Incorporating EITC Participation Data into the SAIPE Program's County Poverty Model*

Wesley Basel and Brett O'Hara
wesley.w.basel@census.gov, brian.j.ohara@census.gov
U.S. Census Bureau, Small Area Estimates Branch

Abstract

The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program produces state and county income and poverty estimates using decennial census data, household survey data, administrative records and population estimates. The IRS has provided the Census Bureau new data on earned income tax credit (EITC) participation. Incorporation and evaluation of this new data in the SAIPE program models could lead to improvements in small area estimates. The EITC is a uniformly administered and widely advertised program. Regional variation in participation rates might be lower than other programs that have been utilized in the SAIPE program models. Due to the wide coverage of the EITC, a particular goal of this research is improving modeling performance for counties with weak participation in other programs, thus reducing errors under the current model.

Keywords: small-area estimation, small-domain areas, SAIPE, poverty, EITC

I. Introduction

The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program produces model-based estimates of state and county income and poverty by age group. The SAIPE program's county estimates are used to produce estimates of child poverty at the school district level. These school district estimates are used by the Department of Education for funds allocation under the No Child Left Behind Act of 2001. The importance of this allocation program mandates a continual effort towards improvement and maintenance of the SAIPE program's intercensal poverty estimates.

The SAIPE program's estimates rely on regression techniques that incorporate data from the decennial census, household surveys, administrative records and population estimates. To produce final estimates, the predicted number of persons in poverty from the regression is combined with direct survey-based estimates of poverty from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS). This document is part of ongoing research to improve SAIPE program methodology.

An area-level model (e.g., counties, states) is used to produce the SAIPE program's estimates. One route of improvement for the SAIPE program's estimates is to find new data that is poverty-related and consistent at the geographic-level of interest. The IRS has provided the Census Bureau new data on potential qualification for the Earned Income Tax Credit (EITC) program. At the time of this study, the Census Bureau had received this new data for the tax year 2003. Incorporation and evaluation of this data in the SAIPE program models could lead to improvements in the small area estimates. Since the EITC program qualifications are designed for lower-income taxpayers, EITC status should be positively correlated with poverty status. Furthermore, the EITC is a uniformly administered and widely advertised program. As a result, regional variation in eligibility rates relative to actual poverty might be lower than other administrative data used in the SAIPE program models.

Our main interest in using the EITC data is improving model performance, in terms of goodness of fit, for counties with weak participation in other programs. Thus, the goal of this evaluation is to improve the predictive power within the SAIPE production model. For the evaluation, we are only considering children ages 5-17 because school-age children are the most relevant age category for the SAIPE program. Section II provides background regarding the EITC program. Section III discusses the data used in this work and basic correlations in the data. Section IV presents results from estimation using the additional EITC variable. Section V concludes and discusses further research with the EITC information.

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

II. EITC Program

The Earned Income Tax Credit (EITC) is a refundable Federal income tax credit for low-income working individuals and families. A refundable credit allows eligible taxpayers to reduce or offset their federal tax liability, and receive any remaining credit beyond their tax bill as a refund. Holt's (2006) research on the history of the EITC program is the basis for this section. The EITC was authorized in 1975 by Congress to provide incentives to work. In 1986, the program was indexed for inflation so that the real value of the credit remained constant. Under the original legislation, all families with children received the same tax credit under the EITC legislation. Between 1994 and 1996, the legislation changed so that families received a higher dollar amount depending on the number of children (1 child or 2 or more children) and some individuals without children could qualify for the credit. For tax year 2002, legislation simplified rules and increased the earned income limit for married couples by $1,000, approximately 3%. In 2004, the definition of "qualifying child" for EITC benefits was simplified, although the expected overall effect is negligible. These cumulative changes in the legislation have made the EITC the largest U.S. income support program.

The changes subsequent to the tax year 2003 data used in this study that might impact the aggregate county observations are predominantly related to inflation indexing of the income limits. Thus, these changes will mostly change the scale, or average quantity of participants, while the correlation with overall poverty is not expected to change substantially in subsequent years. Reduced intertemporal variation compared to other administrative data sources, such as food stamps, might be a useful characteristic of the data for further study.

The dollar amount of the refundable credit depends on the number of "qualifying" children. Under current law, three criteria need to be met to have a qualifying child. These three criteria have undergone only minor changes since 1994. The child must be the taxpayer's child, foster child, sibling or a descendent of a child or sibling. The child needs to have lived with the taxpayer for at least half of the year. The child also has to meet one of the age definitions: less than 19 years old; less than 24 but in school full-time; or permanently disabled at any time of the year, regardless of age. Although some persons without qualifying children are eligible for EITC, the threshold for childless filers is low and the credit amounts are small.

The level of earned income and marital status also affects the amount of the credit. Within the qualification limits, the EITC is proportional to earned income, at a varying rate. The primary components of earned income considered for the EITC are wage and salary earnings and self-employment earnings. Within the phase-in income range for the credit, the EITC has a high subsidy per dollar of earnings, achieves a maximum credit amount, and then phases out by reducing the proportional rate. Within the phase-in range, the credit rates are 7.65 percent of earned income for filers with no qualifying children, 34 percent for one child, and 40 percent for two or more children. This rate is maintained until the credit reaches the maximum set by the tax law each year (which is also inflation adjusted). The maximum for married couples is set higher than for non-married workers.

Given the dollar amounts for 2005, the following example is based on a single parent with 2 children. During the phase-in of the credit, the first $11,000 of earned income receives a tax credit of 40 cents to the dollar. Between 11,000 and 14,000, the maximum credit amount is $4,400. During the phase-out period, this family would receive a declining total subsidy until being completely eliminated at $35,000. As a reference point, when this hypothetical family of 3 has income of $15,835 they are out of poverty, according to the official poverty thresholds used by the SAIPE program. Therefore, the population of tax filers that receive EITC benefits is a mixture of people above and below the official poverty threshold.

III. Data

The evaluation of the EITC data proceeds within the context of the current county-level poverty model for school-age children. Unless otherwise noted, "children" refers to children ages 5-17, excluding tax concepts. For tax-derived variables, the age range for children varies by program or concept. All the variables discussed are log-transformed in the regression. The dependent variable in the model is the number of related children in poverty, by county. For this dependent variable, the SAIPE program creates a 3-year weighted average of the CPS ASEC estimated number of people in poverty by county and age. For this research, the direct estimates of children in poverty are centered on the tax year of 2003. Most counties do not have any household interviews in the ASEC and thus cannot be used in fitting the regression model.[1]

The explanatory variables are all aggregated to the county-level and do not contain personal identifiers. Included are: population estimates of children under age 18 from the Population Division of the U.S. Census Bureau; poverty estimates of children based on Census 2000; total number of child exemptions claimed in the IRS data; total number of child exemptions in the IRS data that are in poverty; and total households receiving food stamps. The population and Census 2000 data are age-specific, under 18 and 5-17 related respectively. Child tax exemptions derive from an IRS definition of child tax deductions on the 1040 tax form, and thus do not have a strict age definition. Child tax-poor exemptions are tabulated child exemptions where the taxpayer(s) has an adjusted gross income below the poverty threshold for a household size indicated by the dependents on the return.

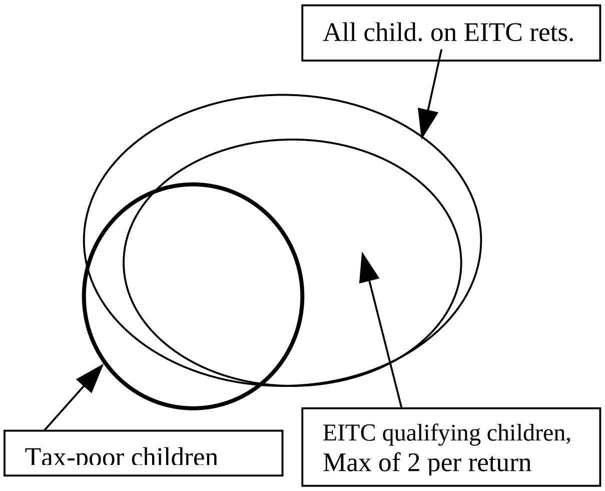


Figure 1: Relation of EITC Qualifying Children to Other Tax Concepts

In addition to the IRS data already included in the model, the EITC data will be tested and evaluated for its predictive power to improve the current model. The EITC data received is based on the IRS 1040 form. If claiming a child or children for the EITC, Schedule EIC must be completed. One or two children may be listed; the maximum credit amount allows for two children. Furthermore, the income threshold for EITC qualification is substantially higher than the poverty threshold for most family sizes. Therefore, child tax-poor exemptions are, on average, a subset of EITC qualifying children for tax units with two children or less. Any additional children may or may not be in the poverty universe as Figure 1 shows. Currently, we do not have data to show (1) the extent of overlap between EITC non-qualifying children[2] in (a) poor families and (b) non-poor families and (2) the percent of children in EITC tax units that are poor.

Not being able to differentiate how many children are in a tax unit that receives EITC is a drawback. Sixty-nine percent of EITC qualifying children are in the category of “2 or more”. This high percentage would be worrisome if the majority of EITC returns have 3 or more child exemptions. Examination of the 2003 data shows that only 10 percent of EITC returns have 3 or more child exemptions, which is acceptable for the purposes of this research.

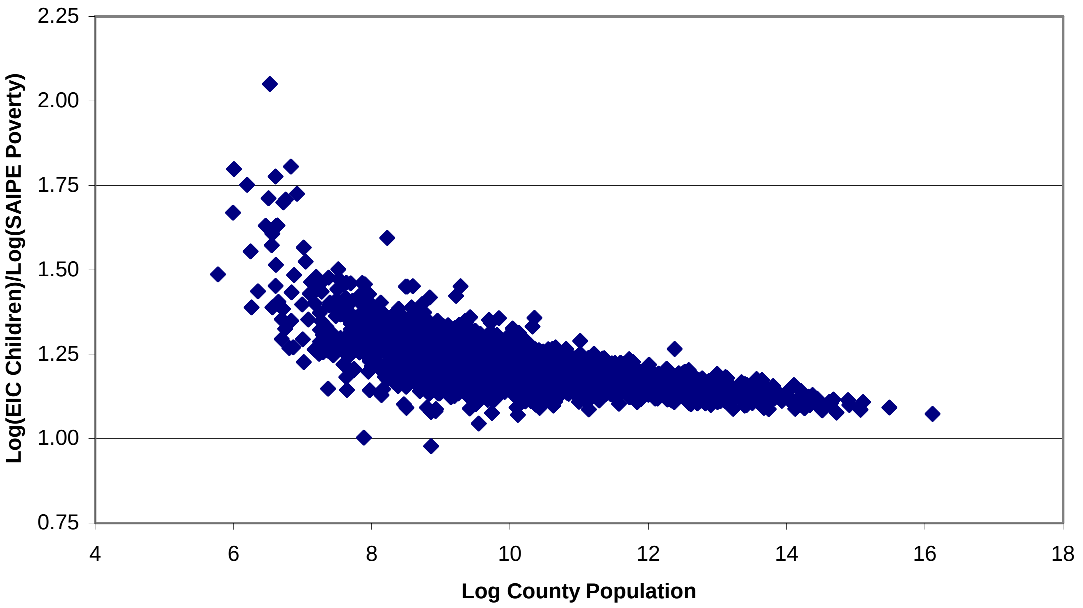
Although the EITC has identical rules and eligibility requirements nationally, the participation rate for EITC varies across cities and states (Berube 2006). This variation may be due to differential outreach efforts, the number of eligible filers differing between counties and states, as well as other county or state specific factors. This may be problematic for using EITC as a predictive variable. Chart 1 shows the number of EITC qualifying children as a ratio to the official SAIPE estimate of poverty for children ages 5-17, plotted against total population, all in logs. The variation in this ratio appears to decrease as county size increases, similar to trends seen in other concepts. The outliers among the smallest counties are not problematic for evaluation, since there is no CPS sample in these counties. The issue might require adjustments before utilizing for out-of-sample prediction, however.

Furthermore, there is a slight downward trend evident in the chart, which is exaggerated by the outliers in smaller counties. As discussed, we expect the difference between EIC children and poverty levels to be comprised mostly of children from households with incomes below the EIC cutoff, yet above the official poverty threshold. This lower income group should display similar trends to households in poverty, which also shows a negative correlation with population size.

[1] More details on the data and specification of the SAIPE program model can be found at http://www.census.gov/hhes/www/saipe/.

[2] EITC non-qualifying children are defined for this paper as child exemptions, which are not qualifiers for the EITC credit. These can be either children in excess of the maximum of 2 EITC qualifiers, or they may not be eligible for the credit qualification for other reasons, such as residency.

Chart 1: Comparison of EIC Qualifying Children to SAIPE Poverty Estimates



The focus of this research is improving the current model that already includes tax information. Holt (2006) estimates that only 35 percent of poor people participate in EITC because the elderly and taxpayers under age 25 are unable to participate unless they have qualifying children, many poor people have no taxable income, and the EITC qualifying threshold is very low for childless tax units. This implies that EITC may perform poorly in our model for predicting all poor but perform better for predicting child poverty. The added information between the number of poor and the number participating in EITC may be marginal, but useful, information.

We have two EITC concepts available for evaluation, level of EITC qualifying income for those filers that might qualify, and the number of EITC qualifying children for the same set of filers. The number of EITC qualifying children concept was chosen for evaluation, being analogous to the level concepts already contained in the model. We included this concept in the model in two alternate specifications. First, we simply include the EITC data as tabulated (Model A in Table 1). The second specification for the EITC variable is the number of EITC child participants less the number of poor child exemptions (Model B in Table 1). We interpret this new variable as mainly comprising the number of low-income, non-poor EITC filers, even though some poor exemptions will not be subtracted out due to large families (only the first two children are counted). This second specification eliminates most of the overlap between EITC qualified children and poverty exemptions, as illustrated in Figure 1.

IV. Methodology

The SAIPE methodology combines direct estimates from the survey with regression-based (indirect) estimates to produce final estimates. Fisher (1997) discusses the methodology for creating final estimates using a shrinkage estimator. This section will briefly discuss the regression and our methodology for evaluating the EITC data within the SAIPE production model. We are not interested in creating the final shrinkage-based estimates, since it is unnecessary for evaluating the effect of EITC data in the regression.

The specific class of regression used was an empirical best linear unbiased predictor (EBLUP). The specification is log-levels and the regression technique is a weighted least squares approach incorporating survey error and modeling error. Variance estimates for the error terms are iterated maximum likelihood. The weighted regression approach is necessary to control for the different CPS ASEC (Annual Social and Economic Characteristics supplement) sample sizes in each county (Asher & Fisher 2000), plus additional sources of heterogeneous variance effects. Equation 1 shows the functional form of the regression model.

$$y = X\beta + u + e \qquad (1)$$

where: y is a vector of the county estimate of the number of related children aged 5-17 in poverty from the ASEC, logged

X is the matrix for the explanatory variables in the regression, logged

β is the vector for the corresponding regression coefficients

u is the vector of model error

e is the error vector for the regression, otherwise referred to as sampling error

The National Research Council (Citro & Kalton 2000) evaluated the SAIPE county model along several criteria. Some of these evaluation criteria are relevant for our inclusion of a new variable to the regression model. The first step in this evaluation will look at individual regression coefficients to note the significance of new and existing coefficients and improvement to the overall fit of the model. When the new variable is included in the model, the coefficients of other variables change. The second step is examination of the changes for the coefficients of the original independent variables. For instance, the total child exemptions and the total child population are similar concepts and have closely related numbers. Thus, even if we find a significant model coefficient for the EITC concept, it might be at the expense of explanatory significance of other variables already in the model.

The third step is to see if the standardized residuals show any tendencies by the size of the counties and compare these tendencies to the original model as a base. This is to see if improvements are made using EITC data where other administrative data may be sparse. Furthermore, we check residuals against the direct estimates for the largest counties in the database, to ensure no degradation in prediction occurs at the higher end.

Passing all three steps of this evaluation is not always required, since if the alternative specification does particularly poor or well with respect to one criterion, the other criteria may be discounted. A coefficient for the newly included variable that is insignificant at the 5% level will cause rejection of the model from further evaluation. But even given a significant coefficient for the newly included variable, in general, the alternate specification will be rejected if 1) no significant improvement in R-squared occurs (under 0.005 at least) and 2) substantial changes occur to the other coefficients (either a change from significant to insignificant t-statistic, or 10% or more reduction in the actual value would be problematic) and 3) no reduction in small county outliers are observed. If strong improvements are noted as measured by at least one of these criterion, while little or no change registered by the other criteria, then the model would be proposed as an alternative for outside review. More complex results of one criterion reporting improvement and one degradation, such as a large increase in R-squared (over 0.05, for example), with a simultaneous loss of significance of one of the original explanatory variables, would be handled on a case-by-case basis.

Since EITC participation is highest and most stable for households with children, the EITC data is best correlated with child poverty. Thus, our evaluations focus exclusively on the school-age poverty model.

V. Results

Estimation results for the two alternate specifications including the EITC qualifying children concept in the SAIPE regression model are shown in Table 1, along with the official regression estimates for tax year 2003. As previously stated, the specification of this model is in log-levels. In the second alternate specification tax-poor exemptions are subtracted from EITC qualifying children before the logs are taken, so this specification cannot be interpreted as a simple difference of two explanatory variables, in terms of significance tests and similar evaluations. Rather the two alternates should be evaluated as two unique nonlinear specifications.

As seen in the table, the EITC concept shows minimal significance in both specifications, using an approximate critical level of 2.0. From a strictly predictive performance standpoint, even small levels of significance would warrant inclusion of the new concept. Such a criterion is highlighted by the similarly small values for t-statistics on other variables in the original model.

The R-squared value displays a very small improvement, analogous to the small T-statistic value. This R-squared was calculated with the usual mean-corrected definition of explained sum of squares divided by total sum-of-squares. In this weighted regression context, a strict equivalence with an F-type significance test does not exist, due in part to the iterated estimates for the variance weights.

Table 1: Regression Coefficients for County-Level Poverty Model Ages 5-17, Calendar Year 2003

Variable Description*	Original Model	EITC Model A	EITC Model B
Constant	-0.88 (5.8)	-0.99 (6.3)	-0.77 (4.9)
All children	0.79 (2.2)	0.77 (2.2)	0.74 (2.1)
All children in 1999 poor households, from Census 2000	0.20 (2.4)	0.36 (2.6)	0.19 (2.2)
Persons receiving food stamps	0.20 (3.9)	0.14 (2.6)	0.15 (2.8)
All child tax exemptions	-0.67 (2.0)	-0.69 (2.1)	-0.66 (2.0)
Child tax exemptions in poor households	0.50 (4.4)	0.26 (1.8)	0.50 (4.4)
EITC Qualifying Children	—	0.18 (2.1)	—
EITC Qual. Children less child tax-poor exemptions	—	—	0.09 (2.0)
R-Squared	0.772	0.774	0.773

T-statistics in parentheses.
Counties in sample : 1233
*Dependent & explanatory variables are in logs

In terms of comparing the two alternate specifications for the inclusion of the EITC concept, the R-squared value favors the unaltered concept (Model A), but by a likely insignificantly small amount. Examining the concept's effect on the original explanatory variables, the second alternate specification (Model B) is preferred. Particularly in regards to the original tax-poor term, the coefficient value is barely altered in this alternative. In Model A, the significance of the tax-poor variable is reduced by a substantial degree, and thus clearly fails criterion two outlined above. Correlation between the EITC concept and the tax-poor variable is 97% in this specification, while it is below 90% in the second. Note that since logs are taken after the difference, Model B is not an exact algebraic alteration of the first. Also note that preference for this altered concept is based solely on interpretation of the original explanatory variables in the model. Based on purely predictive power, that is, the minimum squared residuals represented by R-squared, the Model B would be preferred.

For this second alternative (Model B), all original explanatory variables remain significant, but note the large reduction in the value, or marginal effect, of the food stamp coefficient. This 25% reduction in the marginal effect represents a clear failure of the second criterion for evaluation. Without a substantial improvement in small county outliers, this model does not appear to be a candidate for further review.

Since there is no compelling motivation for inclusion of the EITC variable evident in the goodness of fit, we undertook a closer evaluation of the alternate models' ability to predict poverty in the smaller counties. Administrative data, such as food stamps, for smaller counties frequently show higher volatility due to differences in data collection processes and the higher impact outlier households can have on smaller counties. As discussed previously, it is hoped that uniform data processing and eligibility standards would reduce such volatility for EITC compilations, when compared to other input data.

To examine this possibility, standardized residuals, or z-scores, were calculated for each specification, defined as the residual divided by the standard error of the regression. A simple plot of these z-scores versus county population showed no discernible difference between specifications, which was expected, given the nearly identical R-squared values. Tallying outlier counties also produced no substantial differences, as the number of counties with z-scores over 2.0 were identical for all specifications, and the number with z-scores over 3.0 were only one less for the second alternative. No differences among population groups were evident.

Finally, the residuals were examined for the ten largest counties, which comprise the counties of sufficient size to provide a reasonable direct estimate from the CPS ASEC. In these counties, the current official version of the SAIPE prediction model was superior to either EITC specification, as the residuals were smaller for the SAIPE model for every county. This seems odd, given the R-squared advantage of the alternative specification, but the residual differences were not significant. Furthermore, the R-squared advantage indicates the average residuals were lower across all 1,233 counties, not for any given group of counties. So the SAIPE model performs better for the 10 largest counties, and the alternatives are better over some group of smaller counties.

VI. Conclusion

From a purely predictive standpoint, inclusion of the EITC concept cannot be rejected. The t-statistics shown in Table 1 are approximately valid for large samples, and thus the EITC variable is weakly significant in the two alternative specifications. Even with the difference transformation of the concept entailed in Model B, however, the marginal effect of the food stamp variable is reduced by 25%. For Model A, there is even more serious degradation in the significance of the original explanatory variables.

Reviewing the three criteria we set up for evaluating these models, both alternatives have inconclusive results for the goodness of fit criterion, with negligible increases in R-squared. Both alternatives displayed no reduction in small county outliers. Both models showed substantial reductions in either the significance or marginal effect of the original explanatory variables. So there is clear failure on the two criteria of integrity of the original model, and small county outliers, with weak failure to reject for the goodness of fit criterion. The conclusion is these alternatives are not candidates for further review, given this evaluation based on one year of data.

One substantial shortcoming of this evaluation is the inability to evaluate the predictive power for counties outside of the CPS sample. In previous model evaluations, the poverty levels obtained from the decennial census were utilized as the benchmark. But we currently have EITC data available for tax year 2003 only. Comparisons with previous-year official estimates, or the 2000 Census values, were examined, but eventually rejected as being difficult to interpret given the numerous additional sources of intertemporal variation. A second evaluation will be performed, however, when county-level data from the American Community Survey become available.

Another source of potential improvement would be through reduced intertemporal volatility for the EITC potential participation concept, particularly in comparison to other administrative data, and thus more meaningful predictors. Once additional years of EITC data are available, such evaluation will proceed. At this time, our evaluation concludes that there is insufficient improvement in single-year predictive power to motivate alteration of the official model. Additional years of data should be evaluated jointly before the possibility of including the EITC data is rejected conclusively.

VII. References

Asher, J. & Fisher, R. (2000), “Alternate CPS Sampling Variance Structures for Constrained and Unconstrained County Models: Additional Research,” Small Area and Income and Poverty Estimates Program, U.S. Census Bureau, Technical Report.

Berube, A. (2006), “The New Safety Net: How the Tax Code Helped Low-Income Working Families During the Early 2000s,” Brookings Institution, Research Brief.

Citro, C. & Kalton, G. (2000), *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*, National Academy Press.

Fisher, R. (1997), “Methods Used for Small Area Poverty and Income Estimation,” *Proceedings of the Section on Government Statistics*, American Statistcial Association, pp. 177-182

Holt, S. (2006), “The Earned Income Tax Credit at Age 30: What We Know,” Brookings Institution, Research Brief.