio
Average Family Income FH/Direct

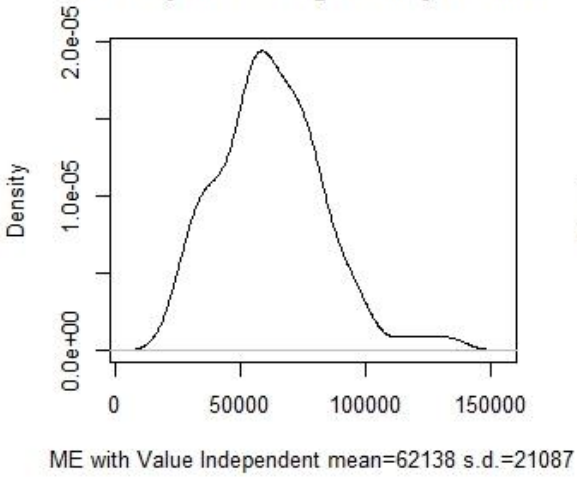
Table 14: Measurement Error (ME)



Density Direct Average Family Income



Density ME Average Family Income



Density Root MSE Ratio