# Research and Development for Methods of Estimating Poverty for School-Age Children*

Ciara Nugent Sam Hawala†

**Abstract**

The American Community Survey (ACS) does not publish single year estimates for areas with a population of less than 65,000. The majority of school districts in the United States contain fewer than 2,000 relevant school-age children. As a result, direct survey estimates for school districts are often highly variable and unreliable due to small sample sizes. This paper looks at two different approaches to modeling that address the problem of small samples in small areas. These small sample sizes can also cause a high frequency of zeros estimates. First, we look at a reweighting method proposed by Schirm and Zaslavsky (1997). Secondly, we look at a censored Fay-Herriot model, similar to that in Slud and Maiti (2011). This model takes up the specific case of no observations of poverty in the ACS sample, resulting in a zero estimate: a problem for over 25% of school districts.

**Key Words:** Survey Weights, Small Area, Censoring, Reweighting, SAIPE

## 1. Introduction

Direct survey estimates for small areas are often unreliable. To produce more reliable estimates, various modeling techniques have been designed and implemented. Models can borrow strength from other data sources, different time periods, and different geographic areas. While modeling seeks to improve inferences from data, it is still dependent on the data it uses. If there are problems with the input data, the model should be modified accordingly, or a new model proposed. The Small Area Income and Poverty Estimates (SAIPE) program at the U.S. Census Bureau uses modeling to produce various estimates of and relating to poverty, including the number of children 5-17 in families in poverty at the state and county level. The SAIPE program also publishes estimates for school districts; however these estimates are not directly modeled. There are several known problems with the input data to the SAIPE model when implementing it at the school district level. First, some of the auxiliary administrative data are unavailable, such as the number of participants in the Supplemental Nutrition Assistance Program (SNAP). Other auxiliary data cannot be given a Census geographic identification, or be geocoded to a specific Census block, preventing any linking to data collected by the Census Bureau. In particular, tax data are not 100% geocoded to school districts, and have very poor geocoding rates in certain areas. Another problem comes from the reliability of the direct survey estimates themselves. Estimates of poverty in school districts are often based on a sample of less than five children, and over 25% of school districts have zero poor children in sample, resulting in a zero estimate. See

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

†Social, Economic and Housing Statistics Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Table 1 for a summary of the range of ACS sample sizes by population count, as determined by the Population Estimates Program at the U.S. Census Bureau.

**Table 1**: Table of ACS Sample Sizes of Related Children 5-17 Binned by Population of Children 0-17

| | Pop $< 1000$ | | Pop 1000-5000 | | Pop $> 5000$ | |
|---|---|---|---|---|---|---|
| | ACS Sample | ACS Sample in Poverty | ACS Sample | ACS Sample in Poverty | ACS Sample | ACS Sample in Poverty |
| Minimum | 0 | 0 | 1 | 0 | 28 | 0 |
| Median | 9 | 1 | 33 | 4 | 122 | 18 |
| Maximum | 70 | 23 | 191 | 57 | 14249 | 4022 |

While research is being done into finding solutions for all of these problems, this paper focuses on the problem of small sample sizes. One possible solution is to try to augment the sample by drawing in information from other geographic areas. This is the approach taken in reweighting. In the reweighting process, you take the sample size in a larger aggregate, such as a group of school districts, and use the entire aggregate to produce estimates for each school district. Reweighting is presented in Section 2. Another approach is to modify the current SAIPE county model to include the auxiliary information that is available from the school districts that have zero children in poverty in sample. This is the approach of the censored model, presented in Section 3. Both of these methods try to address the problems posed by the data.

## 2. Description of Reweighting Effort

Since most surveys only sample a fraction of the population, each observation is assigned a survey weight so that estimates of the whole population can be derived from the sample. Schirm and Zaslavsky (1997) propose a method of reweighting whereby these survey weights are reapportioned so that estimates for one geographic area can borrow strength across several geographic areas. Consider a group of $i$ areas. In the reweighting process, area specific weights are calculated for each observation in the group resulting in a set of $i$ new weights for every observation, one for every area in the group. Area weights are calculated using a Poisson regression fitted to determine the prevalence of each type of observation in that area. An observation type is defined by a set of demographic characteristics. The new area weights are then used to calculate estimates for an area using all of the observations in the group. For further detail on the reweighting algorithm, please see Schirm and Zaslavsky. Because school districts tend to have small sample sizes, using reweighting to augment each school district sample seems like a reasonable approach to improving poverty estimates. The method was ultimately found to be unsuitable for use on ACS unit level data, however the failures of this approach provide useful information on when reweighting is appropriate, and the nature of the ACS data itself.

One of the first questions to address when constructing the reweighting algorithm is how to group the areas together. The method of grouping has a significant impact on whether or not the algorithm converges. Areas can be grouped by geography, population, or any other measurable characteristic. Initially, school district pieces, or the portions of school districts residing within county boundaries, were grouped together by county. However, the number of school district pieces in a county is highly variable, ranging from counties with only one school district piece to counties with more than 20 school district pieces. It is undesirable to have too many or too few school district pieces in a county. If there are only two or three school district pieces in a group, there are a limited number of observations from which to draw strength. As an extreme example, if there is only one school district piece, there are no other pieces to draw strength from, and all that is left is the original survey estimate. On the other extreme, if there were too many school district pieces, the algorithm seems to have trouble converging. The lack of convergence could be due to the number of school district pieces involved, or to the characteristics of the school districts being grouped together, which can be highly variable both in terms of sample size and demographic composition. Another problem with grouping by counties is that many school districts cross county lines. This means that school districts must be evaluated as school district pieces, rather than as complete school districts. Also, because one school district can have pieces in several different counties, if any one of those counties fails to converge, the estimate for the entire school district is missing. As an alternative to grouping by counties, school districts were grouped by population count within states. This allows one to control the number of school districts in each group, and groups together areas with similar sample sizes. Additionally, by removing the restriction of county lines, one can consider all school districts as whole entities, rather than as pieces. Moving away from county groupings drastically increased the number of school districts in converging groupings.

Failure to converge was an overall problem with implementing the algorithm. The algorithm can only be used if it converges, or if it looks like it is heading towards convergence. This does not always happen. One main reason for failure to converge was outliers in the survey weights. The ACS survey weights are highly variable. For the sample that was considered, related children 5-17, the survey weights range from 1 to 925, with 90% of the survey weights falling between 15 and 185 and 99% of the survey weights falling between 8 and 270. Often, when a group of school districts did not converge, a look at the unit level data would reveal an observation with a survey weight of 200 or more. These high survey weights seem to skew the results of the reweighting process, placing a disproportionate amount of emphasis on these observations. Adding to this problem is the fact that survey weights for those in poverty are higher, on average, than survey weights for those not in poverty. For related children 5-17, about 3.6% of the survey weights are greater than 200. If you look at related children 5-17 in poverty, nearly 5% of the survey weights are greater than 200. So, the population of interest, those in poverty, is more likely to cause a failure of convergence in the algorithm.

Another cause for failure to converge is areas that do not have anyone with a certain demographic in sample. Looking at Race/ Ethnicity as the demographic, there are numerous school districts that will show an ACS estimate of 0, while the 2010 Census may show a population of 100 or more. These areas cause a problem because survey weights are redistributed according to the demographic traits of the observations to produce area weights in the reweighting process. If an entire demographic entity is missing from area A, observations with that characteristic in other areas cause problems during the calculation of area A weights because the survey weights of those observations have nowhere to go. This makes it harder for the algorithm to converge.

Due to the problems encountered with extreme weights and missing demographic groups, calibration was explored as a possible solution. The ACS weights are a result of selection probability, adjustment for nonresponse, and further adjustments to calibrate to various control totals, see U.S. Census Bureau (2011) for more details. However, adjustments to control totals are not made at the school district level, nor are they necessarily made across quantities of interest. There are many different ways to perform calibration, and a comparison of different methods is an entire paper onto itself. The method proposed here seemed the most straightforward and easy to implement given the information that was available. Calibrations were made at the school district level for various demographic counts available from the 2010 Census. Counts available from the 2010 Census include age of the householder (35 or older, and younger than 35), presence of children under five, number of related children, race/ethnicity groups (White, Black, Hispanic, and other), age groups (5-13, 14-17), and tenure. Race/Ethnicity was chosen as the means of calibration due to the correlation observed between Race/Ethnicity and Poverty. To handle school districts that have zero estimates of a given demographic in the ACS, but positive counts in the 2010 Census, the following method was employed. From the population of related children 5-17 in the 2010 Census, the counts of Whites, Blacks, Hispanics, and Others were tabulated for each school district. The weights from the ACS were calibrated to these totals. If a school district has an ACS count of zero for a particular Race/Ethnicity group, an additional entry was created for that school district. This new entry has a weight equal to the Census count of the missing race ethnicity group. To calculate the poverty for the new entry, you would calculate the percentage of those in poverty for that race/ ethnicity group across the entire state in which the school district resides. In this way the statewide percentage by race/ethnicity group stands in for the missing data.

To test the reweighting and calibration methods, the 2010 Census counts were used as a barometer. Because poverty counts are no longer available from the Census, the algorithm was altered to estimate tenure, specifically the number of children aged 5-17 who live in homes owned by their family. Comparisons were made between the ACS direct survey estimates, estimates derived from reweighting alone, estimates derived from calibration alone, and estimates derived from both calibration and reweighting. The results were inconclusive. Some areas came closer to the 2010 Census counts using the reweighting weights, and others showed no improvement over the direct estimates by any alternate method. There was no clear indication of when the reweighting estimates were likely to be better. Future research could look into different means of calibration. One possibility is to calibrate across finer classifications, such as Race $\times$ Sex. Any further research will provide an even deeper understanding of the ACS data and the ACS weighting process.

# 3. A Censored Fay-Herriot Model

## 3.1 Background

The current SAIPE county model uses a Fay-Herriot Model of the log count of children 5-17 in families in poverty. The model combines ACS 1-year direct estimates with additional administrative and Census data in a regression. The synthetic estimate is then combined with the direct survey estimate in a shrinkage equation. Details on the current SAIPE county model can be found on the SAIPE website (U.S. Census Bureau 2010). Areas with no poor children in sample are excluded from analysis in the SAIPE model, because counts of zero result in an undefined log count, which cannot be analyzed in a regression or used as direct estimates in a shrinkage equation. None of the information from these areas, including the available administrative and Census data, is used to determine the parameters in the modeling process. This is not as big of a problem at the county level where a little less than 200 out of 3142 counties, or about 6%, have ACS direct estimates of zero every year. However, at the school district level the number of areas with zero poor in the ACS sample jumps to more than 25%. These areas are not randomly dispersed throughout the sample. Areas with ACS direct estimates of zero have smaller population counts, on average, than areas that do not. Additionally, there exists a negative correlation between sampling variance and sample size. Areas with a sample size of zero would be expected to have a higher sampling variance, on average, than areas with larger sample sizes. A better model should result from including the information available for ACS direct estimates of zero.

Slud and Maiti (2011) have proposed a left-censored Fay-Herriot model as an alternative to the current SAIPE model. While the model described below does not exactly follow the model of Slud and Maiti, it does use a censoring framework to incorporate the information available from observations with an ACS estimate of zero.

### 3.2 Model

This model starts with the current SAIPE county model:

$$y_i = Y_i + e_i$$

$$Y_i = \mathbf{x_i}'\boldsymbol{\beta} + u_i$$

where:

- $y_i =$ log(ACS direct estimate of the number of related school-age children in poverty in county $i$)
- $Y_i =$ log(true number of related school-age children in poverty in county $i$), which is unknown
- $\mathbf{x_i} = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})'$, a vector of known explanatory variables for county $i$
  - $x_{i1} =$ log(Census 2000 number of poor school-age children in county $i$)
  - $x_{i2} =$ log(number of participants in SNAP in county $i$)
  - $x_{i3} =$ log(estimated population 0-17 in county $i$)
  - $x_{i4} =$ log(number of child tax exemptions reported by families in poverty in county $i$)
  - $x_{i5} =$ log(number of child tax exemptions reported in county $i$)
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$, a vector of regression coefficients to be estimated
- $u_i =$ the unobservable random effect for county $i$, and $u_i \overset{iid}{\sim} N(0, {\sigma_u}^2)$ where $\sigma_u^2$ is the model error variance
- $e_i =$ the sampling error for county $i$, and $e_i \overset{ind}{\sim} N(0, {\sigma_{e_i}}^2)$ where ${\sigma_{e_i}}^2$ is the sampling variance for county $i$, the sampling variance is assumed known.

The left-censored model will require further notation, which we outline here:

- $\delta_i =$ a categorical variable that is 1 if the observation is censored, and 0 if it is not

$$\delta_i = \begin{cases} 1 & \text{if the ACS direct estimate is zero} \\ 0 & \text{otherwise} \end{cases}$$

- $k =$ a threshold such that if $Y_i$ is less than $k$, there is a probability that the ACS will not record a measurement

Now consider the distribution of $y_i$ given $Y_i$ in relation to the censoring threshold:

$$\begin{array}{llll} \text{If} \;\; Y_i \geq k, & y_i = Y_i + e_i & \text{with probability} & 1 \\ \text{If} \;\; Y_i < k, & y_i \text{ is censored} & \text{with probability} & p\left(\delta_i = 1 | Y_i < k\right) \\ & y_i = Y_i + e_i & \text{with probability} & 1 - p\left(\delta_i = 1 | Y_i < k\right) \end{array}$$

*3.2.1 Likelihood and Posterior Distributions for Censored Model*

Take the prior distribution to be independent and diffuse:

$$p\left(\boldsymbol{\beta}, \sigma_u^2\right) \propto c$$

where $c$ is a constant.

The posterior distributions are as follows:

$$\pi\left(\beta|\mathbf{Y}, \sigma_u^2\right) \propto \prod_{i=1}^{m} \left\{ \exp\Big( -\frac{(Y_i - \mathbf{x_i}'\boldsymbol{\beta})^2}{2(\sigma_u^2)} \Big) \right\}$$

$$\pi\left(\mathbf{Y}|\mathbf{y}, \boldsymbol{\sigma_e^2}, \boldsymbol{\beta}, \sigma_u^2, \boldsymbol{\delta}, k\right) \propto \prod_{i=1}^{m} \Bigg[ \exp\Big( -\tfrac{(Y_i - \mathbf{x_i}'\boldsymbol{\beta})^2}{2(\sigma_u^2)} \Big) \Big\{ \tfrac{1}{\sigma_{e_i}} \exp\Big( -\tfrac{(y_i - Y_i)^2}{2({\sigma_{e_i}}^2)} \Big) \Big\}^{I[Y_i \geq k]} \left\{ \left\{ \tfrac{1 - p\left(\delta_i = 1 | Y_i < k\right)}{\sigma_{e_i}} \exp\Big( -\tfrac{(y_i - Y_i)^2}{2({\sigma_{e_i}}^2)} \Big) \right\}^{I[\delta_i = 0]} \left\{ p(\delta_i = 1 | Y_i < k) \right\}^{I[\delta_i = 1]} \right\}^{I[Y_i < k]} \Bigg]$$

$$\pi\left(\sigma_u^2|\mathbf{Y}, \boldsymbol{\beta}\right) \propto \prod_{i=1}^{m} \left\{ \frac{1}{\sigma_u} \exp\Big( -\frac{\left(Y_i - \mathbf{x_i}'\boldsymbol{\beta}\right)^2}{2(\sigma_u^2)} \Big) \right\}$$

where $\Phi$ is the cumulative distribution function of the standard normal density.

*3.2.2 Model Computation*

Using the posterior likelihoods defined in the previous section, the regression coefficients were estimated using a Gibbs Sampler. The $Y_i$ and $\sigma_u^2$ were estimated using a random walk Metropolis algorithm. Each chain was run for 150,000 iterations. The first 100,000 iterations were discarded, and then every tenth iteration was kept, resulting in 5,000 draws for each chain. Six chains were run in parallel, resulting in 30,000 draws from the posterior distribution. For each of the variables, calculations of Gelman's $R$, see Gelman et. al. (2004), were below 1.1. All of these calculations were performed using the R Language, R Development Core Team (2010).

## 3.3 Simulation

As a first means of testing the censored model, a simulation was used. There were two main motivations for using a simulation. First, in a simulation the true values and survey estimates are not merely assumed to be normally distributed, they are known to be normally distributed. Second, the true values for all of the variables are known, so comparisons can be made to the truth. The data were simulated once, and a censored Fay-Herriot and an uncensored Fay-Herriot were both run on the simulated data. Comparisons were made based on which model came closer to estimating the true values of the variables.

#### *3.3.1 Approach to Simulation*

For the simulation, data were taken from the 2010 SAIPE county data. Since the model is intended for school districts, the data were subset to include only those counties with a log count in poverty less than or equal to 10. From this subset, 150 counties were chosen at random. These 150 counties constitute the sample. The sample variance and explanatory variables were kept from the observations. The true value, $Y_i$, was generated according to $Y_i \stackrel{iid}{\sim} N(\mathbf{x_i}'\boldsymbol{\beta}, \boldsymbol{\sigma_u}^{\mathbf{2}})$ with $\boldsymbol{\beta} = (-0.45, 0.16, 0.12, 0.8, 0.75, -0.82)'$ and $\sigma_u{}^2 = 0.02$, a constant. The survey estimate $y_i$ was then generated according to $y_i \stackrel{ind}{\sim} N(Y_i, \sigma_{e_i}{}^2)$ with $\sigma^2_{e_i}$ set equal to the corresponding sampling variance. These values were chosen because of their similarity to the coefficients of the SAIPE county model. Once all of these values were set, observations with a simulated $Y_i < 6$ were censored with a probability of $0.6$. The censoring threshold and percentage were chosen ad hoc to give a reasonable overall censoring rate while not censoring 100% of values below the threshold. The data set was simulated once. Out of 150 observations 21 were censored, for a 14% censoring rate. The range of $\mathbf{Y}$ was 3.083 to 10.552 and the range of $\mathbf{y}$ was 0 to 10.456.

#### *3.3.2 Results of Simulation*

The results from Table 2 show that while the estimates of the coefficients and the model variance differ for the uncensored model and the censored model, neither truly outperforms the other in terms of reproducing the true values. The average absolute error for the parameters in Table 2 for the uncensored model is 0.204, and the average absolute error for the censored model is 0.208.

**Table 2**: Table of the True and Estimated Values of $\boldsymbol{\beta}$ and $\sigma_u{}^2$

| | True Value | Uncensored Average | Uncensored Absolute Error | Censored Average | Censored Absolute Error |
|---|---|---|---|---|---|
| $\beta_0$ | -0.450 | -0.875 | 0.425 | -0.845 | 0.395 |
| $\beta_1$ | 0.160 | 0.135 | 0.025 | 0.202 | 0.042 |
| $\beta_2$ | 0.120 | 0.144 | 0.024 | 0.137 | 0.017 |
| $\beta_3$ | 0.800 | 1.289 | 0.489 | 1.283 | 0.483 |
| $\beta_4$ | 0.750 | 0.698 | 0.052 | 0.595 | 0.155 |
| $\beta_5$ | -0.820 | -1.221 | 0.401 | -1.175 | 0.355 |
| $\sigma_u{}^2$ | 0.020 | 0.032 | 0.012 | 0.029 | 0.009 |

If you look at the model estimates of $Y_i$, the censored model outperforms the uncensored model. To estimate $Y_i$ for the censored values from the uncensored model, the only information available is the synthetic estimate, which is what SAIPE uses for such observations in their model. For each $\boldsymbol{\beta}$ generated from the posterior distribution, $\mathbf{x_i}'\boldsymbol{\beta}$ was taken as the estimate of $Y_i$ if the value was censored. Table 3 and Table 4 show how often the model captures the true value in 50%, 75%, and 90% confidence intervals. The confidence intervals are based on the 30,000 posterior draws for each $Y_i$. For example, if at least 5% of the posterior draws are greater than the true value, and at least 5% are less than the true value, then the true value is considered to be in the 90% confidence interval.

**Figure 1**:

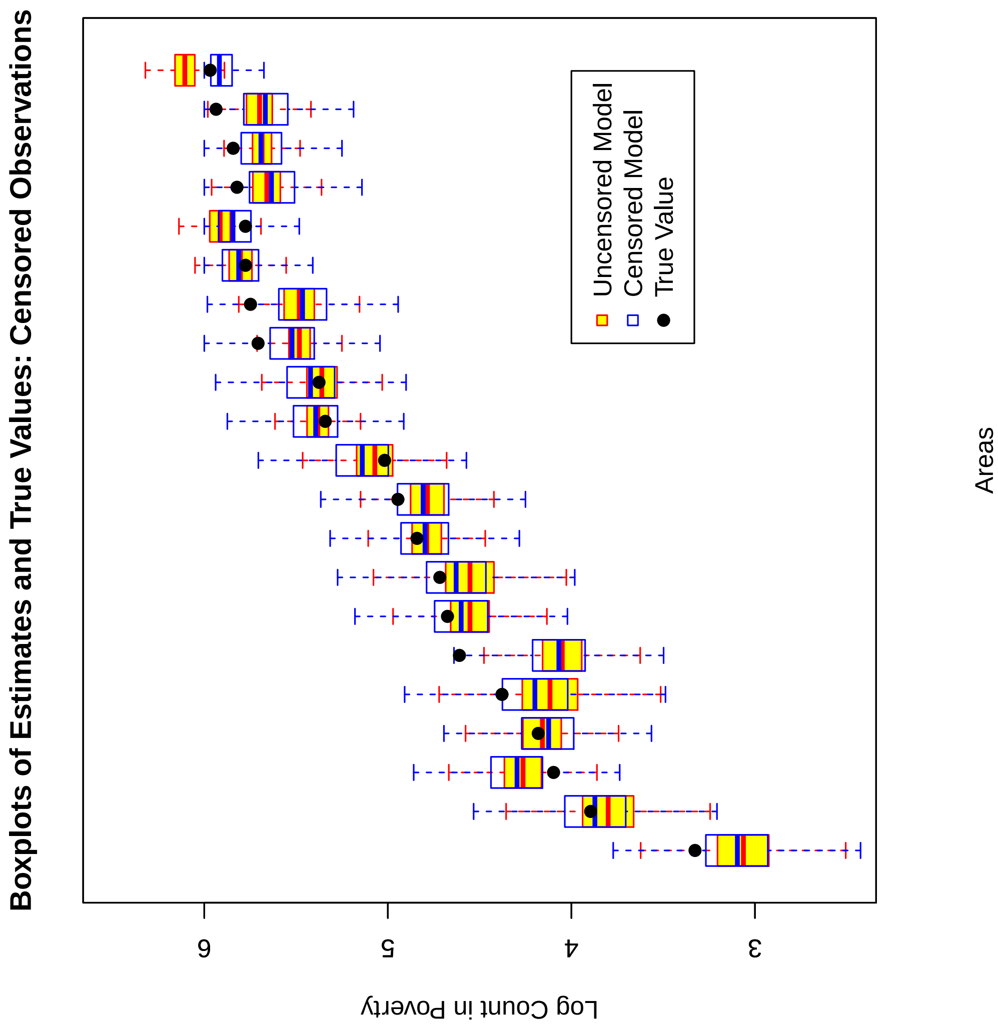

Boxplots of Estimates and True Values: Censored Observations
Uncensored Model
Censored Model
True Value
6
5
4
3
Log Count in Poverty
Areas

**Table 3**: Accuracy of Confidence Intervals (All Observations)

| | 50% Confidence Interval | | | 75% Confidence Interval | | | 90% Confidence Interval | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | In | Out | Coverage Rate (%) | In | Out | Coverage Rate (%) | In | Out | Coverage Rate (%) |
| Uncensored | 94 | 56 | 63 | 123 | 27 | 82 | 139 | 11 | 93 |
| Censored | 95 | 55 | 63 | 132 | 18 | 88 | 145 | 5 | 97 |

**Table 4**: Accuracy of Confidence Intervals (Censored Observations)

| | 50% Confidence Interval | | | 75% Confidence Interval | | | 90% Confidence Interval | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | In | Out | Coverage Rate (%) | In | Out | Coverage Rate (%) | In | Out | Coverage Rate (%) |
| Uncensored | 7 | 14 | 33 | 10 | 11 | 48 | 14 | 7 | 67 |
| Censored | 11 | 10 | 52 | 18 | 3 | 86 | 19 | 2 | 90 |

Figure 1 shows boxplots from the distribution of the estimated true values, $Y_i$, of the censored observations from the uncensored model and the censored model. The uncensored model produces smaller posterior variances. This is expected since the uncensored model estimates are purely synthetic. Moreover, the uncensored model seems to be more biased than the censored model, see Figure 1 and Table 4.

### 3.4 Application to School District Data

#### *3.4.1 Setting up the Model for School Districts*

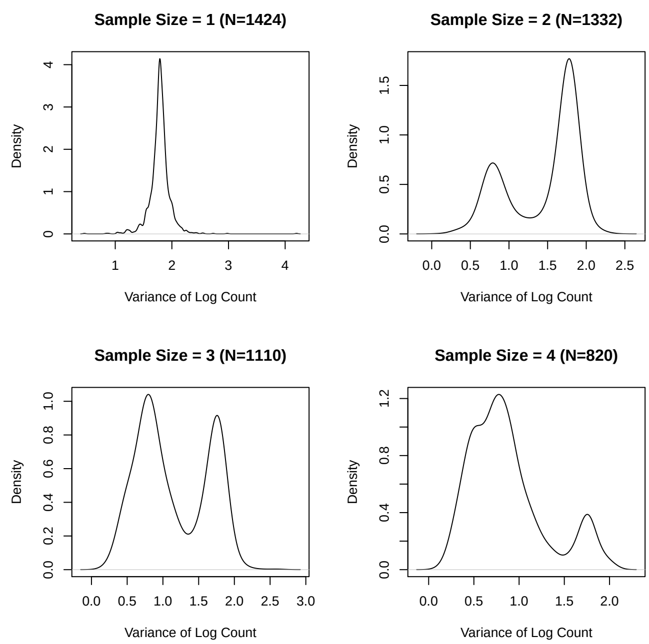
When applying the model at the school district level, some changes need to be made. First, not all the covariates in the SAIPE county model can be used in the school district model. The number of participants in SNAP is not used because those data are not available below the county level. Also, there are a number of school districts that do not have a count of poor school-age children from Census 2000. This is usually the result of a school district that did not exist in 2000, although it can indicate that there were zero poor children in sample in the Census 2000 long form. To obtain a more timely account of the number of children in poverty, Census 2000 counts of school-age children in poverty were combined with the corresponding counts from the 5-year ACS. The values were combined in the following manner. If the 5-year ACS count was zero, the log of the Census 2000 count was used. If the Census 2000 count was zero, the log of the ACS 5-year count was used. If both counts were greater than zero, the log counts were combined such that the log of the ACS 5-year count contributed 60% and the log of the Census 2000 count contributed 40%. Although Census 2000 is considered more reliable than the ACS 5-year estimate, the Census information is over a decade old. Due to the timeliness of the ACS 5-year estimate in comparison to Census 2000, more weight was given to the ACS 5-year estimate. After combining the ACS 5-year with Census 2000, very few school districts had a resulting count of zero. These observations were deleted. The covariates for the school district model are as follows:

- 0.6 X log(ACS 5-year number of poor school-age children in school district $i$) + 0.4 X log(2000 Census number of poor school-age children in school district $i$), except as described above

- log(estimated population 0-17 in school district $i$)
- log(number of child tax exemptions reported by families in poverty in school district $i$)
- log(number of child tax exemptions reported in school district $i$)

Additionally, because geocoding of tax data is poor in certain areas, and there is currently no established method of assigning nongeocoded tax returns to school districts, only areas with high geocoding rates were considered for the simulation. The decision was made to use Connecticut and Rhode Island, which both have over a 90% geocoding rate for their tax returns.

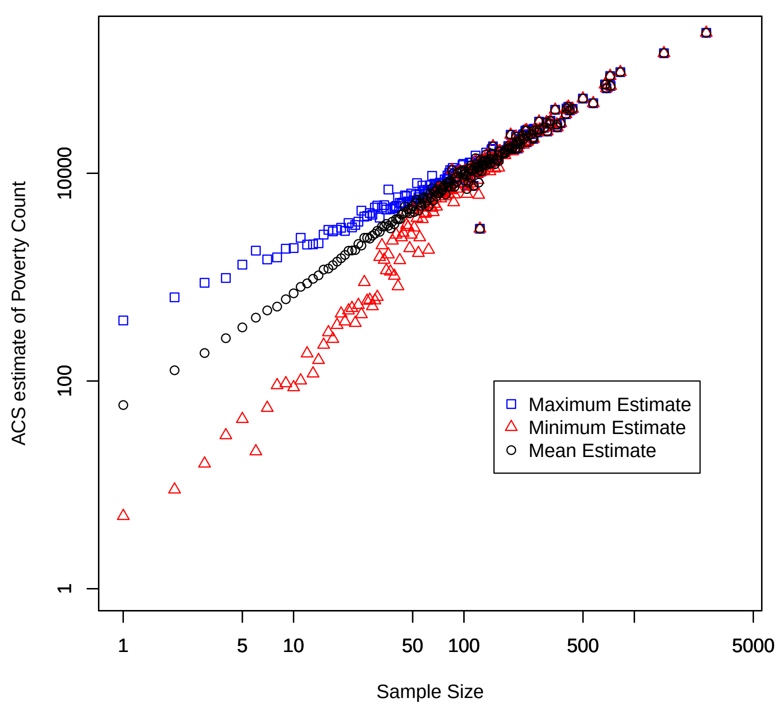
**Figure 2**: Variance Distribution by Sample Size




To execute the model, it is necessary to have a sampling variance for all of the observations. The sampling variance of the log counts was calculated by using the replicate weights from the ACS file. This method does not provide a variance for areas with an ACS estimate of zero. In an effort to compute the variance, the previous two years of data were examined, 2008 and 2009. If there was an estimate in either of those years, the variance of the most current estimate was plugged in for those areas. This brought the number of school districts with no variance down to about 11% in the chosen area of Connecticut and Rhode Island. However, that is still a significant portion of the school districts being considered. To derive a variance estimate for the remaining school districts, the variance distributions for all school districts in the nation with sample sizes of one, two, three, and four were examined, see Figure 2. All the graphs show a peak at around 1.8. The variance seems to converge on 1.8 as the sample size gets smaller and smaller, so if an area had no estimate of the variance, it was assigned a variance of 1.8. Other ways to estimate the sample variance are a topic for future research.

The censored model is dependent on the censoring threshold. So, a threshold for the model needs to be set. There is no obvious choice for the threshold. The ACS weights themselves are extremely variable. For example, when there is only one poor school age child in sample the ACS estimates range from 5 to 383 with an average of 59. When there are two poor children in sample, the ACS estimates range from 9 to 640, with an average of 127. Figure 3 shows the range of direct estimates by sample size, plotting the maximum, minimum, and mean ACS estimate for each sample size. It might be helpful to use a linear regression, and use the intercept as the threshold. However, Figure 3 shows that the data does not appear linear for small sample sizes. As a proxy, the maximum estimate for a sample size of one is taken as the cutoff; this results in a threshold of 5.95. Better ways of estimating the threshold is a possible consideration for future research, including the possibility of modeling the threshold as an unknown.

**Figure 3**: Plot of the Maximum, Minimum, and Mean ACS Estimate of Poverty Count Against Sample Size (Log Scale)




## 3.5 Further Research on the Censored Model

Work continues on the censored model. The model has been fully implemented on school district data and is currently being analyzed. These results will be presented later. Additionally, research is being done on assigning tax returns to Census blocks so that poor geocoding is no longer a limitation.

## REFERENCES

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, Washington, D.C.: Chapman & Hall.

National Research Council (1998), "Small-Area Estimates of School-Age Children in Poverty," Constance F. Citro, Michael L. Cohen, and Graham Kalton, eds. Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics. Washington, DC: National Academy Press.

R Development Core Team (2010), "R: A Language and Environment for Statistical Computing," http://www.R-project.org, R Foundation for Statistical Computing, Vienna, Austria.

Schirm, A., and Zaslavsky, A. (1997), "Reweighting Households to Develop Microsimulation Estimates for States", *Proceedings, Section on Bayesian Statistical Science*, Section on Survey Research Methods, American Statistical Association.

Slud, E. V. and Maiti T. (2011), "Small-area estimation based on survey data from a left-censored Fay-Herriot model", *Journal of Statistical Planning and Inference*, 141, 3520-3535.

U.S. Census Bureau (2010), "2006 - 2009 County-Level Estimation Details," last modified December 8, 2010, http://www.census.gov/did/www/saipe/methods/statecounty/20062009county.html

U.S. Census Bureau (2011), "American Community Survey Accuracy of the Data (2010)," last modified September 14, 2011, http://www.census.gov/acs/www/Downloads/data_documentation/Accuracy/ACS_Accuracy_of_Data_2010.pdf