Calculating Standard Error Estimates on American Community Survey Data with Variables Imputed from Outside Sources*

Daniel Scheer, Mark Levitan, Christine D'Onofrio,
John Krampner, and Todd Seidel

New York City Center for Economic Opportunity
Department of Poverty Research

Abstract

The New York City Center for Economic Opportunity (CEO) is engaged in developing an alternative poverty measure, based on the National Academy of Sciences' (NAS) recommendations, which employs the American Community Survey (ACS) as its main data set. While the ACS is a rich data source for measuring pre-tax cash income, it lacks data on several important components of family resources required by the NAS proposal. In order to estimate a NAS-style poverty rate for NYC, CEO must impute these variables into the ACS from outside data sets. By incorporating outside data into the ACS, however, we introduce two additional sources of error into our estimates: (1) error from the model parameters used for imputation; and (2) sampling error from the outside data sets.

This paper describes a modification of the Census-recommended ACS variance estimator, which is designed to capture the additional error introduced through imputing variables from outside data sources. It will compare the results of this modified variance estimator with variance estimates derived from the Census-recommended approach. We find that the modified variance estimator increases the standard error of our Citywide poverty estimates by roughly 80 percent, while the change in standard errors for different subgroups varies substantially. Implementing this new procedure would raise concerns over whether the resulting standard error estimates would be too large to be useful for inference.

Key Words: Survey data, poverty measurement, statistical matching, variance estimation

1. Introduction

The New York City Center for Economic Opportunity (CEO) is engaged in developing an alternative poverty measure for New York City, based on the National Academy of Science's (NAS) recommendations (Citro and Michael (eds.), 1995). CEO employs the Census Bureau's American Community Survey (ACS) as its main data set because it provides a large annual sample for New York City (NYC). While the ACS is a rich data source for measuring pre-tax cash income, it lacks data on several important components of family resources required by the NAS proposal, including the cash value of Food Stamps, housing assistance, medical out-of-pocket spending (MOOP) and childcare costs.[1] In order to estimate a NAS-style poverty rate for NYC, CEO must impute these variables into the ACS

*We are indebted to Michael Cohen, John Czajka, Joe Salvo, and Alan Zaslavsky for their insightful comments on earlier drafts of this work.

[1]In addition to these items, CEO employes a tax simulation program to estimate the dollar values of net taxes, and must estimate the cash-equivalent value of other costs and programs like School Meals, energy assistance, and commuting costs. Since the tax simulator and the estimation of commuting cost are rules-based and estimated mechanically from reported income, we do not consider them "imputed values." School meals and energy assistance are small components of the poverty measure and are not likely large contributors to the added variance from imputation.

from outside administrative and survey data sets. By incorporating outside data into the ACS, however, we introduce additional sources of error into the resulting poverty rate estimates that are not accounted for by the variance estimation method the Census Bureau has developed for the ACS.

This paper describes a modification to the Census-recommended ACS variance estimator, which is designed to capture the additional error introduced through imputing variables from outside data sources. It compares results of this modified variance estimator with variance estimates derived from the Census-recommended approach. A final section discusses the findings and poses questions for future work.

2. Variance Estimation in the ACS

The ACS relies on a complex systematic sampling design[2] and weighting adjustment (U.S. Census Bureau, 2009). While systematic sampling designs have many advantages, they lack an unbiased, design-based variance estimator (Opsomer et al., 2010). To address this, the ACS uses Fay's successive difference replication (SDR) estimator. The SDR method involves creating R replicate factors (where R is a multiple of four), which are used to multiply the full-sample weight, producing replicate weights (Fay and Train, 1995). Fay's method is based on comparisons of neighboring observations, ensuring that the order of selection inherent in the systematic sampling process is taken into account (Gbur and Fairchild, 2002). In the case of the ACS, 80 replications are provided.

The ACS Public Use Micro Sample file contains household and person-level identification variables, a set of respondent-reported variables (e.g., income, marital status, etc.), the ACS full-sample weight, and the 80 replicate weights. In order to calculate a mean for a given variable, Y, the observations are weighted by the full-sample weight. To estimate the variance of the mean of Y, the ACS documentation suggests the following formula:

$$V_{SDR}(\bar{Y}) = \frac{4}{80}\sum_{r=1}^{80}(\bar{Y_r} - \bar{Y_o})^2$$

where $\bar{Y_o}$ is the mean of Y weighted by the full-sample weight and $\bar{Y_r}$ is the mean of Y weighted by replicate weight r:

$$\begin{array}{l} \bar{Y_o} = \frac{1}{n}\sum_{i=1}^{N} w_i y_i \\ \bar{Y_r} = \frac{1}{n}\sum_{i=1}^{N} f_{i,r} w_i y_i \quad \text{for} \quad r = 1, ..., 80 \end{array}$$

The full sample weight is given by w_i and replicate weight, r, is given by $f_{i,r}w_i$. The replicate factors, $f_{i,r}$, are derived from an $R \times R$ Hadamard matrix, $A = [a_{i,j}]$, where $a_{i,j} = \pm 1$. Since the number of observations in the ACS is much larger than the number of replicate weights, rows of the Hadamard matrix must be assigned to multiple observations. The Census Bureau uses an algorithm to assign pairs of rows from the Hadamard matrix to each record in the ACS. This assignment repeats every 780 records until all records in the ACS receive a pair of rows from the Hadamard matrix (Census Bureau, 2009). Each replicate factor $f_{i,r}$ is then a linear combination of the r^{th} column entry of the two rows assigned to observation i:

$$f_{i,r} = 1 + (2)^{-\frac{3}{2}} a_{R1,r} - (2)^{-\frac{3}{2}} a_{R2,r} \quad \text{for} \quad i = 1, ..., N$$

We refer to this method as the Census-SDR method.

[2]Systematic samples involve selecting elements from an ordered sampling frame. Typically, every k^{th} element is selected, where k is the total population divided by the sample size.

3. Imputing Missing Variables into the ACS

As noted above, the ACS does not contain all of the information necessary to construct a NAS-style poverty measure. Rather, data from outside sources (both administrative and survey) must be combined with the ACS to create a synthetic data set. CEO's reports refer to this data set as the "American Community Survey Public Use Micro Sample as augmented by CEO."

Integrating this additional data into the ACS can be accomplished through a broad category of methods known as "statistical matching" (D'Orazio et al., 2006). Given two data sets, A and B, where A and B share common variables X, and B contains a variable of interest Y, a synthetic data set, C, can be created by matching observations in A and B based on their common characteristics within the vector of variables X. The goal of the match is for the joint distribution, $f(Y, X)$, in the donor data set to be preserved in the synthetic data set (D'Orazio et al., 2006).

An appropriate statistical match requires that two assumptions be fulfilled. First, the data sets A and B must be independent and drawn from the same population.[3] Second, the Conditional Independence Assumption (CIA) must hold. Consider two samples, A, with variables (X, Y) and B, with variables (X, Z) that are to be matched into a synthetic data set containing the variables (X, Y, Z). In this context, the CIA implies $Y|X \perp Z$; that is, the variable Y is independent of Z, conditional on the matching variables, X. It is worth noting that the CIA cannot be tested directly given the data sets A and B. Rather, it is assumed that the model for $f(X, Y)$ is correct.

CEO has developed a variety of models that estimate the effect of Food Stamps, housing assistance, childcare costs and Medical out-of-pocket expenditures (MOOP) on total family resources and poverty status.[4] In the following sections, we provide a brief description of these models.

3.1 Food Stamps

The ACS indicates whether anyone in a household received Food Stamps in the prior 12 months. The dollar amount of benefits, however, is not provided. To estimate Food Stamp benefits, we make use of New York City Human Resources Administration Food Stamp records by statistically matching Food Stamp cases in this administrative data to "Food Stamp Units" we construct in the ACS data. These units are sub-divisions of the ACS household, which more closely mimic the composition of Food Stamp cases observed in the administrative data. The administrative data includes all cases in New York City that were active for any period between July and June of the appropriate year. This period is chosen because it represents the mid-point in the ACS rolling sample, helping to ensure that the administrative data is comparable to the ACS data. To preserve comparability with our poverty universe, individuals in group quarters were removed from both the administrative data and the ACS sample. For each case in the administrative data, we sum the total of Food Stamp payments over the previous year. We then take a random sample of the administrative data, which is used to impute into the ACS.

In order to match ACS and administrative observations, we rely on a predictive mean match (PMM) technique, originally developed by Rubin (1987) and Little (1988). The

[3] It should be noted that the data we use for our childcare and medical spending models come from national, rather than local New York City data, and thus violate this assumption. We currently do not have access to local data, but we believe that using national data is justified since spending on these items is not significantly driven by New York City-specific factors.

[4] For more detailed descriptions of the development of the models used to generate the data used in this paper, see Levitan, et al., 2013.

PMM algorithm has three steps:

1. Estimate a regression model of Y on X in the donor data set, B.
2. Use the estimated coefficients to compute predicted values for Y in the host data set, A.
3. Match host observations to the donor observation with the closest predicted value and donate the $actual$ value of Y to the new, synthetic data set.

We developed a regression model to predict the yearly value of Food Stamps in the administrative data. Given data limitations within the administrative data, the regression model is restricted to the following variables:

Table 1: Regression Variables for Food Stamp Model

Case size
Number of Children
Elderly Case Head? (Y/N)
Elderly or Disabled Member in the Case? (Y/N)
Age of the Case Head
Income Reported on Food Stamp Application (Three categories)

We use the age of the case head variable as a proxy for employment status. Since many families apply for Food Stamps during a period when they have little or no income, the distribution of income reported on Food Stamp applications is highly skewed, with zero values up to the 75th percentile. Given this unusual distribution, we recode the income variable into three categories: (1) less than the 75th percentile; (2) 75-89th percentile; and (3) at or above the 90th percentile of the income distribution of the Food Stamp sample.

The ACS and administrative values are matched via PMM, with the added constraint that both the host and donor cases are in the same Community District.[5] This additional match criterion is designed to capture neighborhood effects that are not explicitly in the model. Once an administrative case donates its value to an ACS case, it is removed from the donor pool.

3.2 Childcare Expenses

The CEO poverty measure treats childcare spending as a non-discretionary expense that is subtracted from income. Because the American Community Survey provides no information on childcare spending, we rely on weekly childcare expenditure data reported in the Census Bureau's Survey of Income and Program Participation (SIPP). We restrict the SIPP data to only include families with all parents working at least part of the year, at least one child 12 years of age or younger, and that live in an urban area.

Following Iceland and Ribar (2001), we estimate separate regression models for single- and two-parent families. The variables are measured at the family level, except for race and education, which are measured for the designated parent. In order to achieve the best possible fit to the data and deal with potential non-linear relationships, we employ a generalized additive regression model (GAM). A GAM is a regression model that allows different

[5]In New York City, Community Districts are geographical subdivisions, quite similar to the Census Bureau's Public Use Microdata Areas identified in the ACS.

functional forms for each independent variables. Some of the variables used in the regression model are included as dummy variables, while others are fit non-parametrically, using smoothing spline functions (Keele, 2008). The model uses the variables shown in Table 2.

Table 2: Regression Variables for Childcare Cost Models

Earned Income*
Number of Children 0-5*
Number of Children 13-17*
Number of Adults in Family*
Female Income as a Proportion of Family Income*
Receiving Food Stamps? (Y/N)
Completed High School? (Y/N)
Completed Some College? (Y/N)
Completed College? (Y/N)

* Variables are fit non-parametrically.

The regression models are used to compute predicted values in the SIPP and ACS, and the observations are matched via PMM, with the added constraint that both the host and donor cases have the same marital status. The matched weekly childcare values are then multiplied by the number of weeks worked in the prior 12 months in order to estimate annual costs.

3.3 Medical Out-of-Pocket Expenditures (MOOP)

The cost of medical care is also treated as a non-discretionary expense that limits the ability of families to attain the standard of living represented by the poverty threshold. MOOP includes health insurance premiums, co-pays and deductibles as well as the cost of medical services that are not covered by insurance. In a manner similar to that for childcare, we use an imputation model to match MOOP expenditures by families in the Agency for Healthcare Research and Quality's Medical Expenditure Panel Survey (MEPS) to families in the ACS sample.

All variables used in the imputation are measured at the family level or for the head of the family unit. Income, family size, number of children, and age of the householder are measured as continuous variables, while the race, education, working status and insurance status categories are included as binary variables. Similar to the childcare model, we estimate MOOP with a GAM regression. The regression contains the following variables:

Table 3: Regression Variables for MOOP Model

Family Income*
Family Size*
Number of Children in Family*
Age of Household Head*
Race/Ethnicity Dummy Variables
Public Insurance (Y/N)
No Insurance (Y/N)
Work Full-Time? (Y/N)
Completed High School? (Y/N)
Completed College or Greater? (Y/N)
Elderly Household Head? (Y/N)
Public Insurance and Elderly? (Y/N)
No Insurance and Elderly? (Y/N)

* Variables are fit non-parametrically.

Observations in the MEPS and ACS are matched via PMM, with the added constraint that both the host and donor cases have the same elderly and insurance status.

3.4 Housing Adjustment

While the high cost of housing makes New York City an expensive place to live, some New Yorkers do not need to spend as much to secure adequate housing as the threshold implies. Many of the City's low-income families live in public housing or receive a housing subsidy, such as a Section 8 housing voucher. A large proportion of New York's renters live in rent-regulated apartments. Some homeowners have paid off their mortgages and own their homes free and clear. We make an upward adjustment to these families' incomes to reflect this advantage.

The ACS does not provide data on housing program participation, however. To address this, we turn to the unique-to-New York City Housing and Vacancy Survey (HVS), a triennial Census survey which collects detailed information on geographic, demographic, and housing-related characteristics of housing units and their occupants. By matching renter households in the ACS to renter households in the HVS we are able to impute the missing housing program status and the out-of-pocket expenditures data to the ACS. Our "cold-deck" matching routine is based on a set of household and head of household characteristics that identify corresponding households between the ACS and HVS. Listed below are characteristics used for matching renter households in the matching algorithm:

Table 4: Variables for Housing Status Cold-Deck

Neighborhoods: Community District or Public Use Microdata Area
Race/Ethnicity Dummy Variables
Householder 65 or older? (Y/N)
Rank of Equivalized Household Income
Rank of Contract Rent
Number of Bedrooms in the Household
Household Composition
Household had Wage Income? (Y/N)

Our initial match is an attempt to match on all eight household characteristics. If we do not find a matching household in the HVS, we incrementally remove or relax characteristics and attempt to match again. Our goal is to preserve the geographical, racial, and family composition distribution of the housing statuses found in the HVS. Because the distribution of participation in means-tested housing assistance (in particular the location of public housing) varies by neighborhood, we attempt to match as many households as possible within the same neighborhood. We then move to adjacent neighborhoods and finally neighborhoods within the same borough.

4. Incorporating Additional Error from Imputation into Standard Error Estimates

Imputing missing variables from outside data sets introduces two additional sources of error into our poverty rate estimates. First, the matching procedures rely on regression models; since these models are estimated, their parameter estimates contain error.[6] Second, data are imputed into the ACS from samples. Since these data come from samples, the imputed values have associated sampling error. Thus, any imputed value can be viewed as a single draw from a population of possible imputed values. Treating the imputed values as reported data in the ACS and applying the Census-recommended formula for calculating variances in the ACS would likely lead to an underestimate of the variance, as there is some additional error associated with each of the imputed values.

CEO contracted with Mathematica Policy Research to develop a framework for incorporating imputation error into our standard error calculations. Their proposed framework, based on the Law of Total Variance (LTV), aims to estimate the standard error of the CEO poverty rate across a set of alternative imputations.[7] The Law of Total Variance is a formula for estimating the variance of a random variable P, given a random variable I:

$$V[P] = E[V[P|I]] + V[E[P|I]] \tag{1}$$

According to (1), the variance of P is equal to the expected value of the variance of P given I plus the variance of the expected value of P given I. In the context of the CEO poverty measure, P is the estimated poverty rate and I represents a range of alternative sets of imputed values.

In order to estimate $E[V[P|I]]$ and $V[E[P|I]]$, we need to generate a set (of size M) of samples of I that is based on varying values for the imputed variables. Each set is comprised of a vector of values for each of the four imputed variables described above. The M alternative imputation sets therefore create a sample of the range of possible imputation sets from which the original imputation set could have been drawn.

Generating alternative values for the imputed variables requires two steps. First, for each alternative imputation of each variable, we reestimate the regression models after resampling the donor data set. This technique is similar to a bootstrap: from the donor data set with N observations, we draw a sample of size N, with replacement, in order to generate a simulated sampling distribution of the data. For example, if the original donor data set for a particular variable contained 10 observations, the resampled data set would also contain 10 observations, with some of the original observations dropped and other observations repeated. The regression is rerun on this simulated data set. This procedure yields a distribution of regression coefficients, and consequently, a distribution of predicted values on which donor and host observations are matched. Resampling and reestimating

[6] Since the housing imputation model is not based on a regression model, there is no error from parameter estimates to account for.

[7] This framework was outlined in Frank Potter, John Czajka and Eric Grau. "Imputation Variance Questions." Memo. December 17, 2010.

the regression models accounts for the first source of additional error noted above, namely error from the imputation model.

After generating a set of new regression models based on the resampled data, values from the original donor data are matched into the host data set using a predictive mean *neighborhood* match (PMN). A PMN match is similar to a PMM. Donor and host cases are matched based on predicted values. But, rather than matching the single closest host and donor cases, host cases are matched to a neighborhood of the six closest donor cases, based on the absolute difference in predicted means.[8] One of these six closest donor cases is then selected at random, and the actual value is donated to the synthetic data set. This perturbation in the matching technique accounts for the second source of additional error noted above, that the imputed data comes from samples (Singh, et al, 2002).

Calculating the variance of the poverty rate estimate given the M alternative imputation sets then proceeds in several steps. First, we calculate $E[V[P|I]]$, the expected value of the variance of the poverty rate, given the M alternative imputation sets. For each alternative imputation set, i, we calculate the variance of the poverty rate estimate, using the formula:

$$V[Pov_i] = \frac{4}{80}\sum_{j=1}^{80}(Pov_{i,j} - Pov_{i,o})^2 \tag{2}$$

where $Pov_{i,o}$ is the poverty rate estimate for imputation set i using the full sample weight, denoted as o, and $Pov_{i,j}$ is the poverty rate estimate for imputation set i using replicate weight j, for $j = 1, 2, \ldots, 80$. After computing the M variances, we take the average of these variances across the i imputation sets:

$$E[V[P|I]] = M^{-1}\sum_{i=1}^{M} V[Pov_i], \quad \text{for} \quad i = 1, ..., M \tag{3}$$

Second, we calculate $V[E[P|I]]$, the variance of the average poverty rate across the group of imputation sets. To do so, we compute the average poverty rate for each imputation set, i, across the 80 replicate weights, given as $\overline{Pov}_{i,j}$. We then take the variance of the averages:

$$V[E[P|I]] = (M-1)^{-1}\sum_{i=1}^{M}(\overline{Pov}_{i,j} - \overline{Pov}_{i,o})^2, \quad \text{for} \quad i = 1, ..., M \tag{4}$$

where $\overline{Pov}_{i,o}$ is the average of the poverty rates using the full sample weights across the group of imputation sets. Summing the terms $E[V[P|I]]$ and $V[E[P|I]]$, as defined above, yields the estimate of the variance of the poverty rate. The standard error of the poverty rate, which we use for significance testing, is the square-root of this value. We refer to this method as LTV-SDR.

5. Results

We are mainly interested in two aspects of this variance procedure. First, how many alternative imputation sets are required in order to achieve a consistent estimate? Second, how much larger are the LTV-SDR standard error estimates compared to the Census-SDR ACS standard error estimate using only the original set of imputations? We investigated these questions by implementing the LTV-SDR procedure using M = 5, 10, 15, and 20 alternative imputation sets for the New York City sample as a whole, as well as for several sub-group populations.

[8] While CEO uses absolute difference to define neighborhoods, PMN techniques can be generalized to other distance functions, including Euclidian and Chebyshev. See D'Orazio et al., 2006 for a discussion of distance functions.

5.1 The Required Number of Alternative Imputations

The question of how many alternative imputations are required can be answered with Rubin's formula for relative efficiency: RE $= 1/(1 + F/M)$, where F is the fraction of missing data and M is the number of alternative imputations (Rubin 1987). This formula measures the efficiency of the point estimates for a given number of alternative imputations. In the application discussed in this paper, F is equal to 100 percent. This means that achieving point estimates 90 percent as efficient as those with fully observed data requires nine alternative imputations; for 95 percent efficiency, the required number of alternative imputations is 19.

The efficiency formula is designed for problems where some fraction of the data is missing. In this application, all of the data is unobserved. Given this, we also computed standard error estimates using 5, 10, 15, and 20 alternative imputation sets for the 2008 and 2011 ACS samples. The largest gain in variance comes from moving from zero alternates to 5 alternates. Between 15 and 20 alternates, there is little change in the standard errors. Between the efficiency formula and these simulations, we feel confident in using 20 alternative imputations for computing standard errors.

5.2 Comparing the Variance Estimation Procedures

Table Five compares the standard error estimates generated by the Census-SDR method with those generated by the LTV-SDR method, based on 20 alternative imputation sets. There is a wide range in the percentage change in the standard error estimates over the Census-SDR estimate, ranging from -3.5 percent to 77.0 percent in 2008 and from -1.0 percent to 79.3 percent in 2008.[9] Despite the variability across sub-groups, the standard error estimates appear to be consistent over the two sample years. Such consistency lends credence to this statistical approach.

We expected to see the largest increases among groups that rely more heavily on imputed values, but this was not the case. For example, persons in single-parent families saw relatively small changes in their standard error estimates (-3.5 percent and -1.0 percent in 2008 and 2011, respectively) while the overall sample, persons aged 18-64, and Non-Hispanic Whites saw the largest increases.

One possible explanation has to do with the composition of the sub-groups we are considering. Since our imputation procedures are based on individual/household characteristics, there is a fairly narrow range of possible imputation values for a given observation. If a sub-group is fairly homogeneous, then considering multiple imputation sets may not increase the standard error estimate much. In contrast, for more heterogeneous groups, such as individuals 18-64, there is more variation in individual/household characteristics and, correspondingly, more variation in possible imputation values.

In order to test this hypothesis, we constructed a measure of variability of the different imputed values across the alternative values for each of the sub-groups. For each of the 20 alternative values of each imputed value, we calculated the sample standard deviation and took the average standard deviation across the set of imputed values. We then divided this by the average mean across the set of imputed values, yielding relative standard errors across the imputation set. We restricted the analysis to observations up to 150 percent of the poverty threshold. It is important to subset the data in this way, as poorer households are more likely to have non-zero values for the imputed variables.

[9] It was surprising to find cases in which the LTV-SDR procedure, which incorporates the imputation variance, yields *smaller* standard errors than the Census-recommended approach. However, in the instances where the LTV-SDR estimates were smaller the decrease is slight and not meaningfully different than no change from the Census-SDR estimate.

Table 5: **Census-SDR versus LTV-SDR Standard Error Estimates**

2008

	Census	LTV	Percent Change
Overall	0.323	0.571	77.0
Age $<$ 18	0.636	0.836	31.4
Age 18-64	0.327	0.574	75.7
Age $\geq$ 65	0.526	0.775	47.3
One Parent	1.251	1.208	-3.5
Two Parent	0.625	0.752	20.2
White	0.429	0.658	53.2
Black	0.605	0.854	41.2
Asian	0.995	1.063	6.9
Hispanic	0.738	0.892	20.9

2011

	Census	LTV	Percent Change
Overall	0.337	0.604	79.3
Age $<$ 18	0.645	0.840	30.2
Age 18-64	0.334	0.598	79.0
Age $\geq$ 65	0.637	0.835	31.1
One Parent	1.142	1.130	-1.0
Two Parent	0.654	0.821	25.4
White	0.446	0.718	60.8
Black	0.630	0.861	36.8
Asian	1.158	1.147	-1.0
Hispanic	0.745	0.906	21.7

Source: American Community Survey Public Use Micro Sample as augmented by CEO.

Table Six contains the relative standard errors for each imputed variable and sub-group. The table provides some evidence to support the contention that more homogeneous groups have less variability in imputed values. For example, single-parent families have relative standard errors for childcare costs of 3.87 in 2008 and 4.10 in 2011, compared with 6.45 and 6.90, respectively, for the overall sample. Similarly, persons 65 and older have a lower relative standard error for MOOP than the full sample.

Table 6: **Relative Standard Errors for Imputed Values by Sub-Group**

2008

	Food Stamps	Housing	MOOP	Childcare
Overall	1.83	1.57	1.57	6.45
Age $<$ 18	1.52	1.44	1.59	4.54
Age 18-64	2.01	1.72	1.55	7.65
Age $\geq$ 65	1.72	1.28	1.49	21.47
One Parent	1.34	1.20	1.66	3.87
Two Parent	1.98	1.87	1.42	7.20
White	2.33	2.01	1.37	9.77
Black	1.71	1.44	1.51	5.33
Asian	2.40	1.93	1.58	8.62
Hispanic	1.56	1.36	1.72	5.79

2011

	Food Stamps	Housing	MOOP	Childcare
Overall	1.52	1.53	1.64	6.90
Age $<$ 18	1.23	1.40	1.76	4.80
Age 18-64	1.66	1.68	1.63	8.15
Age $\geq$ 65	1.54	1.30	1.34	15.84
One Parent	1.09	1.13	1.83	4.10
Two Parent	1.48	1.79	1.55	6.84
White	1.98	2.05	1.36	8.59
Black	1.49	1.42	1.61	5.82
Asian	2.15	2.00	1.59	9.67
Hispanic	1.20	1.24	1.89	6.31

Source: American Community Survey Public Use Micro Sample as augmented by CEO.

Note: Standard Errors calculated for individuals whose income is less than or equal to 150 percent of the threshold.

Some groups with small changes in their standard error estimates have large relative standard errors for the different imputed values. In particular, while Asians have higher relative standard errors for Food Stamps, housing and childcare than the overall sample, they have the smallest increase in the standard error of their poverty rate estimate. As a result, we also computed the extent to which different groups rely on imputed values relative to values reported in the ACS in their overall estimate of income. Table Seven reports the ratio of imputed income to reported income across the different sub-groups.

Table 7: **Reliance on Imputed Values by Sub-Group**

2008

	Food Stamps	Housing	MOOP	Childcare	All Values
Overall	0.113	0.297	0.116	0.007	0.532
Age $<$ 18	0.173	0.413	0.100	0.013	0.700
Age 18-64	0.093	0.251	0.118	0.004	0.465
Age $\geq$ 65	0.074	0.234	0.170	0.000	0.478
One Parent	0.235	0.549	0.114	0.020	0.917
Two Parent	0.063	0.166	0.085	0.003	0.317
White	0.049	0.150	0.176	0.003	0.378
Black	0.177	0.453	0.133	0.010	0.773
Asian	0.041	0.175	0.094	0.002	0.314
Hispanic	0.135	0.328	0.078	0.008	0.548

2011

	Food Stamps	Housing	MOOP	Childcare	All Values
Overall	0.189	0.355	0.107	0.006	0.658
Age $<$ 18	0.265	0.440	0.085	0.012	0.802
Age 18-64	0.169	0.281	0.107	0.004	0.562
Age $\geq$ 65	0.128	0.264	0.140	0.001	0.533
One Parent	0.370	0.776	0.117	0.018	1.282
Two Parent	0.138	0.167	0.075	0.004	0.384
White	0.088	0.154	0.146	0.003	0.391
Black	0.221	0.561	0.13 0	0.008	0.920
Asian	0.105	0.142	0.093	0.002	0.343
Hispanic	0.250	0.407	0.074	0.007	0.739

Source: American Community Survey Public Use Micro Sample as augmented by CEO.

Notes: Ratios calculated as the average of the imputed values over the non-imputed income for the particular group. Ratios calculated for individuals whose income is less than or equal to 150 percent of the threshold.

As is shown in the final column of Table Seven, Asians have the smallest reliance on imputed income of all the sub-groups, with ratios of 0.314 and 0.343 in 2008 and 2011, respectively. These ratios counteract the impact of high relative standard errors for the imputed values. Thus the change in the standard error of the poverty estimate for a particular group is a function of both the variability of the imputed values within the group and the average reliance on these imputed values.

6. Hypothesis Tests

The preceding sections have outlined the LTV-SDR procedure and discussed its properties. The procedure is relatively simple to implement and incorporates the additional variance introduced through imputation. Further, its properties are consistent across the two years of data tested in this paper. The difference between the the Census-SDR and LTV-SDR standard error estimates is a function of both the within group variability of imputed values and the extent to which groups rely on imputed variables.

An important question to consider is how useful this procedure is for applied analysis. Local area poverty measurement often involves comparatively small samples, particularly among certain sub-groups. This sample-size issue makes it difficult to detect statistically significant changes in poverty rates over time. Thus since the LTV-SDR procedure will generally increase the size of the standard error of the poverty estimates, it will likely exacerbate this problem.

The table below shows statistical significance tests for the change in poverty rates from 2008 to 2011. It compares the margin of error (MOE) of the difference between the poverty rates at the 90 percent confidence-level[10] using the Census-SDR and LTV-SDR methods. Interestingly, in spite of the generally larger standard produced by the LTV-SDR method, all of the differences found to be statistically significant by the Census-SDR method are also significant using the LTV-SDR method.

Table 8: **Poverty Rate Change, 2008 - 2011**

	2008	2011	Difference	MOE (Census-SDR)	MOE (LTV-SDR)	Significant? (Census-SDR)	Significant? (LTV-SDR)
Overall	19.0	21.3	2.23	0.82	1.37	Yes	Yes
Age < 18	23.1	24.7	1.58	1.63	1.95	No	No
Age 18-64	16.9	19.9	3.03	0.79	1.36	Yes	Yes
Age ≥ 65	22.7	22.4	-0.29	1.36	1.87	No	No
One Parent	35.5	34.7	-0.75	3.01	2.72	No	No
Two Parent	15.6	18.7	3.17	1.64	1.83	Yes	Yes
White	13.2	15.4	2.17	0.97	1.60	Yes	Yes
Black	20.8	21.5	0.66	1.65	2.00	No	No
Asian	22.4	26.5	4.10	2.64	2.57	Yes	Yes
Hispanic	23.5	25.3	1.77	1.79	2.09	No	No

Note: "MOE" is the margin of error of the difference at the 90 percent level.

The results for the table above should be interpreted with caution. The period between 2008 and 2011 saw a steep rise in overall poverty associated with the Great Recession. The change in overall poverty in New York City during this period was 2.23 percentage points. This is considerably larger than the change from any other period between 2005 and 2011.[11] Using the LTV-SDR method will result in only large changes in poverty rates (greater than 1.4 percentage points) reaching the level of statistical significance. Using this method, none of the year-to-year changes in poverty rates from 2005 - 2011 would be statistically significant.

7. Conclusion

Integrating outside data into the ACS introduces an additional source of error that will not be captured by the survey's standard Census-SDR variance estimation method. The method and results presented in this paper represent an initial, formal attempt to measure this variance. The modified method for variance estimation introduced in this paper (based on the Law of Total Variance) shows promise as an alternative estimation technique.

The major concern with incorporating the additional variance from imputation is that the resulting standard errors will be too large to be useful for inference. Indeed, the LTV-SDR procedure yields standard errors that are nearly 80 percent larger for the overall City population, though the increase for various subgroups is somewhat less. There is, however, reason to regard these estimates as an upper bound on the standard error estimate. CEO imputes the four variables discussed in this paper separately from different data sources. As such, the alternative imputation values are also imputed independently from each other.

[10] Consistent with poverty reports published by the Census Bureau, we use the 90 percent confidence-level to assess statistical significance.

[11] CEO's poverty measurement work for New York City only covers the 2005 - 2011 period.

This means that the procedure for capturing the additional variance is not accounting for the correlations between imputed variables, some of which are likely to be large. If the imputed variables are correlated with each other, then the range of alternative imputation values will be constrained, and the resulting standard error estimates will be smaller. It is important for future work on this problem to deal with the correlation issue in order to produce a more accurate standard error estimate.

REFERENCES

Citro, Constance F. and Robert T. Michael (eds). (1995) Measuring Poverty: A New Approach. Washington, D.C.: National Academy Press.

D'Orazio, Marcello and Marco Di Zio and Mauro Scanu. (2006). Statistical Matching: Theory and Practice. Wiley Series in Survey Methodology. West Sussex.

Fay, Robert E. and George F. Train. (1995) "Aspects of Survey and Model-Based Post-Censal Estimation of Income and Poverty Characteristics for States and Counties." *Proceedings of the Section on Government Statistics.* Alexandria, VA: American Statistical Association, pp.154-159.

Gbur, Phillip M. and Lisa D. Fairchild. (2002). "Overview of the U.S. Census 2000 Long Form Direct Variance Estimation." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association, pp. 1139-1144.

Iceland, John and David C. Ribar. "Measuring the Impact of Child Care Expenses on Poverty." Paper presented at the 2001 Population Association of America (PAA) meetings in Washington, D.C., March 29, 2001.

Keele, Luke John. (2008). Semiparametric Regression for the Social Sciences. West Sussex, England: J. Wiley and Sons.

Levitan, Mark, Christine D'Onofrio, John Krampner, Daniel Scheer, Todd Seidel. (2013) *The CEO Poverty Measure, 2005-2011*. New York City Center for Economic Opportunity.

Little, R.J.A. (1988) "Missing-Data Adjustments in Large Surveys." *Journal of Business and Economic Statistics*, vol. 6, no.3, pp. 287-296.

O'Donnell, Sharon and Rodney Beard. Imputing Medical Out of Pocket Expenditures using SIPP and MEPS. Presented at the American Statistical Society Annual Meetings, August 2009.

Opsomer, J.D., M. Francisco-Fernández, and X. Li. (2010). "Model-Based Non-Parametric Variance Estimation for Systematic Sampling."

Potter, Frank, John Czajka and Eric Grau. "Imputation Variance Questions. Memo. December 17, 2010.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons.

Singh, A.C., E.A. Grau, and R.E. Folsom (2002) "Predictive Mean Neighborhood Imputation With Application To The Person-Pair Data Of The National Household Survey On Drug Abuse" *Proceedings of the Survey Research Methods Section.* Alexandria, VA: American Statistical Association, pp.154-159.

U.S. Census Bureau. (2009). *Design and Methodology: American Community Survey*. U.S. Government Printing Office, Washington, D.C.