SAIPE County Poverty Models Using Data from the American Community Survey *

David Powers, Wesley Basel, and Brett O'Hara; david.s.powers@census.gov
U.S. Census Bureau, Small Area Estimates Branch; Washington, DC 20233

Abstract
The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program produces model-based estimates of income and poverty using data from Census 2000, the American Community Survey (ACS), administrative records, and intercensal population estimates. This work assesses SAIPE county poverty models under log rate and log count data transformations. Equivalent rate and count models are described as the basis for comparison, which differ only in their corresponding dependent variables. Scale invariance and homoskedasticity are assessed, and "goodness of fit" comparisons are made in terms of root mean square error (RMSE). The estimation results support use of the log count model.

Key Words: poverty, SAIPE, ACS, Fay-Herriot, Census, small area

1. Introduction

The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program annually produces estimates of the number of related children ages 5 to 17 in poverty for states, counties, and school districts. The SAIPE estimates improve upon direct survey estimates of poverty by borrowing strength from administrative records, intercensal population estimates, and decennial census data.

Starting with the 2005 SAIPE estimates, the SAIPE program switched to using data from the American Community Survey (ACS) in place of data from the Annual Social and Economic Supplement of the Current Population Survey (CPS ASEC) in the dependent variables of its models. Due to the much larger sample size and much wider county coverage of the ACS, variances of the SAIPE estimates were generally reduced. This switch to the ACS was made after an external review that took place in September of 2007. Documentation for this change is available online at *http://www.census.gov/hhes/www/saipe/techdoc/methods/05change.html*. An extensive report discussing the change is at *http://www.census.gov/hhes/www/saipe/techrep/report.pdf*.

The dependent variable of the SAIPE county production model is the logarithm of the number of related children ages 5-17 in poverty. This concept will be referred to as "5-17 related poverty" from here forward. Accordingly, the independent variables (i.e., explanatory variables, predictors or regressors) are also in logarithm form. We refer to this model as the "log-level" model. Another general type of model considered in the original SAIPE county model evaluations was the "log-rate" model (NRC 1998, page 46). In this model, the dependent variable is the logarithm of a direct survey estimate of the county poverty rate of related children ages 5-17. In its basic form, the explanatory variables are also log-rate. One theoretical advantage of using rate transformations is greater stability of input data ratios in situations in which levels vary over a wide range or in which levels are difficult to measure.

This work assesses estimation results of log-level models relative to those of log-rate models. Both types of models reflect the general form suggested by Fay and Herriot (1979). The model is specified as:

$$(1)\ \log(y_i) = \log(Y_i) + e_i \text{ where } e_i \sim ind.\ N(0, \mathrm{v}_i)$$

$$(2)\ \log(Y_i) = \log(x_i')\beta + u_i;\ \ u_i \sim i.i.d.\ N(0, \sigma_u^2),$$ where, for county *i*,

y_i = ACS direct estimate of 5-17 related poverty (number in poverty or poverty rate)
Y_i = true population value of 5-17 related poverty (number in poverty or poverty rate)
$e_i = \log(y_i) - \log(Y_i)$ = sampling error in $\log(y_i)$ as an estimate of $\log(Y_i)$
x_i = matrix of regression variables
β = vector of regression parameters
u_i = random model error (county random effect).

* *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.*

Different versions of the model (log-rate or log-level) are defined by choosing y_i to be either the ACS estimate of the county number in poverty or the county poverty rate and by the corresponding different predictors contained in x_i.

The paper proceeds as follows. *Section 2* reviews the SAIPE model input data, and *Section 3* discusses model forms. *Sections 4* and *5* present the main estimation results and discuss goodness of fit. Concluding remarks are included in *Section 6*.

2. Model Inputs

Table 1 below lists definitions of the variables included in the log-level and log-rate models. “LL6” is the log-level model with six predictors (including the intercept); this is the SAIPE production model. “LR6” is the log-rate model with six predictors; “LL8” is the log-level model with eight predictors; and “LR8” is the log-rate model with eight predictors. The motivation for the eight-predictor models is that they contain identical predictor information, which facilitates the comparison. These models are described further in Section 3.

Log-Level Inputs			*Log-Rate Inputs*	
Short Name	Description		Short Name	Description
Dependent Variable			Dependent Variable	
Log (ACS number in poverty, ages 5-17 related)	Log estimated county number of 5-17 related children in poverty from the 2005 ACS.		Log (ACS poverty rate, ages 5-17 related)	Log estimated county poverty rate of 5-17 related children from the 2005 ACS.
Predictors for "LL6" Model			Predictors for "LR6" Model	
Log (IRS child tax-poor exemptions)	Log number of county tax-poor child exemptions from IRS administrative records, where tax-poor is defined as Adjusted Gross Income (AGI) below the poverty level for a household size defined by the total number of exemptions on the return.		Log (IRS child tax-poor exemption rate)	Log number of county tax-poor child exemptions divided by total child exemptions, from IRS administrative records.
Log (Food Stamp Program participants)	Log number of county Food Stamp Program participants reported in July (data from the USDA Food and Nutrition Service), raked to a control total obtained from state Food Stamp participant data.		Log (Food Stamp rate)	Log number of county Food Stamp Program participants divided by total all-ages population from the PEP intercensal demographic estimates.
Log (PEP population, ages 0-17)	Log county population, ages 0-17, as of July 1, 2005, from the Census Bureau‘s Population Estimates Program (PEP) of intercensal demographic estimates.		Log (PEP population, ages 0-17)	Log county population, ages 0-17, as of July 1, 2005, from the Census Bureau’s Population Estimates Program (PEP) of intercensal demographic estimates.
Log (IRS child tax exemptions)	Log total number of county child exemptions from IRS administrative records.		Log (IRS child filing rate)	Log total number of county child exemptions from IRS administrative records divided by county ages 0-17 PEP intercensal demographic estimate.
Log (Census 2000 poor, ages 5-17 related)	Log estimated county number of 5-17 related children in poverty from Census 2000. Income reference is 1999.		Log (Census 2000 poverty rate, ages 5-17 related)	Log estimated county poverty rate for 5-17 related children from Census 2000. Income reference is 1999.
Additional Predictors for "LL8" Model			Additional Predictors for "LR8" Model	
Log (PEP population, all ages)	Log county population, all ages, as of July 1, 2005, from the PEP intercensal demographic estimates.		Log (PEP population, all ages)	Log county population, all ages, as of July 1, 2005, from the PEP intercensal demographic estimates.
Log (Census 2000 poverty universe, ages 5-17 related)	Log number of county 5-17 related children from Census 2000.		Log (Census 2000 poverty universe, ages 5-17 related)	Log number of county 5-17 related children from Census 2000.
Further information about these input data is available on the SAIPE program’s webpage, http://www.census.gov/hhes/www/saipe/techdoc/inputs/datintro.html				

Table 1: Variable definitions for the log-level (left side) and log-rate (right side) models

The American Community Survey (ACS) is a nationwide survey designed to provide communities timely information about how they are changing. The yearly ACS estimates combine results from twelve monthly surveys (starting in January and ending in December), each referring back twelve months, to span a total of twenty-three months of reference. For the 2005 ACS, the twelve-month spans start as early as January 2004 and end as late as November 2005. Also, the 2005 ACS utilizes population controls from the U.S. Census Bureau’s Population Estimates Program (PEP) for July 1, 2005.

All 3,141 U.S. counties are covered in the ACS. Single-year ACS estimates are publicly available for counties with populations of 65,000 and larger, which corresponds to roughly 775 out of the 3,141 U.S. counties. Unpublished ACS estimates exist for the remaining counties, and these counties are also used in the model fitting to follow. For some counties with small samples, the estimate of the number of related children ages 5-17 in poverty is zero by random chance. Since logs cannot be taken of these estimates, such counties are excluded from the model fitting. This led to the exclusion of 166 counties from the modeling. A further 3 counties were omitted due to having poverty rate estimates of 100% (which leads to direct variance estimates of zero given the population controls). Thus, all reported results are based on 2,972 county observations.

Sampling error variances for the county log number and log rate in poverty are estimated directly using successive difference replication (Fay and Train, 1995), using replicate weights included in the survey micro-data file. The inclusion of covariates for IRS child tax exemptions and PEP population (ages 0-17) together in the model can be interpreted as providing a proxy for a log “tax filers rate” [= log (IRS child tax exemptions) – log (PEP population, ages 0-17)], for which a negative relation to poverty is expected.

3. Models and Restrictions

The SAIPE production model is of log-level form and contains six predictors (including the intercept). This is the "LL6" model in Table 1. A similar log-*rate* model would include rate versions of the same predictors and also a scale term such as population size. This is the "LR6" model in Table 1. However, the LL6 and LR6 models are not directly comparable since some of the rate predictors are created with denominators not included as level predictors in the corresponding level models. For example, the Census 2000 poverty *rate* from the rate model contains both data on poverty (the numerator) and data on the poverty universe (the denominator), whereas the Census 2000 poverty *level* from the level model contains only data on poverty.

In order to make the log-level and log-rate models comparable, we include as predictors two additional denominator terms corresponding to the remaining log-level predictors. These are the LL8 and LR8 models from Table 1. To illustrate, consider the terms related to Census 2000 poverty, as below:

$$\beta_1 \log(\text{CenPov}) + \beta_4 \log(\text{CenPovUniv}) \;=\; \beta_1\{\log(\text{CenPov}) - \log(\text{CenPovUniv})\} + (\beta_4 + \beta_1)\log(\text{CenPovUniv})$$

The expression on the left of the equal sign is in log-level form, and the expression on the right of the equal sign is in log-rate form (since log(y/x)=log(y)-log(x)). A similar example would apply to the IRS tax-poor ratio and its corresponding tax-poor exemptions (the numerator), and tax exemptions (the denominator). Applying this kind of structure for all predictors, the right-hand-side data for the eight-predictor rate and log models would be logically equivalent, despite their different appearance. As a result, the predicted values and residuals would be invariant to whether the explanatory variables are expressed in log-rate (LR) or log-level (LL) form. Only a redefinition of the beta coefficients occurs. A more general statement of this result follows below.

Since the log-rate form involves a linear combination of log(numerator) and log(denominator) terms, the log-rate predictors can be derived as a simple linear transformation of the log-level predictors. Consider the general linear model: $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{u}$ where $\mathbf{X}$ is an *n* x *k* matrix including a constant column. Define a linear transformation matrix, $\mathbf{A}$ *k* x *k* (invertible), with full rank such that the inverse exists. Then consider the general model:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}\,, \quad \text{where } \mathbf{u} \mid \mathbf{X} \sim (\mathbf{0}, \Sigma)$$

Applying the transformation to the right-hand-side matrix:

$$\mathbf{Y} = \mathbf{X}\mathbf{A}\mathbf{A}^{-1}\beta + \mathbf{u} = \mathbf{X}_A\beta_A + \mathbf{u}\,, \quad \text{where } \mathbf{u} \mid \mathbf{X}_A \sim (\mathbf{0}, \Sigma)$$

The transformation has no effect on the conditional distribution of the error term, but does represent a redefinition of the parameters. Any estimators based on the cross-correlation matrices, such as ordinary least squares, weighted least squares, maximum likelihood estimation under normality, etc., will be similarly affected only through this linear transformation. The weighted least-squares coefficient vector, for example, is:

$$\begin{aligned}\tilde{\hat{\beta}}_{WLS} &= \left(\mathbf{X}'_A\Sigma^{-1}\mathbf{X}_A\right)^{-1}\mathbf{X}'_A\Sigma^{-1}\mathbf{Y} = \left(\mathbf{A}'\mathbf{X}'\Sigma^{-1}\mathbf{X}\mathbf{A}\right)^{-1}\mathbf{A}'\mathbf{X}'\Sigma^{-1}\mathbf{Y} \\ &= \mathbf{A}^{-1}\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{A}'^{-1}\mathbf{A}'\mathbf{X}'\Sigma^{-1}\mathbf{Y} = \mathbf{A}^{-1}\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y} = \mathbf{A}^{-1}\hat{\beta}_{WLS}\end{aligned}$$

And the predicted values and residuals will be identical:

$$\begin{aligned}\hat{\tilde{\mathbf{Y}}}_{WLS} &= \mathbf{X}_A\left(\mathbf{X}'_A\Sigma^{-1}\mathbf{X}_A\right)^{-1}\mathbf{X}'_A\Sigma^{-1}\mathbf{Y} = \mathbf{XA}\left(\mathbf{A}'\mathbf{X}'\Sigma^{-1}\mathbf{XA}\right)^{-1}\mathbf{A}'\mathbf{X}'\Sigma^{-1}\mathbf{Y}\\ &= \mathbf{XAA}^{-1}\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{A}'^{-1}\mathbf{A}'\mathbf{X}'\Sigma^{-1}\mathbf{Y} = \mathbf{X}\left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y} = \hat{\mathbf{Y}}_{WLS}\end{aligned}$$

This mapping between the log-level and log-rate predictors is exact with eight predictors. The mapping with six predictors is inexact since in that case the denominators for two of the log-rate predictors do not enter the log-level predictors. The hypothesis tests between the six-predictor (LL6) and eight-predictor (LL8) models discussed in Section 4 provide a formal evaluation of the six- and eight-predictor models.

When modeling across counties, we typically ask if there is an effect of county size on the prediction. In other words, are the results "scale invariant?" More precisely, "Do the scale components in the set of X data, such as population size or poverty universe size, have a net effect on the poverty rate prediction? For the log-level model, the restriction implied by scale invariance is that the estimated coefficients on all non-intercept log predictors sum to one. For the log-rate model, the required condition for scale invariance is that the sum of the estimated coefficients on any non-intercept log scale predictors, like population or poverty universe, sum to zero. A proof of the invariance condition for log-level can be found in Appendix B.2 of (Bell, et al., 2007).

We will denote the restricted log-level models LL6r and LL8r, and the unrestricted log-level models as LL6u and LL8u. We will denote the restricted log-rate models LR6r and LR8r, and the unrestricted log-rate models as LR6u and LR8u. For the LR6 model, this reduces to simply having the coefficient on log(PEP population, ages 0-17) be zero (as shown in the table for LR6r).

4. Estimation

We estimate the four model forms described in Section 3 (i.e., the 6r, 6u, 8r, 8u forms) for both the log-level and log-rate models. Model estimation is via an iterated weighted least-squares algorithm, with Newton-Raphson updating for σ_u^2. Convergence occurs in fewer than 10 iterations in all cases. The observation weights are $(\mathrm{v}_i + \sigma_u^2)^{-1}$, i.e., the inverse of the log total error variance (sum of the sampling and model variances). The v_i is estimated from successive difference replication (see Section 2) and is assumed known.

4.1 Log-Level Model

Table 2 displays regression results for all four versions of the log-level model considered. Restricted and unrestricted refer to the scale invariance restrictions discussed in Section 3.

Predictor		Model			
		LL6r	LL6u	LL8r	LL8u
Intercept	β_0	-0.394 (0.041)	-0.421 (0.057)	-0.218 (0.095)	-0.256 (0.104)
Log (IRS child tax-poor exemptions)	β_1	0.542 (0.044)	0.548 (0.045)	0.544 (0.047)	0.563 (0.052)
Log (Food Stamps)	β_2	0.173 (0.022)	0.173 (0.022)	0.181 (0.023)	0.180 (0.023)
Log (PEP population, ages 0-17)	β_3	1.082 (0.112)	1.050 (0.122)	1.118 (0.113)	1.069 (0.127)
Log (IRS child tax exemptions)	β_4	-1.064 (0.106)	-1.037 (0.114)	-1.062 (0.142)	-1.073 (0.143)
Log (Census 2000 poor, ages 5-17 related)	β_5	0.268 (0.030)	0.268 (0.030)	0.250 (0.035)	0.242 (0.037)
Log (PEP population, all ages)	β_6			-0.122 (0.061)	-0.117 (0.061)
Log (Census 2000 poverty universe, ages 5-17 related)	β_7			0.091 (0.097)	0.141 (0.114)
Degrees of freedom		2,967	2,966	2,965	2,964
Model error variance		0.0216	0.0217	0.0209	0.0211
Akaike Information Criterion (AIC)		4100.3	4101.8	4099.8	4101
Maximum log likelihood		-2044.14	-2043.9	-2041.91	-2041.52
Sum of slopes		1	1.0036	1	1.0055

Table 2: Regression prediction results for four versions of the log-level model. Standard errors are in parenthesis.

Table 3 displays likelihood ratio test results for all the comparisons implied by the four models in Table 2. Asymptotically, the likelihood ratio statistic should be distributed as chi-squared under the respective null hypotheses with the degrees of freedom shown. Wald and Lagrange multiplier tests for these comparisons were also examined and did not yield any different conclusions.

Models	H_0	Likelihood	Degrees of	Critical Value	
		Ratio	Freedom	(α=0.10)	(α=0.05)
LL6r vs. LL6u	$\Sigma\beta_i = 1;\ i \neq 0$	0.48	1	2.71	3.84
LL8r vs. LL8u	$\Sigma\beta_i = 1;\ i \neq 0$	0.78	1	2.71	3.84
LL6r vs. LL8r	$\beta_6 = \beta_7 = 0$	4.46	2	4.61	5.99
LL6u vs. LL8u	$\beta_6 = \beta_7 = 0$	4.76	2	4.61	5.99
LL6r vs. LL8u	$\beta_6 = \beta_7 = 0$ and $\Sigma\beta_i = 1;\ i \neq 0$	5.24	3	6.25	7.82

Table 3: Likelihood ratio tests for the log-level models

In Table 3, the tests of LL6r versus LL6u and LL8r versus LL8u examine the imposition of the scale invariance restriction. The tests of LL6r versus LL8r and LL6u versus LL8u examine the omission of the two additional scale or denominator terms listed towards the bottom of Table 1. The motivation for these additional predictors is discussed in Section 3. The final test in Table 3 is for the joint hypothesis that the additional denominator terms are not needed and, simultaneously, that the scale invariance restriction holds.

The likelihood ratio tests fail to reject four of the five tested null hypotheses. For the test of the reduced predictor set with unrestricted coefficients (LL6u vs. LL8u), the null hypothesis is rejected at the 10% significance level but still not the 5% level. The ordering of these results is nearly coincident with the AIC ordering in Table 2, since for these simple linear models there is little difference between the likelihood ratio and the AIC formula.

In summary, there is little preference between the six- and eight-predictor models, and the null hypothesis of scale invariance is not rejected. Note the test results do not reject the model with six explanatory variables and the scale invariance restriction on the coefficients (LL6r). However, any variance advantage under LL6r will be relatively minor, as can be seen by comparing the standard errors of the regression coefficients between LL6r and LL6u.

4.2 Log-Rate Model

Tables 4 and *5* repeat the tests of restrictions discussed in Section 3, but now for the log-rate model. The same derivation of the equivalence of log-level versus log-rate predictors that motivated the eight-predictor models for log-level can be applied to models with the log-rate dependent variable. Note the eight-predictor model with the log-rate dependent variable (LR8 in Table 4) is similar in concept to the "hybrid models" examined in the second National Academy of Sciences panel interim report evaluating county and school district estimates for Title I allocations (NRC 1998, page 101).

The tests of LR6r versus LR6u and LR8r versus LR8u examine the imposition of the scale invariance restriction, namely, that the sum of the denominator term coefficients equals zero. (See discussion in the latter part of Section 3.) The tests of LR6r versus LR8r and LR6u versus LR8u examine the omission of the two additional scale or denominator terms listed in Table 4. The motivation for these additional predictors is discussed in Section 3. The final test in Table 5 is, again, for the joint hypothesis that the additional denominator terms are not needed and, simultaneously, that the scale invariance restriction holds.

Predictor		Model			
		LR6r	LR6u	LR8r	LR8u
Intercept	β_0	0.235 (0.031)	0.372 (0.048)	0.154 (0.094)	0.242 (0.102)
Log (IRS child tax-poor exemption rate)	β_1	0.624 (0.046)	0.556 (0.049)	0.623 (0.047)	0.576 (0.052)
Log (Food Stamp rate)	β_2	0.166 (0.022)	0.167 (0.022)	0.163 (0.022)	0.167 (0.022)
Log (PEP population, ages 0-17)	β_3	0	-0.020 (0.005)	-0.321 (0.086)	-0.222 (0.098)
Log (IRS child filing rate)	β_4	-0.255 (0.109)	-0.414 (0.117)	-0.507 (0.128)	-0.521 (0.128)
Log (Census 2000 poverty rate, ages 5-17 related)	β_5	0.272 (0.035)	0.289 (0.035)	0.253 (0.035)	0.273 (0.036)
Log (PEP population, all ages)	β_6			0.090 (0.061)	0.079 (0.061)
Log (Census 2000 poverty universe, ages 5-17 related)	β_7			0.232 (0.083)	0.130 (0.097)
Degrees of freedom		2,967	2,966	2,965	2,964
Model error variance		0.0319	0.0303	0.0306	0.0300
Akaike Information Criterion (AIC)		3866.4	3854.8	3856.5	3854.2
Maximum log likelihood		-1927.2	-1920.4	-1920.3	-1918.1
Sum of denominators		0	-0.0201	0	-0.0135

Table 4: Regression prediction results for four versions of the log-rate model. Standard errors are in parentheses.

Models	H_0	Likelihood	Degrees of	Critical Value	
		Ratio	Freedom	(α=0.10)	(α=0.05)
LR6r vs. LR6u	$\beta_3 = 0$	13.62	1	2.71	3.84
LR8r vs. LR8u	$\beta_3 + \beta_6 + \beta_7 = 0$	4.32	1	2.71	3.84
LR6r vs. LR8r	$\beta_6 = \beta_7 = 0$	13.88	2	4.61	5.99
LR6u vs. LR8u	$\beta_6 = \beta_7 = 0$	4.58	2	4.61	5.99
LR6r vs. LR8u	$\beta_3 = \beta_6 = \beta_7 = 0$	18.2	3	6.25	7.82

Table 5: Likelihood ratio tests for the log-rate models

In Table 5, four of the five null hypotheses are rejected. The exception is the borderline result when comparing the unrestricted, six-predictor model to the unrestricted, eight-predictor model. The ordering of these results is, again, nearly coincident with the AIC ordering, which shows preference for the two unrestricted models (LR6u and LR8u) over the two restricted models (LR6r and LR8r). In summary, there is again no clear favorite between the six-predictor and eight-predictor models. However, the null hypothesis of scale invariance is rejected, suggesting a possible scale effect in the log-rate model that was not seen in the log-level model.

5. Discussion of Results

The tests discussed in Section 4 suggest that the unrestricted six-predictor models for both log-level (LL6u) and log-rate (LR6u) perform as well as the alternatives with scale restrictions or with the two additional predictors. We now discuss these LL6u and LR6u estimation results further. The following text refers explicitly back to the "LL6u" column of Table 2 and to the "LR6u" column of Table 4.

5.1 Coefficient Estimates

For the LL6u log-level model (see Table 2), the coefficient estimates on all predictors bear the expected sign, with the coefficient estimates positive except for the coefficient on the IRS child tax exemptions variable, which is negative as expected. The results can be summarized as follows: (i) the coefficient estimates are each individually highly statistically significant, (ii) the log (IRS child tax-poor exemptions) variable is the most important (highest t-statistic), and (iii) the sum of the coefficients on the log (PEP population, ages 0-17) and log (IRS child tax exemptions) variables is not statistically different from zero. This last point reinforces the interpretation of the net effect of these two variables as relating to the effect of a log (tax filer rate) variable.

Most of the results in the LR6u log-rate model (see Table 4) are analogous to those for the LL6u log-level model. First, the regression coefficient estimates are all statistically significant and all bear the expected sign, with a negative coefficient for the log (IRS child filing rate) variable. Second, the most important variable (highest t-statistic) is that for the log (child tax-poor rate). Since the fundamental difference between the log-rate models and the log-level models is the denominator (PEP population, ages 0-17) in the dependent variable, the significance of this coefficient provides some evidence against a pure rate model (such as LR6r and LR8r).

The correlation matrix for the coefficient estimates in the log-level model shows that the correlation for the coefficients on the IRS child tax exemptions and the PEP population, ages 0-17, variables is nearly negative one. Since the estimated values are equal and have opposite signs as well, this is further evidence that the two terms could be combined without appreciably changing the model predictions or standard errors. The correlation matrix for the coefficient estimates in the log-rate model contains a pattern of negative correlations somewhat similar to that for the log-level model, but without any of the correlations approaching negative one. These correlation matrices are shown explicitly in Tables 2.4 and 2.11 of Bell, et al. (2007).

5.2 Diagnostic Checking

Box-whisker plots and scatter plots of the standardized residuals were examined relative to various classification variables, and no systematic patterns were observed. The following classification variables were considered: (1) Census 2000 total population, (2) Census 2000 percent in poverty, (3) Population growth Census 1990 – Census 2000, (4) Population growth Census 2000 – 2005 PEP total population, (5) Census 2000 percent Black, (6) Census 2000 percent Hispanic, (7) Census 2000 percent Asian, and (8) Census 2000 percent group quarters. The box-whisker plots group the categorization variables into quintiles, each with 594 counties of the 2,972 modeled counties. These plots are available in Appendix B of Bell, et al. (2007).

As a test for heteroskedasticity, Spearman's rank correlation coefficients are computed between the squared standardized residuals and four covariates: 2005 PEP total county resident population, all ages (log); 2005 ACS sample size, household count (log); Census 2000 poverty rate, all ages; and Census 2000 percent rural. For the log-level model, the Spearman test statistics confirm the general patterns, or rather lack of patterns, seen in the standardized residual plots. These tests fail to reject the assumption of homogeneous variances along the directions of any of these covariates. However, under the log-rate model, some small, but statistically significant, heterogeneity is detected. It is not an induced problem caused by the reduced list of predictors; the same significant heterogeneity is found for the model with eight predictors (LR8u). The tables of Spearman test statistic values are reported in Bell, et al. (2007): Table 2.5 for log-level, and Table 2.12 for log-rate.

The Spearman's test is a low-powered robust test. Any inaccuracies in the sampling error variances can be mixed in with the estimate of model error variance. We are currently researching improvements to the ACS sampling error variance estimates used through modeling. We forgo higher-powered tests for heteroskedasticity until further research on direct sampling error variance estimates is completed.

5.3 Goodness of Fit

This section continues discussion of the LL6u and LR6u benchmark results, but now with respect to goodness of fit. These models are defined in Section 3. Although the log-rate models have a similar structure to log-level models, direct statistical inferences about comparisons of log-level and log-rate models are not immediate since the two types of models involve different dependent variables. For log-level models, the dependent variable is log (y_i), whereas for log-rate models it is log (y_i)−log (z_i), where z_i is the 2005 ACS direct survey estimate of the 5-17 poverty universe for county *i*.

One approach to comparing log-level and log-rate models would be to construct a simultaneous (bivariate) model of log (y_i) and log (z_i) under which the LL6u and LR6u models (or LL8u and LR8u models) would be nested. However, developing a model for the log poverty universe estimates (log (z_i)) would be somewhat extraneous to the present focus. We instead compare model predictions from the log-level and log-rate models by translating the respective results to a common data transformation. We do so by converting log-rate regression predictions to corresponding predictions of number in poverty by adding an estimate of log(z_i) to the log-rate predictions, and by converting log-level regression predictions to corresponding predictions of log-rate by subtracting log(z_i) from the log-level predictions. We then use these results to assess goodness of fit. The statistic used to compare goodness of fit of the two models is root mean square error (RMSE). The errors in this case are the regression residuals from the fitted models. Comparisons between log-level and log-rate models are made in both a level scale and a rate scale.

From the test results in Sections 4.1 and 4.2, the log-level and log-rate models are comparable under the eight-predictor, unrestricted versions, and are nearly comparable for the corresponding six-predictor versions. Results from the six-predictor models are presented below, as these correspond more closely to the current SAIPE production model. *Table 6* displays RMSEs for the two six-predictor models. This is equivalent to the square root of the unweighted residual mean square for each model. RMSEs are reported for the entire set of observations, as well as for two partitions of counties – one by population size and one by Census 2000 poverty rates – in order to gauge sensitivity of the results. For population size, the partitions are: less than 20,000; 20,000 to 64,999; and 65,000 and over. For Census 2000 poverty rate, the partitions are less than 12.5%; 12.5% to 19.9%; and 20% and over.

RMSE, by model	Comparison on the log-level scale		Comparison on the log-rate scale	
	LL6u	LR6u + log (z_i)	LL6u − log (z_i)	LR6u
Full-sample: 2,972 counties	0.638	0.650	0.597	0.609
Partition by 2005 PEP county total resident population				
> 65k: 775 counties	0.321	0.326	0.317	0.323
20k – 65k: 1,032 counties	0.547	0.556	0.531	0.540
< 20k: 1,165 counties	0.839	0.857	0.770	0.786
Partition by Census 2000 poverty rate, all ages				
> 20%: 936 counties	0.617	0.634	0.547	0.562
12.5 – 20%: 1,000 counties	0.605	0.618	0.561	0.573
< 12.5: 1,036 counties	0.685	0.695	0.670	0.680

Table 6: Root mean-square error (RMSE) comparisons for the LL6u and LR6u models

Comparisons should primarily be made horizontally for Table 6, comparing results for log-level versus log-rate for a given partition of counties. Different subsets of counties will have different mean values, and thus the expected RMSE is different. Even if the difference in means is adjusted, sampling error variances will be much larger for smaller counties in general, and thus larger RMSEs are expected.

The comparisons between log-level LL6u predictions and log-rate LR6u predictions do not show a lower RMSE for the log-rate predictions in any chosen partition. This holds on both the level scale and the rate scale. This suggests the log-rate model *does not fit better* than the log-level model. Note this is a more limited statement than stating the log-level model fits better than the log-rate model. Of note, the individual county-level model predictions produced under either the log-rate or log-level forms, after transformation, are very close to one another in most cases.

The corresponding eight-predictor models, in which the right-hand sides of the rate and level models are equivalent (see Section 3), have the same RMSE relationships. That is, the LL8r model does not have lower computed RMSE than the LR8u model for the same partitions. Thus, again, use of the log-rate model does not appear to raise the goodness of fit relative to use of the log-level model.

6. Conclusion

With the release of 2005 estimates, the SAIPE program utilized ACS direct survey estimates, replacing the CPS ASEC direct survey estimates in the dependent variables of its models. As a result, the SAIPE 2005 poverty estimates generally have reduced variances and smaller published confidence intervals compared with prior years of SAIPE estimates. The SAIPE county models continue to have the log-level form. This work has reviewed some of the estimation and test results that were used to compare the log-level and log-rate forms.

We defined equivalent regression predictors for the log-level and log-rate models and considered the effect of scale or size. For both the log-level and log-rate models, the six-predictor and eight-predictor models perform similarly well. For the log-level model, the null hypothesis of scale invariance is not rejected, and for the log-rate model, in contrast, the null hypothesis of scale invariance is rejected. Comparing the log-level and log-rate models (each with no scale restrictions) directly to one another, the log-rate model is not found to have lower RMSE than the log-level model. This result holds across various population size categories and poverty rate categories. On the basis of goodness of fit, the results do not suggest a need to change from the log-level form of model.

Future work will draw on discussion from the SAIPE external review from September 2007. In particular, research may be done on a binomial logistic model, which could potentially limit further the impact of censored counties with zero poverty estimates. Also, some research may assess the suitability of multilevel and generalized linear mixed modeling to county poverty estimation. Work is currently underway to evaluate county-level ACS sampling variance models and corresponding sampling variance shrinkage estimates (as mentioned in Section 5.2).

References

Bell, William, Wesley Basel, Craig Cruse, Lucinda Dalzell, Jerry Maples, Brett O'Hara, and David Powers. "Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties," prepared in September 2007, <http://www.census. gov/hhes/www/saipe/techrep/report.pdf>

Fay, Robert E., III, and Roger A. Herriot. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association.* Vol. 74, No. 366 (Jun., 1979), pp. 269-277.

Fay, Robert E., III, and George F. Train. "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," *1995 Proceedings from the American Statistical Association*. Available on the World Wide Web, <http://www.census.gov/hhes/www/saipe/asapaper/FayTrain95.pdf>

National Research Council, 1998. *Small-Area Estimates of Children in Poverty, Interim Report 2, Evaluation of 1995 County and School District Estimates for Title I Allocations. Panel on Estimates of Poverty for Small Geographic Areas*. C.F. Citro, M.L. Cohen, and G. Kalton, eds., Committee on National Statistics. Washington, D.C.: The National Academies Press.