Alternate CPS Sampling Variance Structures for Constrained and Unconstrained County Models

Robin Fisher and Jana Asher
SAIPE Technical Report #1
Released December, 1999; Revised July, 2000[1]

Abstract

This document explores alternate CPS county sampling variance functions and the effects of these functions on county-level income and poverty estimates produced by the Small Area Income and Poverty Estimates program (SAIPE) of the U.S. Census Bureau. We will first describe the background of this problem and two alternate modeling strategies for county poor that we explored as part of our research. We will then outline the methodology used to test the effect of different functions of CPS county sample size used in estimating the CPS county sampling variance for the two county poor modeling strategies. Finally, we will examine our results, which suggest a variance proportional to the inverse of the square root of CPS sample size provides the best model of CPS county sampling variance of the options we explored. We finish by describing potential future research topics.

Background

The current SAIPE model for county poor is:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} + \mathbf{e}$$

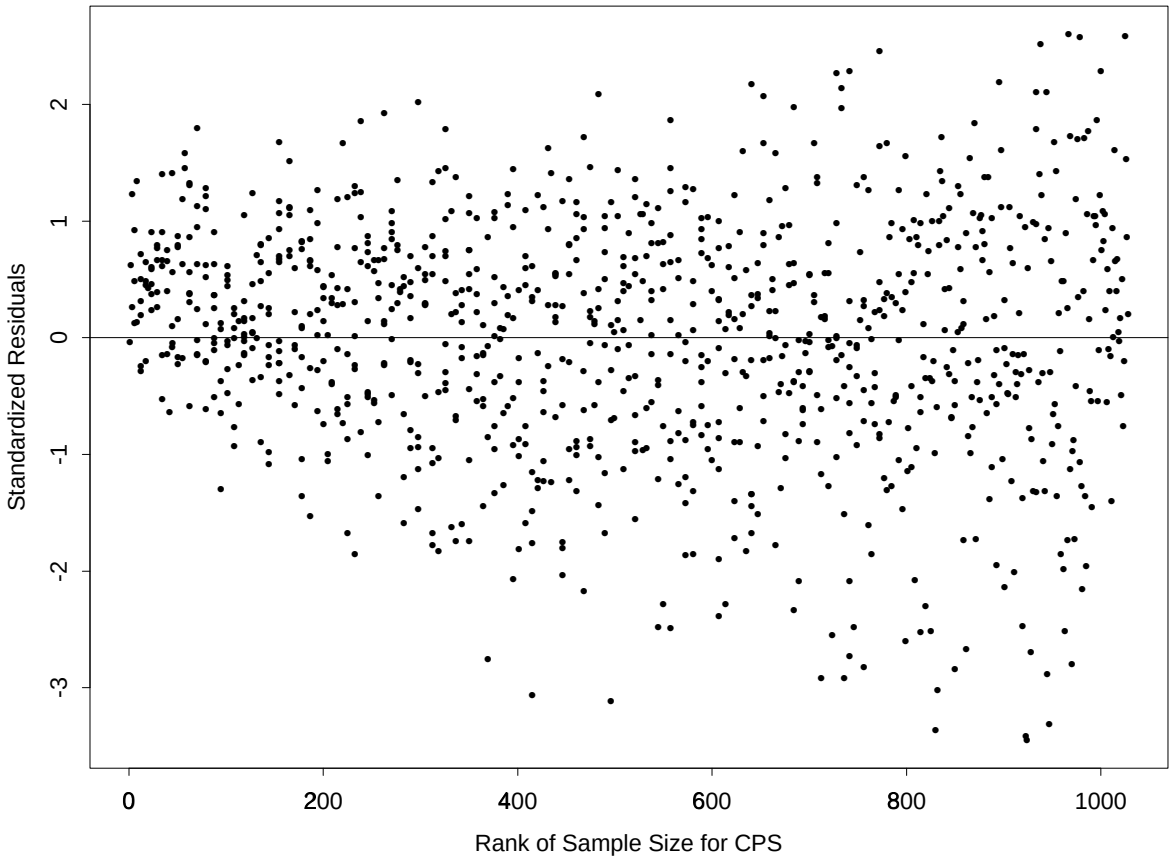
where $\mathbf{y}$ is a vector of log CPS number of poor for counties, $\mathbf{X}$ is a matrix of log values of variables from administrative records, $\mathbf{u}$ represents the model error and is distributed $N[0, \mathbf{V_u}]$, and $\mathbf{e}$ represents the sampling error and is distributed $N[0, \mathbf{V_{ec}}]$. $\mathbf{V_u}$ is assumed to take the form $v_u\mathbf{I}$; $\mathbf{V_{ec}}$ is assumed to be a diagonal matrix whose entries take the form $\frac{\sigma^2}{k_i}$ where k_i, the CPS sample size for county i, has the range (1, 4968). CPS sample size is taken as $\sum k_{ij}$, where k_{ij} is the number of households in CPS sample for county i, year j.

Initial analysis of the standardized residuals of this model suggests a problem; when the residuals are plotted against the rank of sample size, a heteroscedastic pattern is revealed (see Figure 1). The standardized residuals are taken as $\frac{y_i - \hat{y_i}}{s_{y_i}\sqrt{1-h_{ii}}}$, or the data minus the fitted value, divided by an estimate of the standard deviation for the residual.

[1]Changes include a correction of the definition of CPS sample size k_i given in the first paragraph of the Background and Future Research sections, and a clarification of the absolute relative differences given in the chart on page 7.

The standardized residuals represent the sum of the model error and the sampling error divided by joint standard deviation of this sum. Since the model error is assumed to have constant variance, we take the heteroscedasticity of the standardized residuals to suggest a misspecification of the sampling error variance. The funnel-shaped pattern of the standardized residuals suggests that the inverse of sample size underestimates variances for counties with large samples relative to those with small samples in our current modeling procedure.

Figure 1 - Standardized Residuals vs. Rank of CPS Sample Size



The current county level estimation procedure has two steps. The first is only for the estimation of the model error variance. In this step, we estimate the variance of census estimates of poverty with maximum likelihood where the 1990 decennial census direct estimate of poverty is the dependent variable and the sampling error variance is estimated with a generalized variance function. (See U.S. Bureau of the Census (1990) for a description of the generalized variance function.) In the second stage, the model is estimated with CPS poor as the dependent variable and with the model error variance from the census model. The sampling error is assumed to be a constant times the inverse of the

CPS sample size for the county; maximum likelihood estimation is used to determine an estimate of the value of this constant. We call the model created using this two-step procedure the constrained model, as the model error variance is constrained to a value created from census data.

Another option is to assume that the model error is constant, but not estimate it through the census data and sampling error. To do this, maximum likelihood estimation is used to jointly determine v_u and σ^2. We call this model the unconstrained model.

We consider two alternate functions for CPS sampling variance. One function is proposed by Wayne Fuller and Juan Goyeneche in their note "A Different CPS Model Error Specification" (1998). Fuller and Goyeneche use a piece-wise function of the estimated census sampling variance fitted to the residuals of the regression in order to estimate the CPS sampling variance. The other function takes the CPS sampling variance as proportional to the inverse of a power of CPS sample size.

Our task is to find the optimal power of sample size to use as a component of the sampling error, determine whether the constrained or unconstrained model performs better given that variance function, and compare the results obtained with those derived using Fuller and Goyeneche's sampling error function.

Methodology

For this series of experiments, we use the 1990 model for related children age 5-17 in families in poverty. Our methodology can be broken down into two steps:

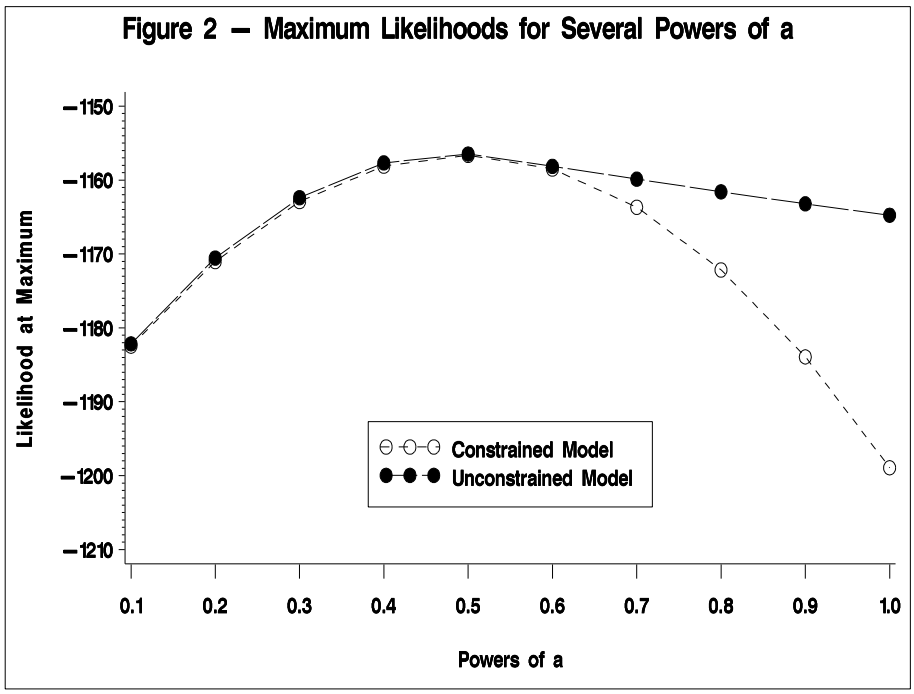
1) Determine an optimal power a for the function $\frac{\sigma^2}{k_i^a}$ of the CPS sample size used in the model of the sampling variance. To do this for the constrained model, we maximize the following log likelihood function jointly for β and σ^2, for each $a \in (.1, .2, .3, .4, .5, .6, .7, .8, .9, 1)$:

$$L(\beta, \sigma^2|a) \quad = \quad \sum_{i=1}^{n} -\frac{1}{2} log(v_u + \frac{\sigma^2}{k_i^a}) - \frac{1}{2} \frac{(y_i - x_i\beta)^2}{v_u + \frac{\sigma^2}{k_i^a}}$$

where:
y_i = CPS poor for county i
x_i = predictors for county i
k_i = CPS sample size for county i
v_u = model variance determined from census data.

We repeat this process for the unconstrained model; the only difference is that we maximize jointly over β, σ^2, and v_u for each value of a.

Figure 2 shows the results of these steps; the value of the likelihood maximized for a particular value a is plotted against a for both the constrained and unconstrained model. In both cases, by visual inspection we determine that the maximized likelihoods take their greatest value at approximately $a = .5$.



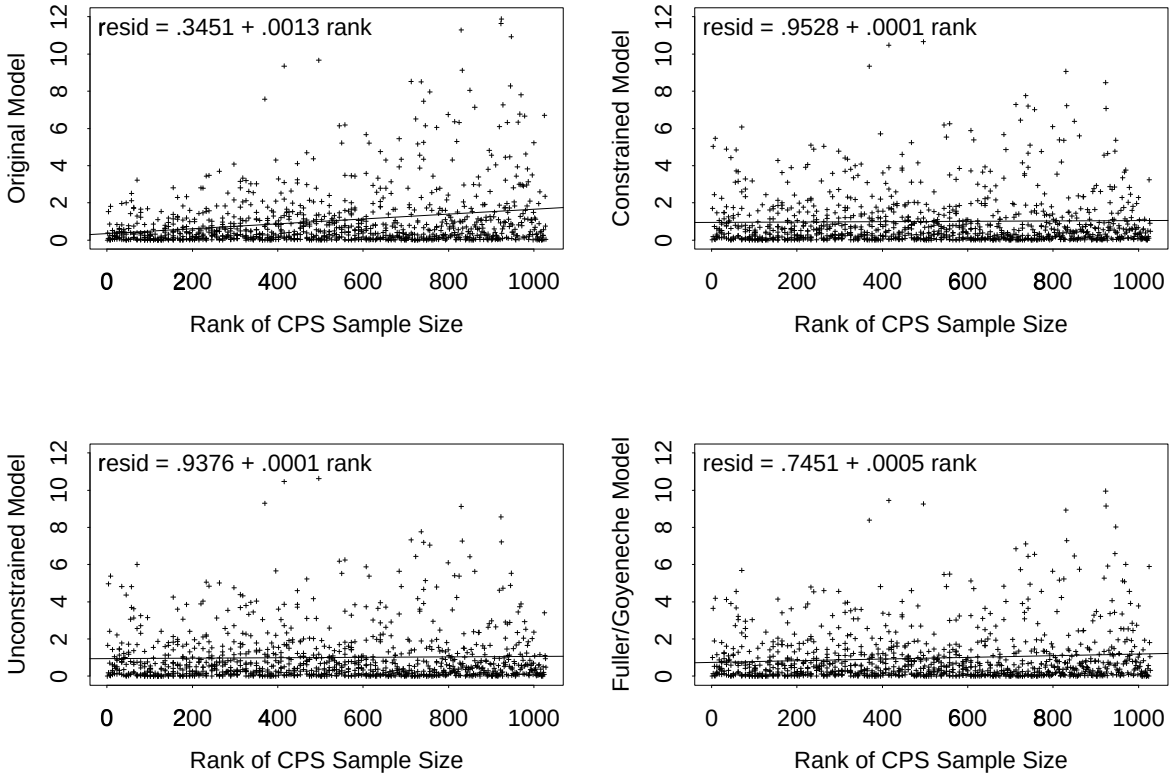
2) Assess the validity of the model using the sampling variance function $\frac{\sigma^2}{\sqrt{k_i}}$ for both the constrained and unconstrained model of log county poor. We also compare these results to the product of the Fuller/Goyeneche sampling variance function and the original sampling variance function. We examine both the plot of standardized residuals versus CPS sample size and a set of regression diagnostics.

Results

To test the heteroscedasticity of the residuals of the four sets of estimates (original, constrained, unconstrained, and Fuller/Goyeneche), we perform a regression analysis of the squared standardized residuals on the rank of CPS sample size for each model. We show these regressions in Figure 3. Because the underlying model assumes normally distributed error terms, the residuals are taken to be normally distributed with mean zero and variance dependent on CPS sample size. The standardized residuals therefore are assumed to be normally distributed with mean zero and variance one, leading to the assumption that the squared standardized residuals follow a chi-squared distribution with one degree of freedom. If these assumptions all hold, the regression lines shown in Figure 3 should have zero slope (indicating the squared standardized residuals have constant variance)

and an intercept of about one (indicating the expected value for the squared standarized residuals to be the expected value of the chi-squared distribution). When the regression line has positive slope, the indication is that the variance is relatively small for estimates corresponding to counties with large samples. This is the situation we see for the current model, shown in the upper left hand corner. The slope is closer to zero in the other three models, especially the unconstrained model and the constrained model.

Figure 3 - Squared Standardized Residuals vs Rank CPS Sample Size



A formal test to determine if the slopes in Figure 3 are significantly different than zero can be performed as follows: Spearman's ρ is determined for the squared standardized residuals and the CPS sample size rank. The hypothesis that the slope equals zero is rejected if the p-value assessed for the test is less than .05. Using this test, the slopes for the original model and Fuller/Goyeneche model regressions are found to be significantly different than zero, while the slopes for the unconstrained and constrained model regressions are not.

A formal test comparing the results of the constrained and unconstrained model regression slopes is performed by applying the same test as follows: if SSR_u is the vector of squared standardized residuals for the unconstrained model, and SSR_c is the vector of squared standardized residuals for the constrained model, then the regression equations can be summed to find:

$$(SSR_u - SSR_c) \quad = \quad (\alpha_u - \alpha_c) + (\beta_u - \beta_c)rank$$

The question of interest is then whether $(\beta_u - \beta_c)$ is zero, or $\beta_u = \beta_c$. Using Spearman's ρ, a p-value of 0 is obtained; therefore we fail to reject the hypothesis that $\beta_u = \beta_c$. Our conclusion is that the unconstrained and constrained modeling procedures produce homogenous standardized residuals, while the original modeling procedure does not. We assume no conclusions for the Fuller/Goyeneche model for reasons outlined below.

These results are particularly interesting for the quality of fit for the unconstrained model. There have been questions about the suitability of the model error variance estimated from the census in the constrained model, especially at the end of a decade. When the CPS sampling variance is assumed to be inversely proportional to the square root of sample size, the constrained model and the unconstrained model give similar results which compare favorably to the current model. The current, constrained procedure is used because it is more robust to problems in the variance model than the unconstrained procedure is, whatever form the variance function takes. As a result, the unconstrained model is not an option given the incorrect specification of the sampling variance in the current SAIPE model. Since the sampling variance function that is inversely proportional to the square root of CPS sample size is more consistent with the data, the robustness property is less necessary. We may therefore be freed from using the model error variance from the census, and can consider using the unconstrained model.

We are additionally interested in how using one of the new variance functions affects the predictions eventually produced by the model. We therefore examine the absolute relative difference between the predictions derived from the original model and those derived from the three alternative models. The absolute relative differences are taken after shrinkage; this is particularly important to note for the Fuller/Goyeneche results.

To implement the Fuller/Goyeneche sampling variance function, we first calculate the variance parameters as in the official model through maximum likelihood estimation. We then use the Fuller/Goyeneche variance function to recaculate the sampling errors. Finally, we do a weighted least squares regression using the inverse of the total error as a weight, which alters the values of the β's. We can then do further analysis of the new estimates and residuals. We should note, however, that if we were to actually use the Fuller/Goyeneche function in a production model, we might wish to apply it in some other way. For example, we might wish to incorporate it directly into the maximum likelihood estimation of β, and σ^2, or we might wish to include the function's changepoints

as parameters in the maximization. For this reason we drew no conclusions from the results of the Spearman's ρ test for the homogeneity of the standardized residuals from the Fuller/Goyeneche results above. The procedure we use here is informative, even so.

The results of our examination follow. These numbers represent the ratio of the sum of the absolute differences between the shrinkage estimate of the alternative model and the shrinkage estimate of the original model to the sum of the shrinkage estimates of the original model, or:

$$RD = \frac{\sum abs(ebest_p - ebest_o)}{\sum(ebest_o)}$$

where *ebest* is the shrinkage estimate before an exponential transformation. We note similar results for the constrained and unconstrained model, and that these variance models lead to about a 1% overall change in the shrinkage estimates before transformation. The Fuller/Goyeneche model leads to about a .7% change. The point is that none of these alternate variance functions cause much change in the actual value of the estimates.

Variance Structure	Absolute Relative Difference
Original	NA
Unconstrained	0.01036
Constrained	0.01063
Fuller/Goyeneche	0.00696

Finally, we are interested in the results of diagnostic testing on demographic factors such as census division, 1990 resident population, percent of population Black or Hispanic, percent in group quarters, and percent poor as given by the 1990 census. By examining the estimates in relation to these groups, we can look for biases for or against these groups in our models. We do so through plots of the standardized, relative, and absolute relative differences of the estimates for a particular variance function from the 1990 census values, divided by categories into boxplots. Results of these tests suggest that all three alternate modeling options (constrained, unconstrained, and Fuller/Goyeneche) do not significantly affect the fit or validity of the county model; the plots for the different variance functions are quite similar for most demographic factors. Figures containing these plots for the unconstrained model and the original model follow in the Appendix (see Figures 4-41, pages 11-29).

We additionally note that the proposed variance functions lead to estimates that appear to produce better results when comparing the standardized differences from the Census against 1990 resident population (see Figures 10-12, pages 14-15, and note that the boxplots seem more uniformly centered on the line at zero for the unconstrained model than

the original model). This is not surprising given the correlation between CPS sample size and county population size. The heteroscedasticity of the standardized residuals plotted against either CPS sample size or county population size is reduced for the proposed variance models in comparison to the original model (see Figures 40-41, page 29, and note the reduction in the "funnel" shape for the proposed model).

Recommendations

The maximized log likelihood functions for the constrained and unconstrained models appear equal at the optimal value of a, the power of the sample size. By visual inspection, that optimal value is approximately .5.

The sampling variance taken as a function of the inverse of $\sqrt{k_i}$ provides a better fit than the sampling variance taken as a function of the inverse of k_i, where k_i is CPS county sample size.

We believe that Fuller and Goyeneche's sampling variance model would be difficult to implement because of the complexity of recalculating two new parameters for each model year and age group: the change points in the function.

As a result of this research, we recommend further study of the unconstrained model taking sampling error to be proportional to the inverse of the square root of CPS county sample size. We believe that the fit of the variance model may change over time as the variation between counties changes over time, therefore use of the unconstrained model for county poor has the benefit of removing our dependency on the fixed-in-time census data. We believe switching to this modeling for future estimate years could be beneficial.

Future Research

Changing our current definition of CPS sample size. The current definition of CPS sample size for county i is $k_i = \sum k_{ij}$, where k_{ij} is the number of households in CPS sample for county i, year j. This definition of CPS sample size does not take into account the sample overlap between years. We could redefine this to model the correlation structure between years or within a sample.

Testing the proposed variance models for different statistics and different years. Our research has focused on number poor of related children age 5-17 in families. To verify our results, we need to repeat our tests for other statistics and the 1993 and 1995 income years, and check the heteroscedasticity of the standardized residuals and the effect of alternate variance models on the estimates.

Testing alternate sampling variance models where the variance depends on the value es-

timated. To date, we haven't explored models where the variance depends on the value estimated. For example, binomial models of the variance may fit well. Several techniques for this type of variance estimation are discussed in Carroll and Ruppert (1988); we would like to explore some of these.

Using a generalized variance function to estimate sample variance. We would base this work on research in progress by Bell and Kramer (1998).

References

Bell, William and Matt Kramer. *Generalized Variance Functions of Sampling Error Variances of Direct CPS County Poverty Estimates.* Unpublished note to the members of the National Academy of Sciences Panel on Estimates of Poverty for Small Geographic Areas, 1998.

Carroll, R.J. and D. Ruppert. Transformation and Weighting in Regression. New York, Chapman and Hall, 1988.

Fisher, Robin. *Methods Used for Small Area Poverty and Income Estimation.* American Statistical Association, Proceedings of the Section on Government Statistics and Section on Social Statistics, 1997, pp 177-182.

Response to the NAS Request For the SAIPE Model Evaluations: Part 2. Part of mailing on August 27th, 1997, to the members of the National Academy of Sciences Panel on Estimates of Poverty for Small Geographic Areas, in preparation for the fifth meeting of the Panel held September 19-20, 1997.

U.S. Bureau of the Census (1990). 1990 Census of Population, Series CP(2). *Social and Economic Characteristics, Missouri, Section 1 of 2.* U.S. Government Printing Office, September 1993, pp C2-C5, C10-C11.

Disclaimer

Robin Fisher and Jana Asher are mathematical statisticians in the Housing and Household Economic Statistics Division (HHES) of the U.S. Census Bureau.

This technical report is based on research and analysis undertaken by U.S. Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the Census Bureau. It is released to inform interested parties of current research and to encourage discussion. The authors wish to acknowlege the support and assistance of William R. Bell and Paul Siegel.

Appendix: Supporting plots.

The following are diagnostic plots for both the official county model and the proposed unconstrained model with sampling variance inversely proportional to the square root of CPS sample size. Diagnostic plots for estimates produced from the constrained model and the Fuller/Goyeneche model are available upon request.

For Figures 4 through 39, each figure contains two plots: the plot on the left contains the results of the proposed unconstrained model, and the plot on the right contains the results of the official county model. Every set of three figures contains first the standardized, then relative, and then absolute relative differences of the estimates from the 1990 census. Figures 40 and 41 are standardized residual plots. The order of the plots is as follows:

Figures 4-6:	Comparison by census division.
Figures 7-9:	Comparison by metropolitan area status as given by the 1990 census.
Figures 10-12:	Comparison by 1990 resident population.
Figures 13-15:	Comparison by the 1980 to 1990 population growth.
Figures 16-18:	Comparison by the percent poor as given by the 1990 census.
Figures 19-21:	Comparison by the percent of population Hispanic in 1990.
Figures 22-24:	Comparison by the percent of population Black in 1990.
Figures 25-27:	Comparison by ERS persistent rural poverty from 1960-1990.
Figures 28-30:	Comparison by ERS economic type for rural counties.
Figures 31-33:	Comparison by percent of population in group quarters as given by the 1990 census.
Figures 34-36:	Comparison for Virginia independent cities.
Figures 37-39:	Comparison by status in Current Population Survey, 1989-1991.
Figures 40-41:	Standardized residual plots for original and proposed models compared by CPS sample size.

These plots were originally developed to test against bias in the estimation procedure for or against particular populations.

Figure 4 - Standardized Difference from the Census, Compared by Census Division

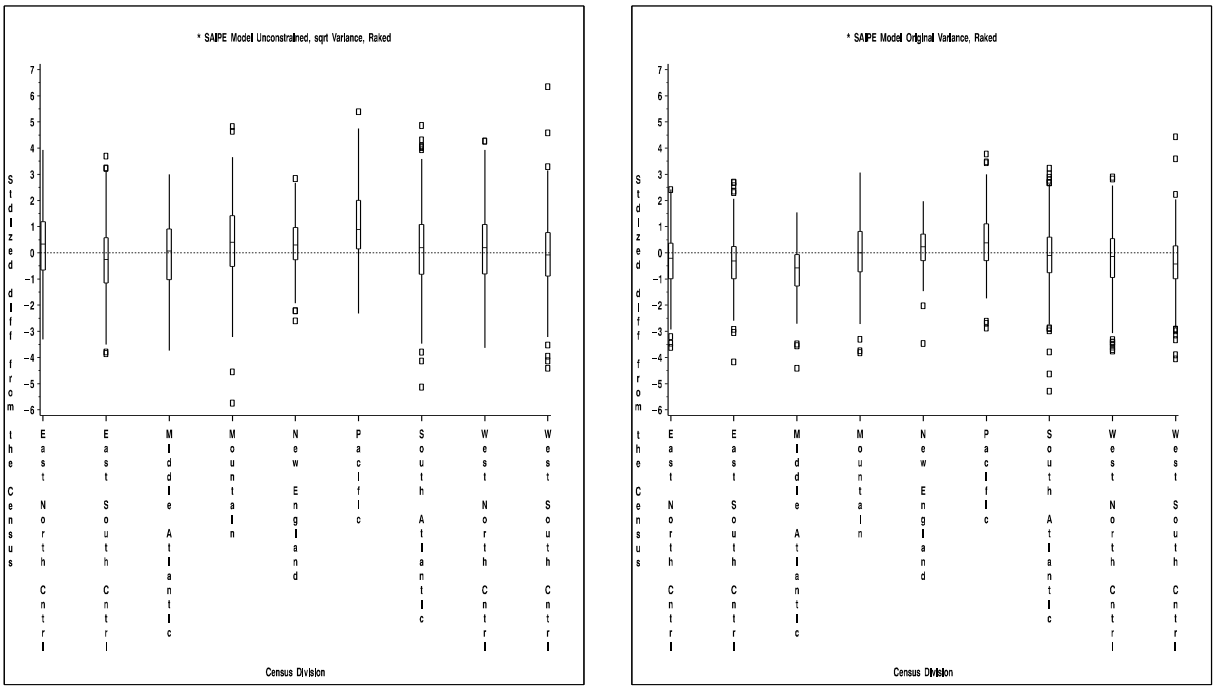


Figure 5 - Relative Difference from the Census, Compared by Census Division

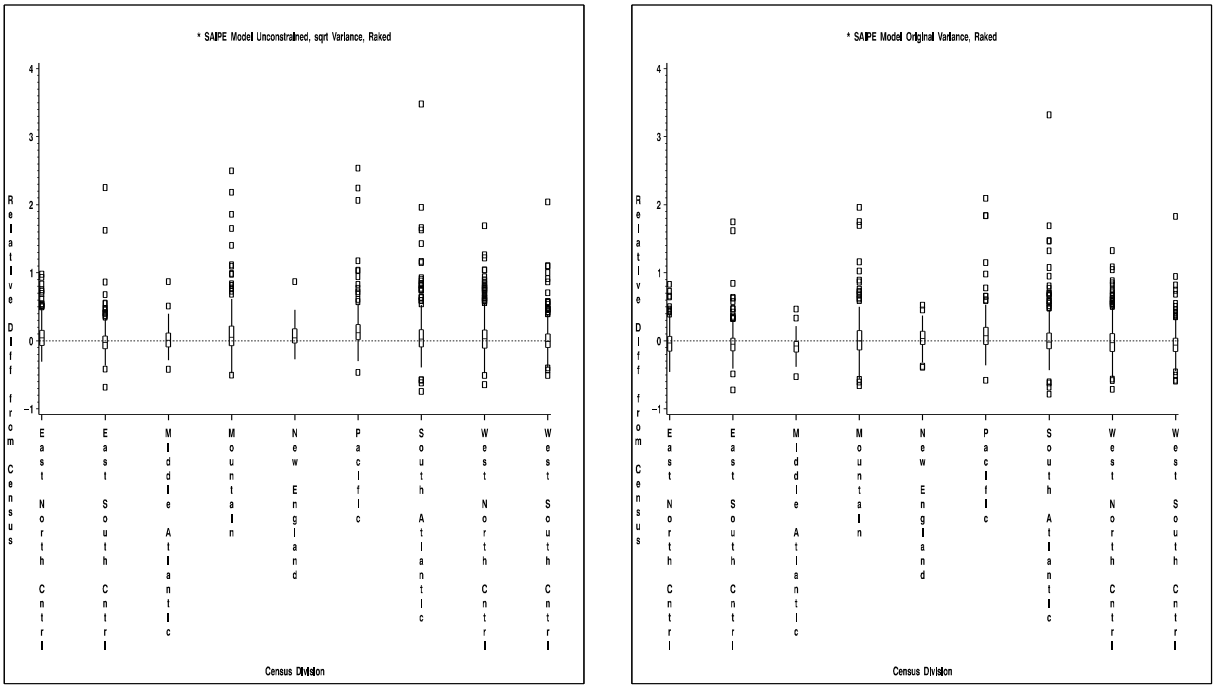


Figure 6 - Absolute Relative Difference from the Census, Compared by Census Division

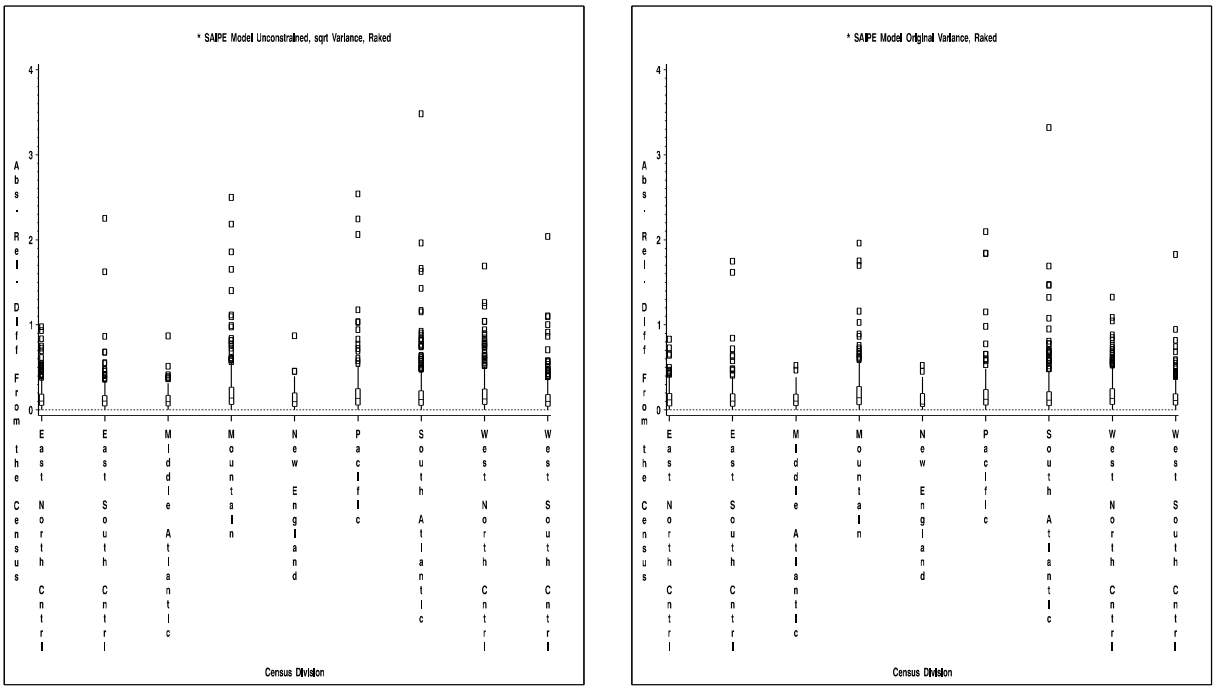


Figure 7 - Standardized Difference from the Census, Compared by Metropolitan Area Status (1990 Census)

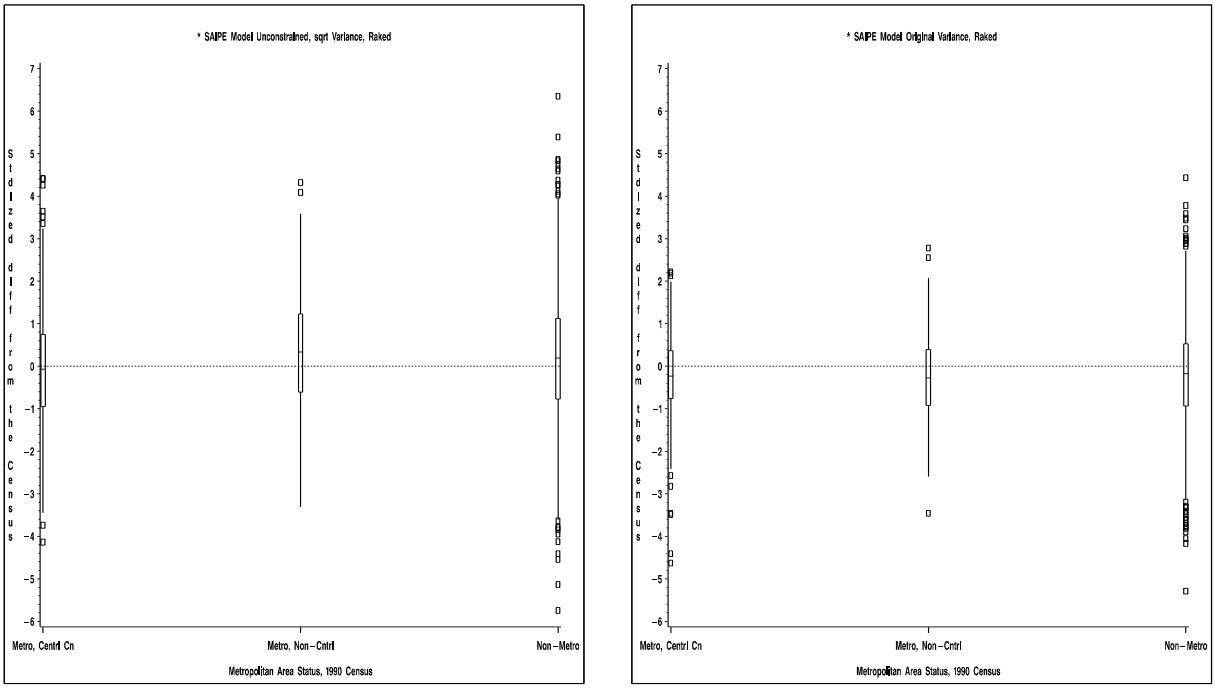


Figure 8 - Relative Difference from the Census, Compared by Metropolitan Area Status (1990 Census)

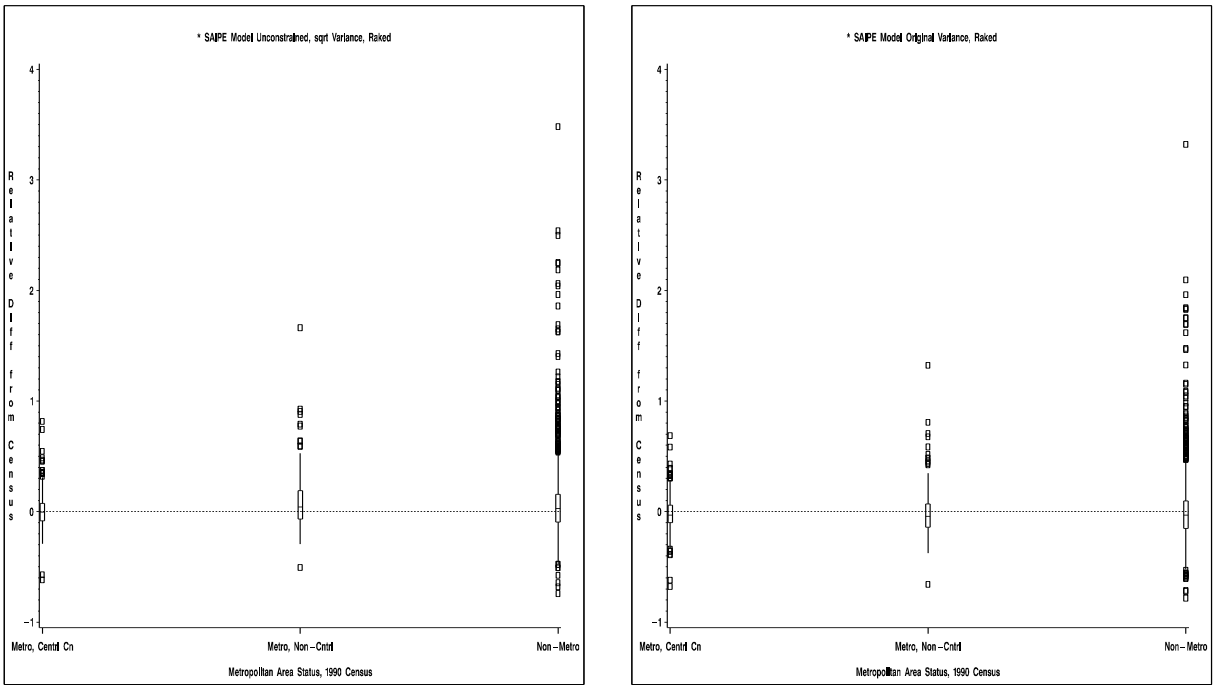


Figure 9 - Absolute Relative Difference from the Census, Compared by Metropolitan Area Status (1990 Census)

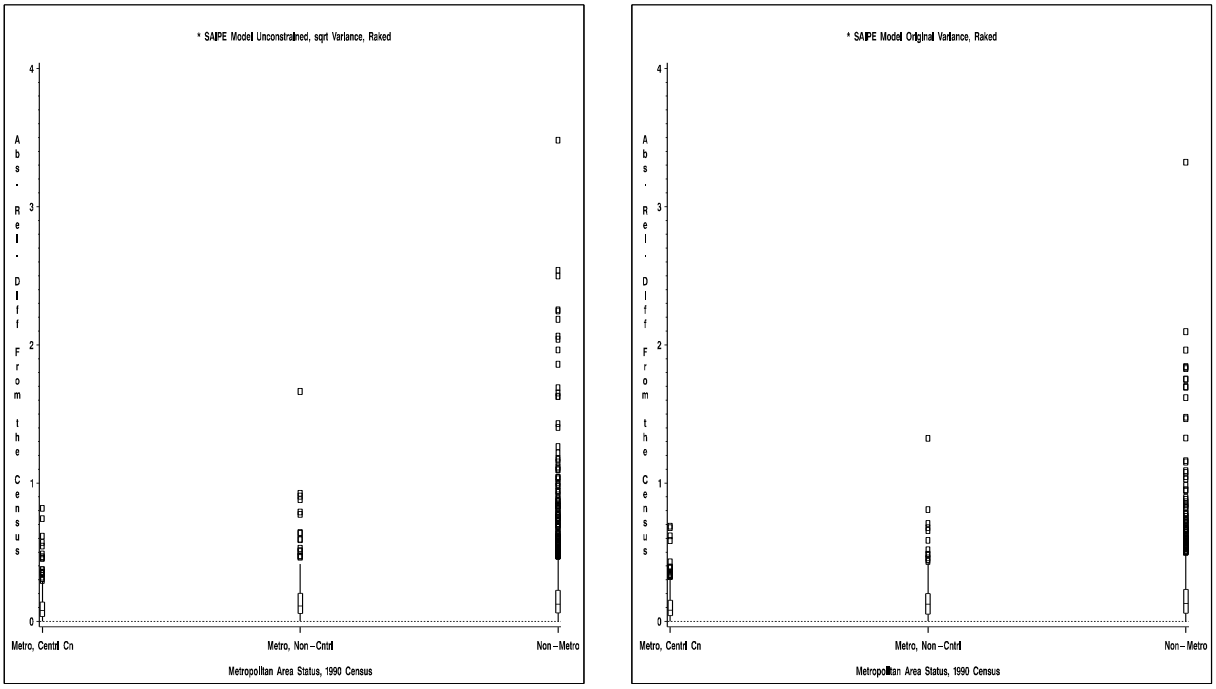


Figure 10 - Standardized Difference from the Census, Compared by 1990 Resident Population

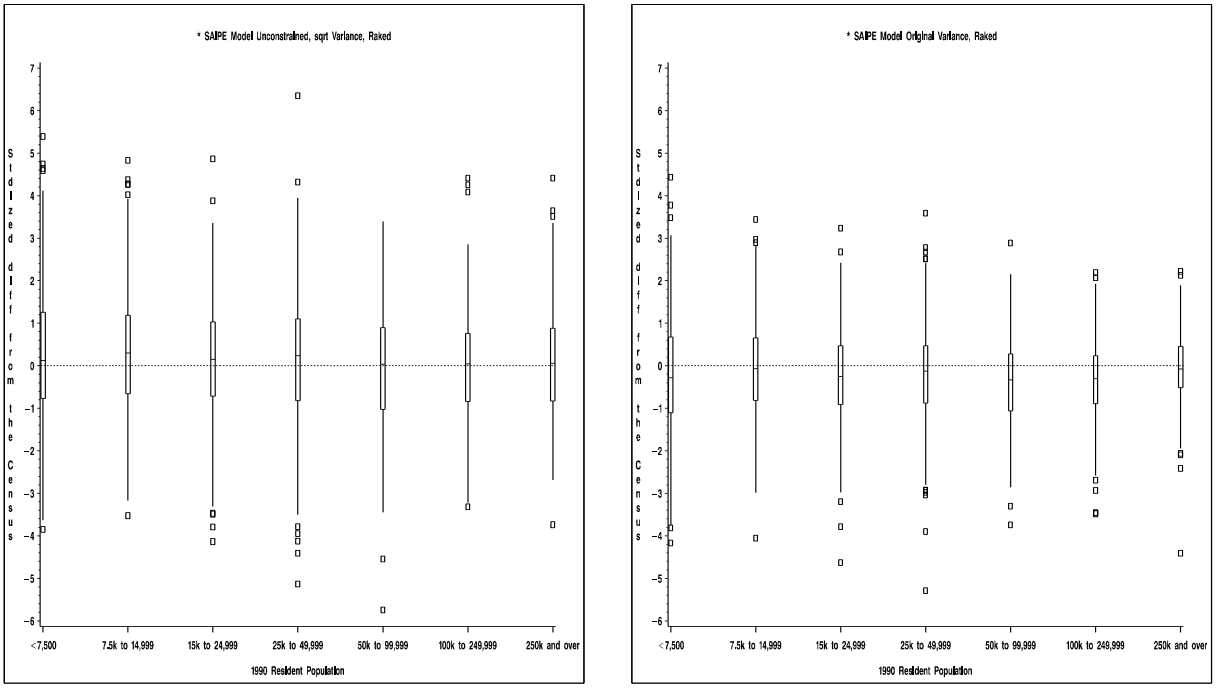


Figure 11 - Relative Difference from the Census, Compared by 1990 Resident Population

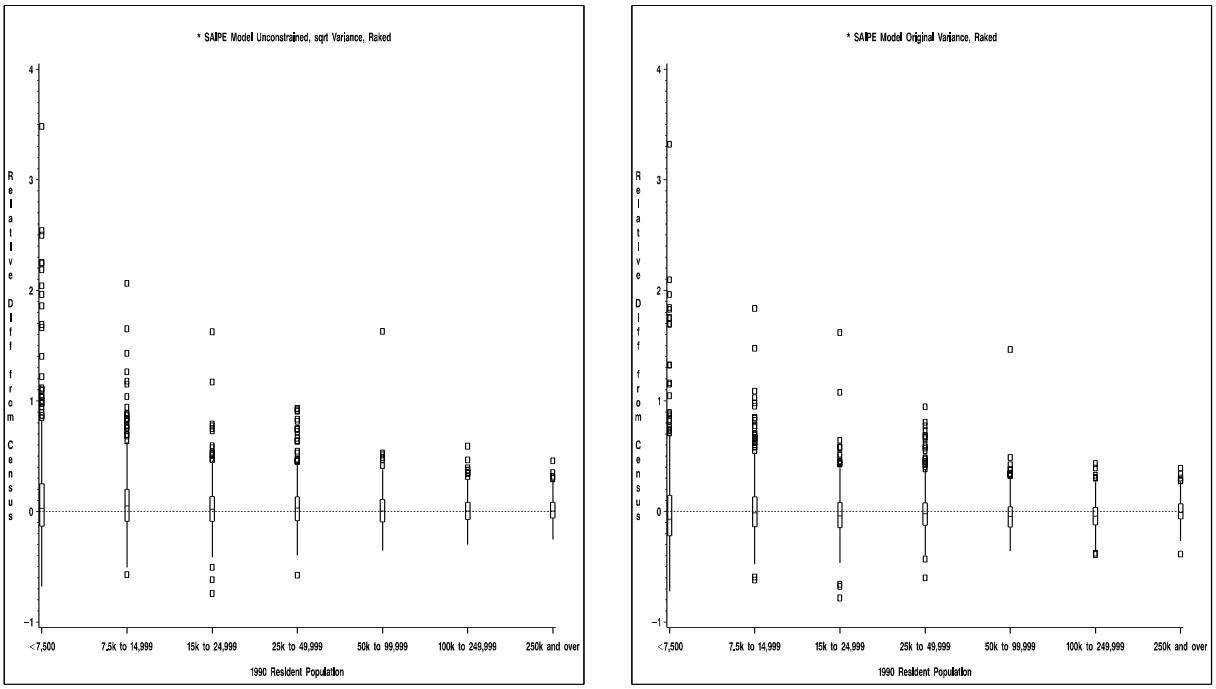


Figure 12 - Absolute Relative Difference from the Census, Compared by 1990 Resident Population

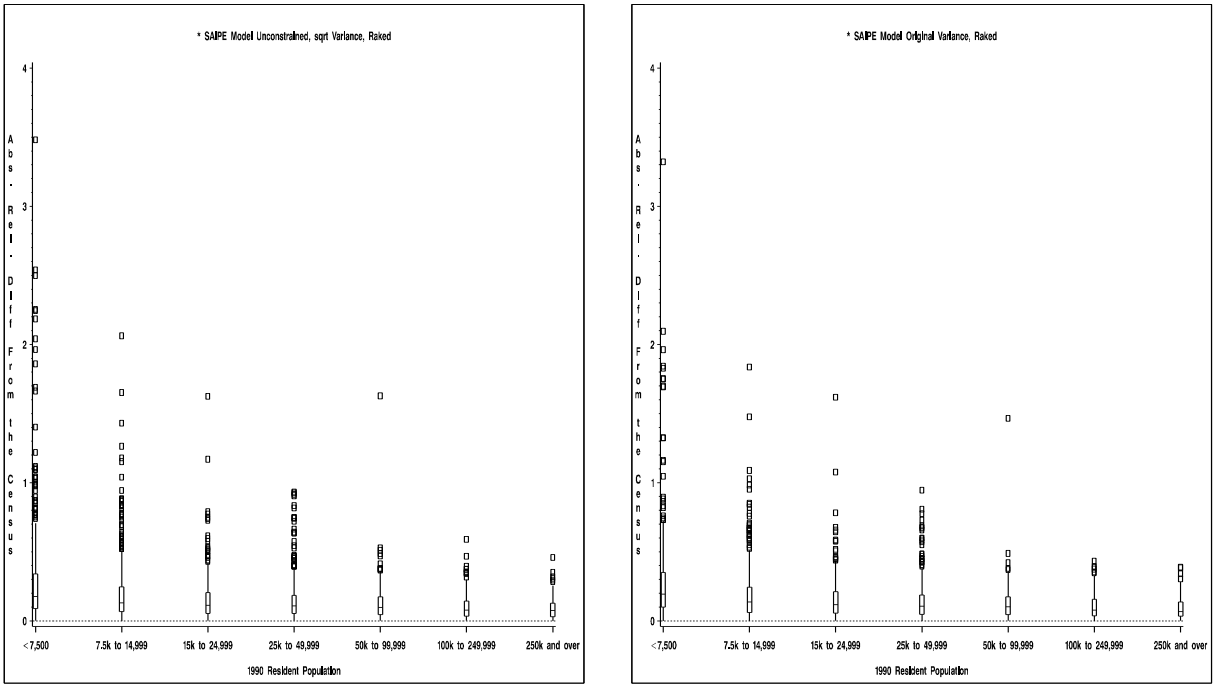


Figure 13 - Standardized Difference from the Census, Compared by 1980 to 1990 Population Growth

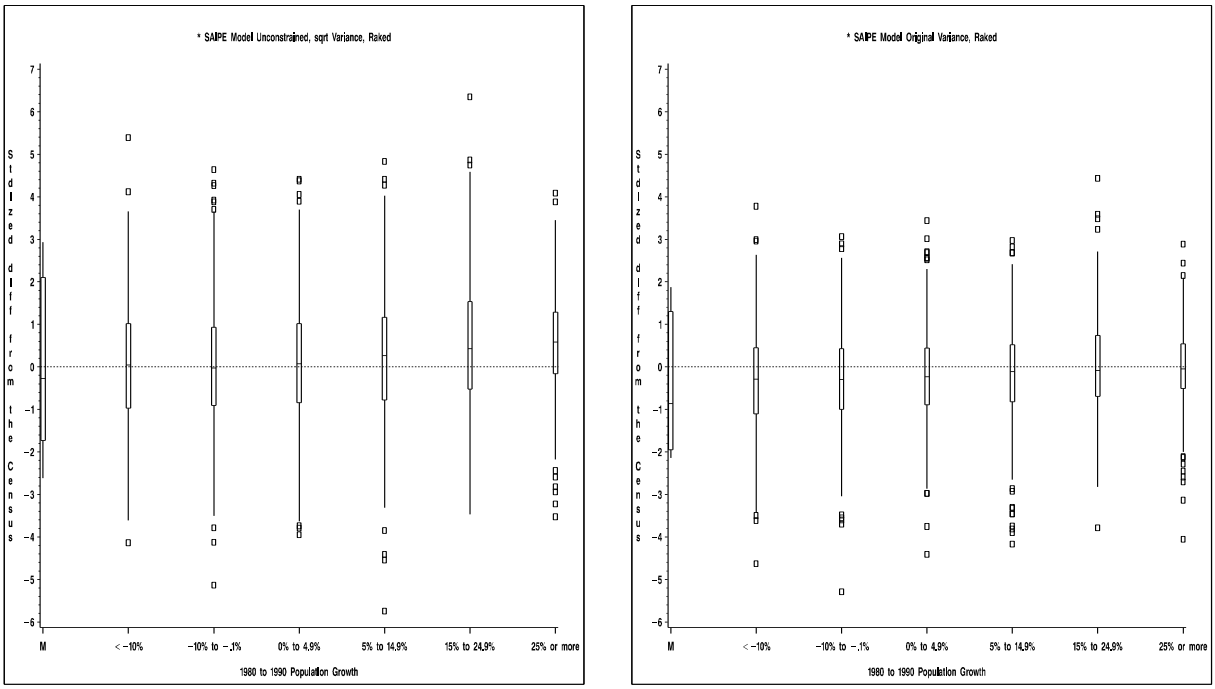


Figure 14 - Relative Difference from the Census, Compared by 1980 to 1990 Population Growth

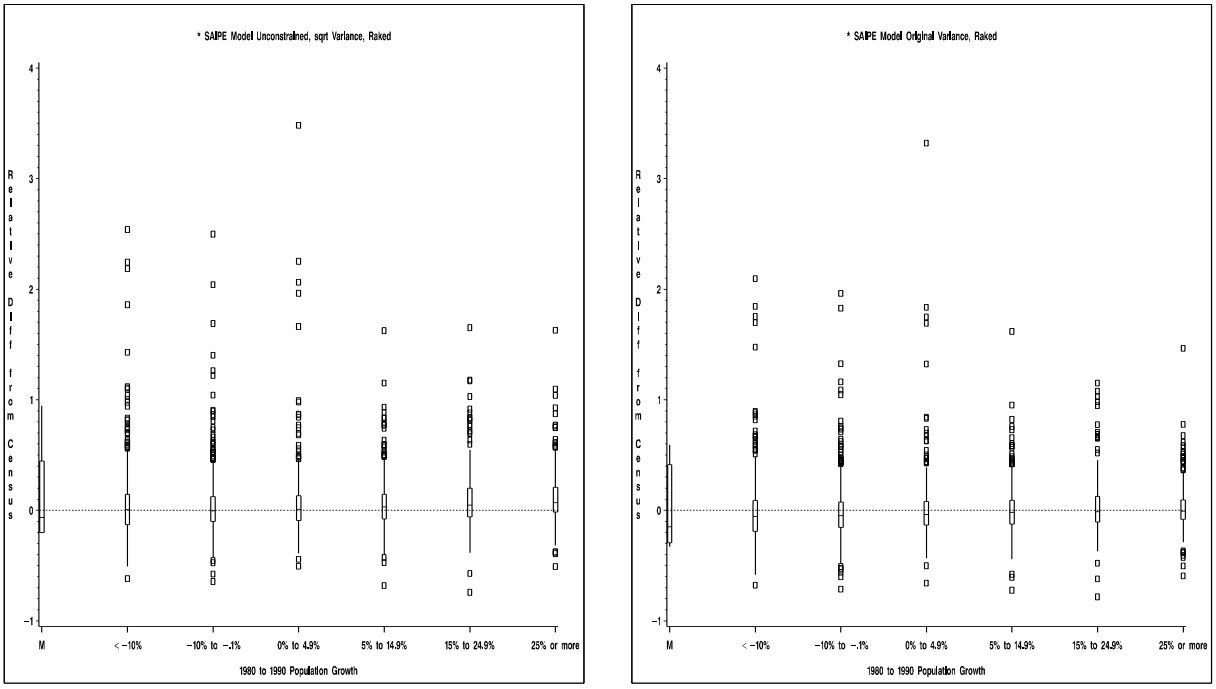


Figure 15 - Absolute Relative Difference from the Census, Compared by 1980 to 1990 Population Growth

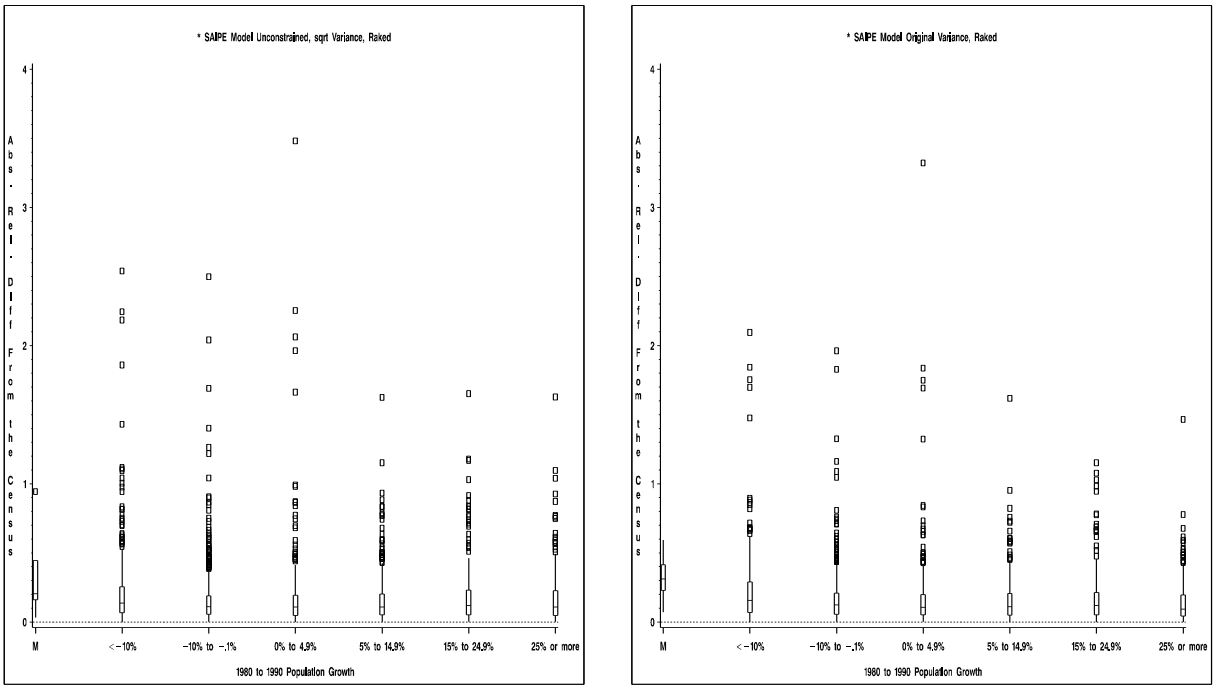


Figure 16 - Standardized Difference from the Census, Compared by Percent Poor (1990 Census)

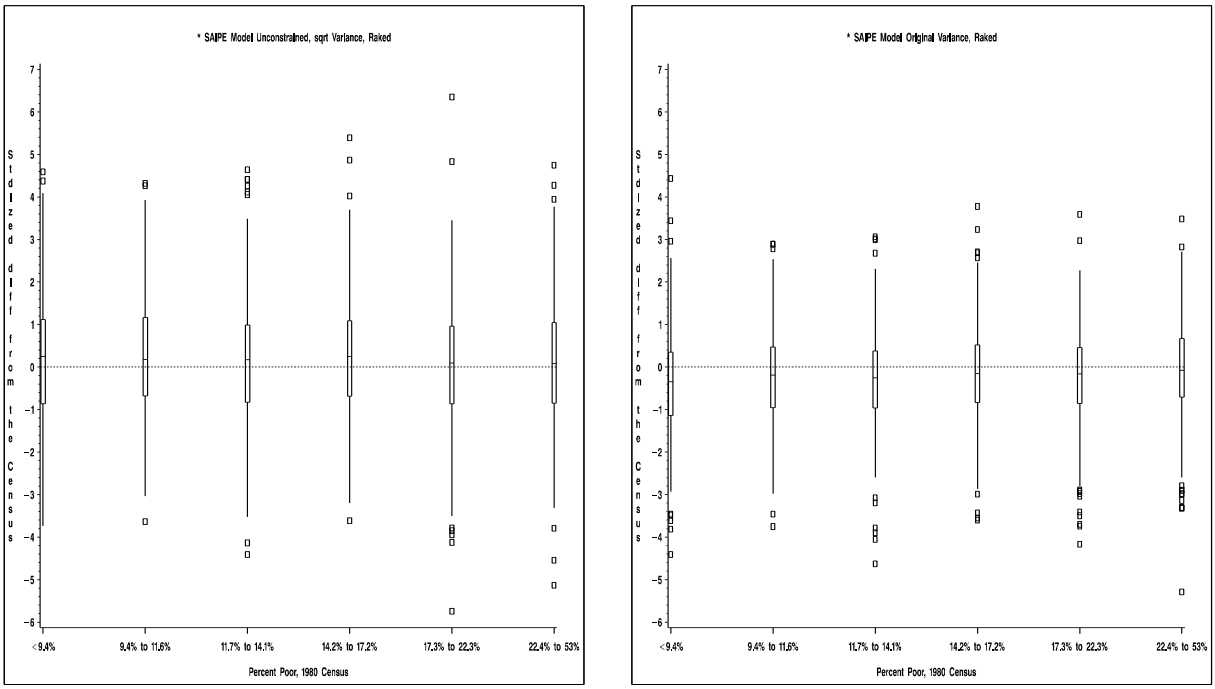


Figure 17 - Relative Difference from the Census, Compared by Percent Poor (1990 Census)

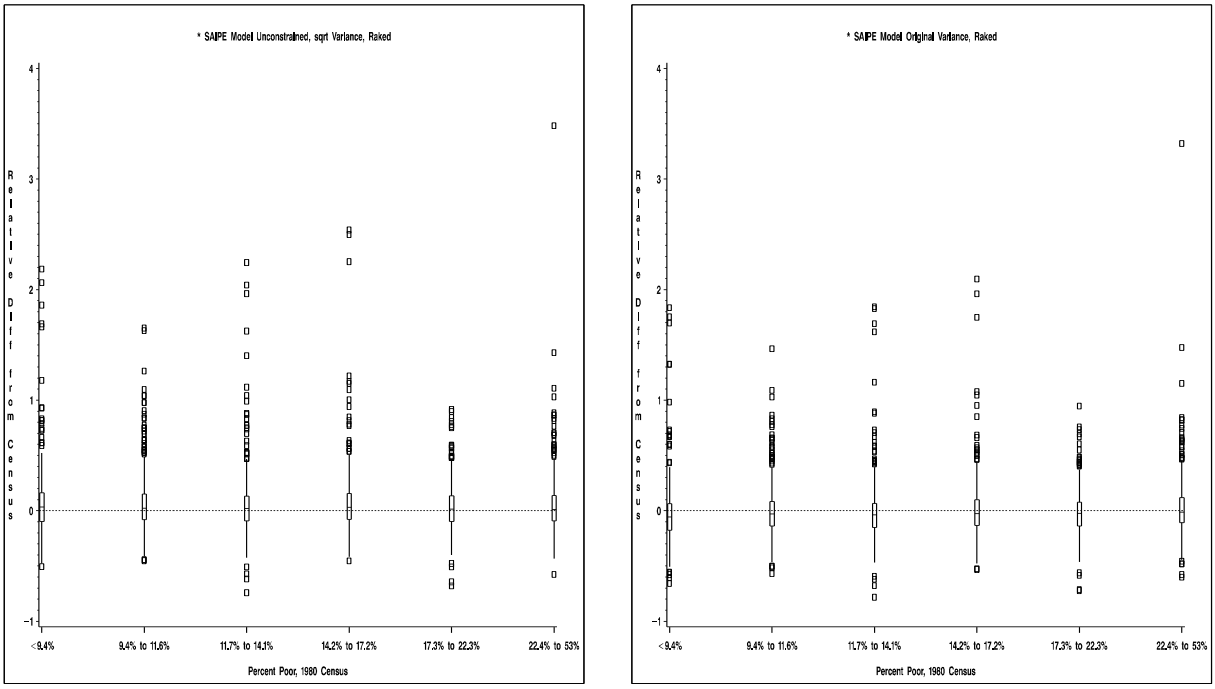


Figure 18 - Absolute Relative Difference from the Census, Compared by Percent Poor (1990 Census)

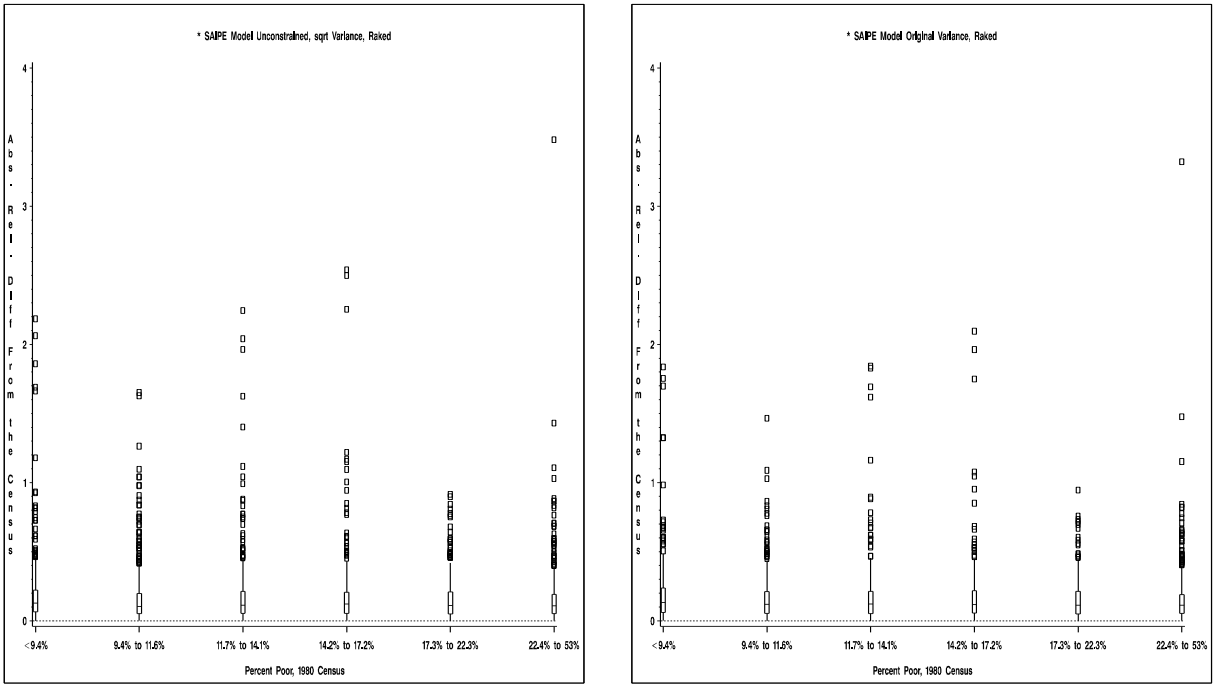


Figure 19 - Standardized Difference from the Census, Compared by Percent of Population Hispanic in 1990

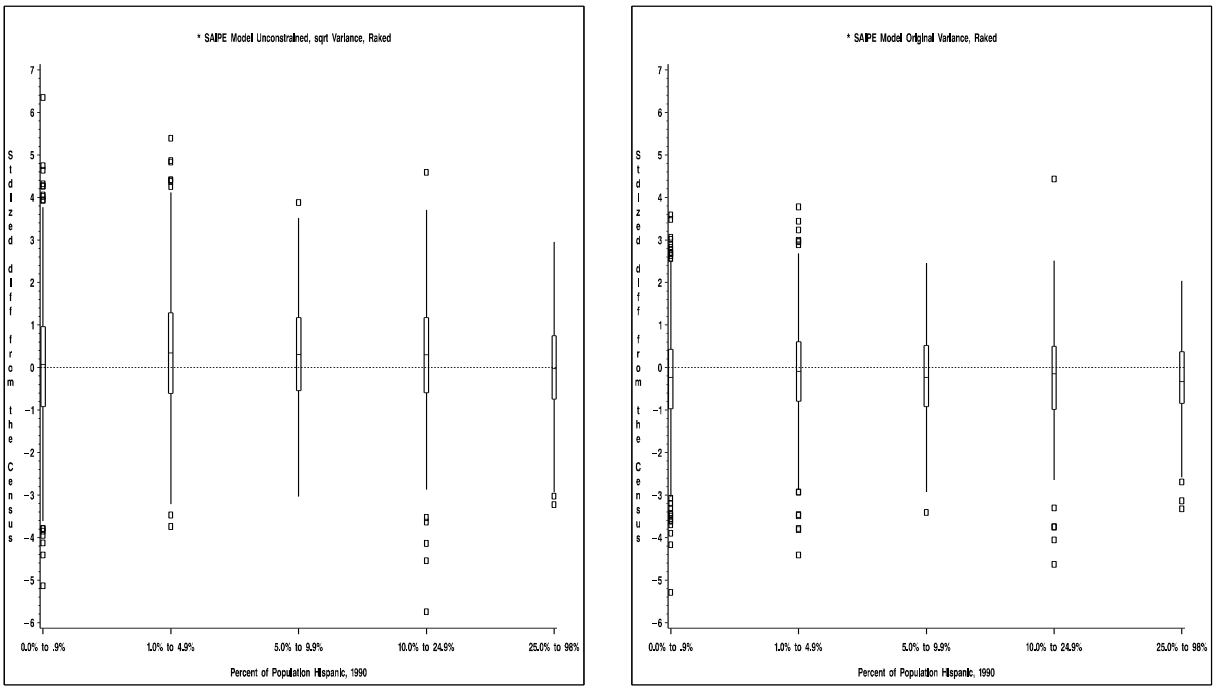


Figure 20 - Relative Difference from the Census, Compared by Percent of Population Hispanic in 1990

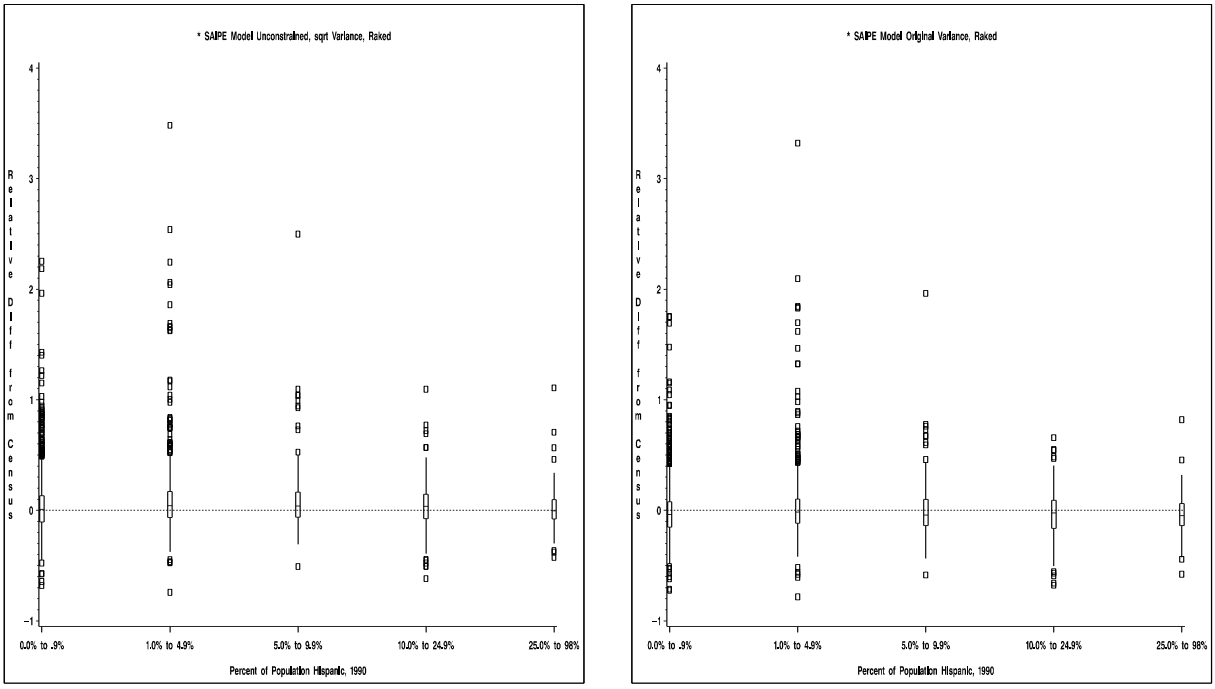


Figure 21 - Absolute Relative Difference from the Census, Compared by Percent of Population Hispanic in 1990

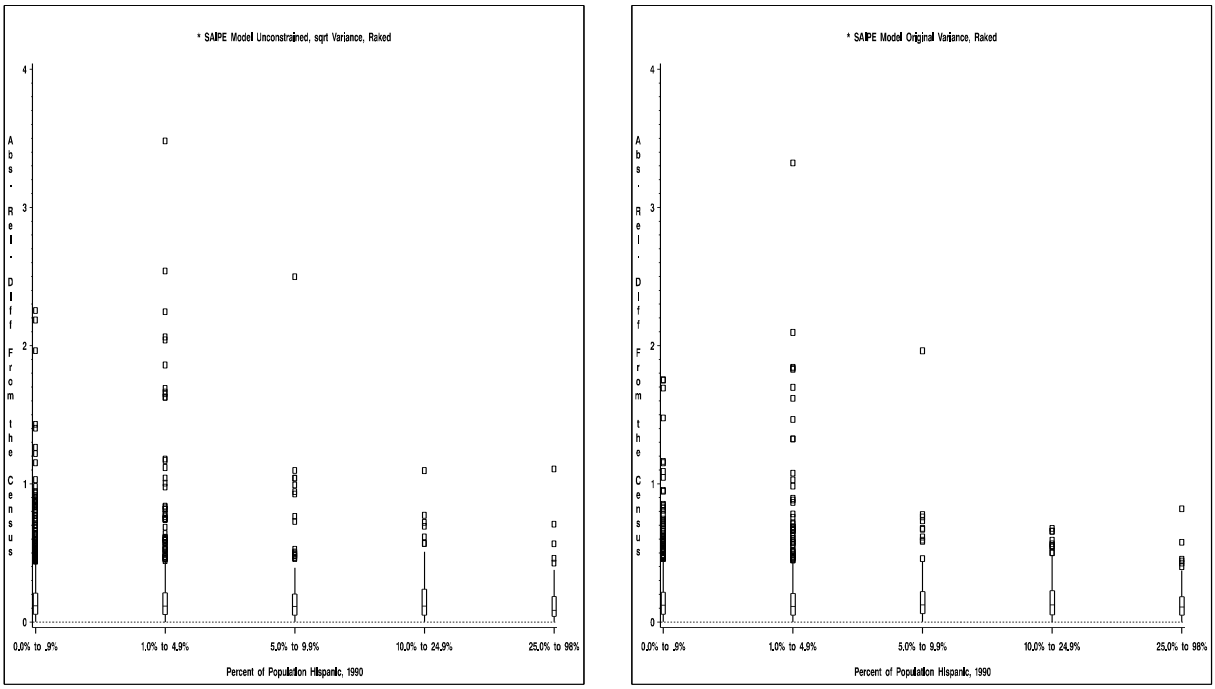


Figure 22 - Standardized Difference from the Census, Compared by Percent of Population Black in 1990

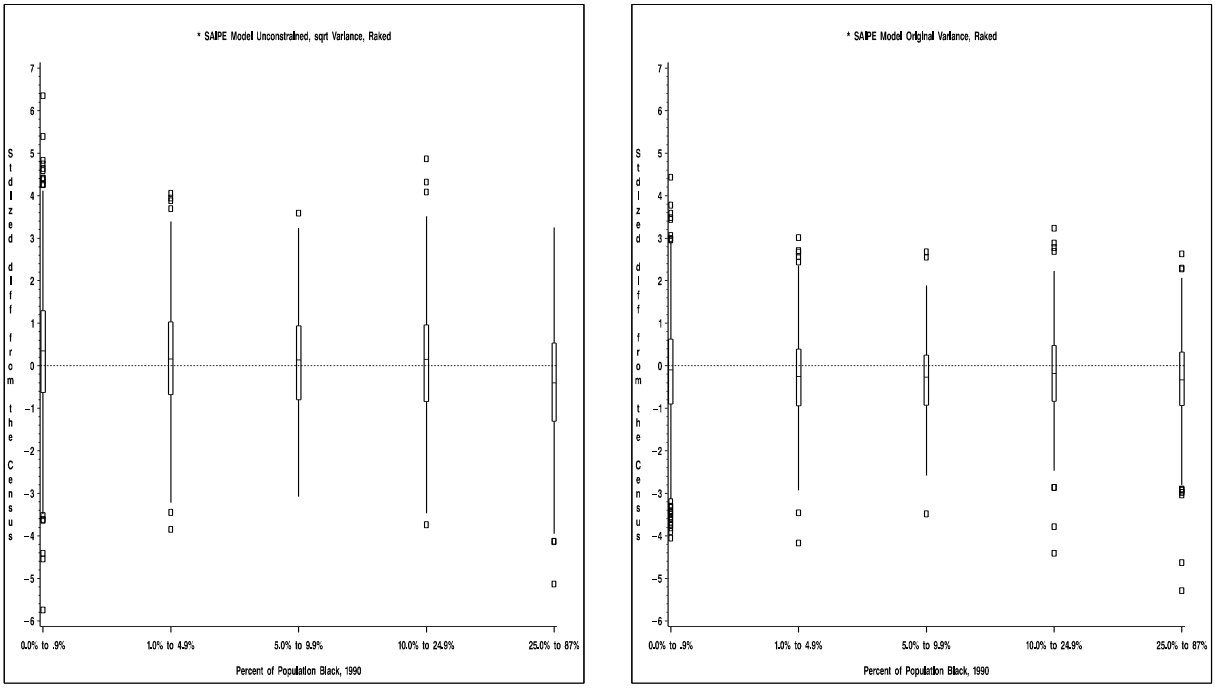


Figure 23 - Relative Difference from the Census, Compared by Percent of Population Black in 1990

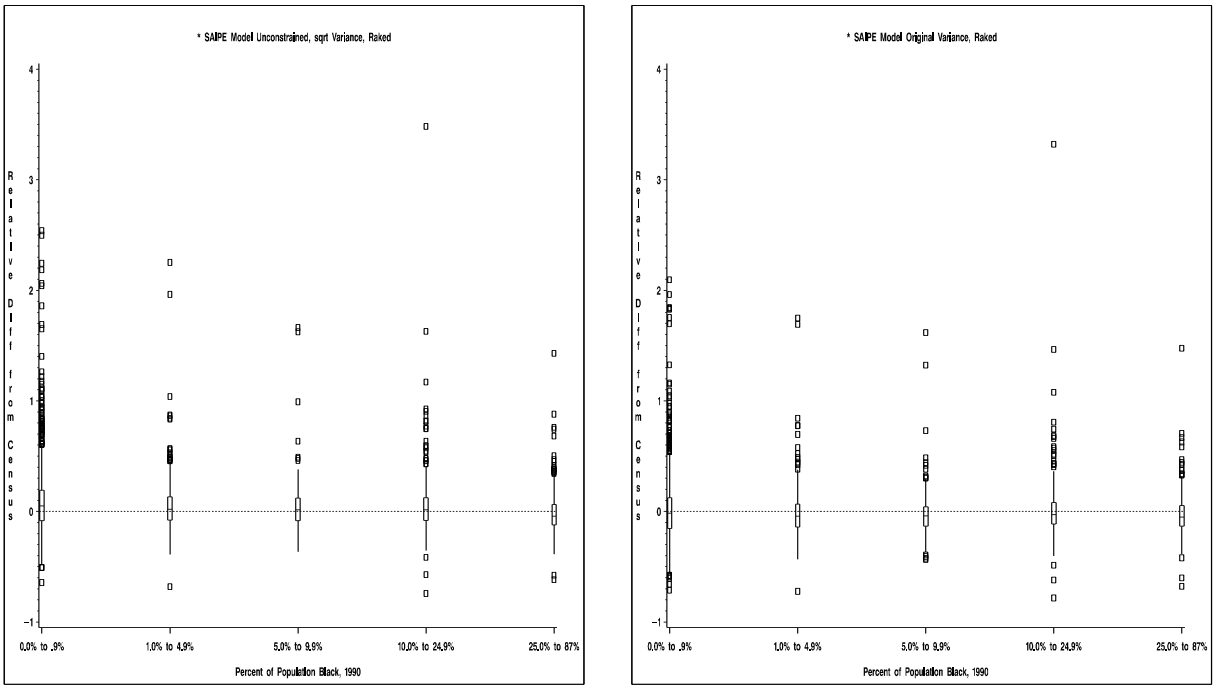


Figure 24 - Absolute Relative Difference from the Census, Compared by Percent of Population Black in 1990

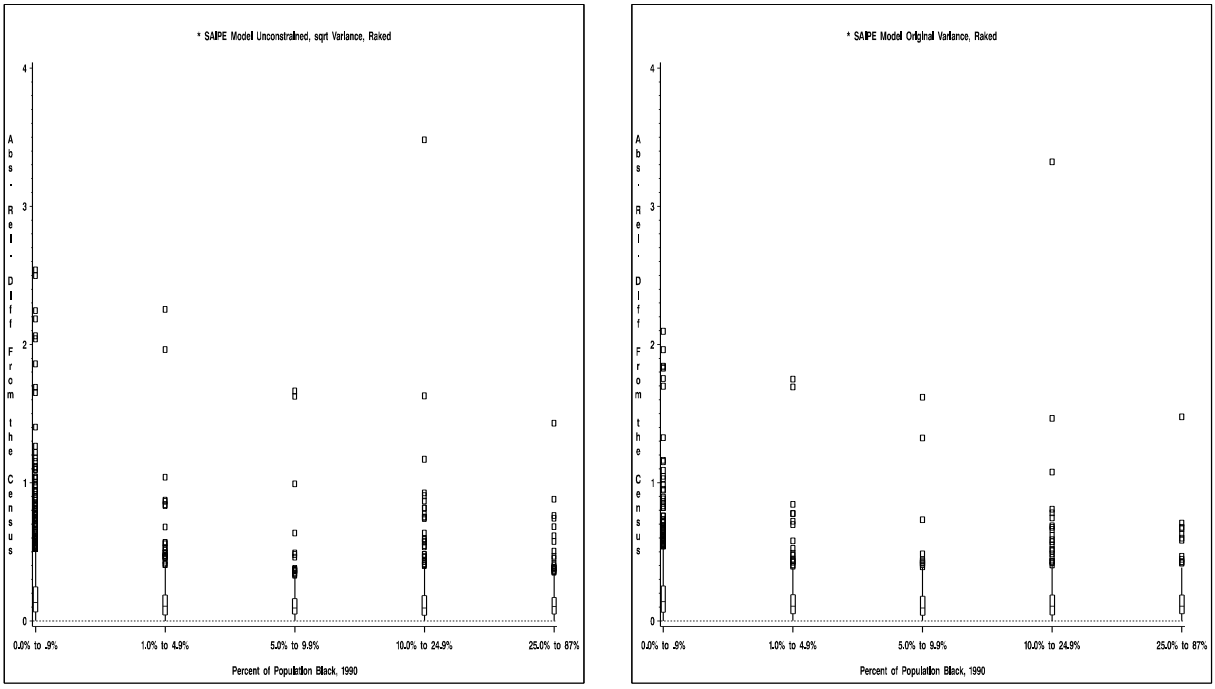


Figure 25 - Standardized Difference from the Census, Compared by ERS Persistent Rural Poverty 1960-1990

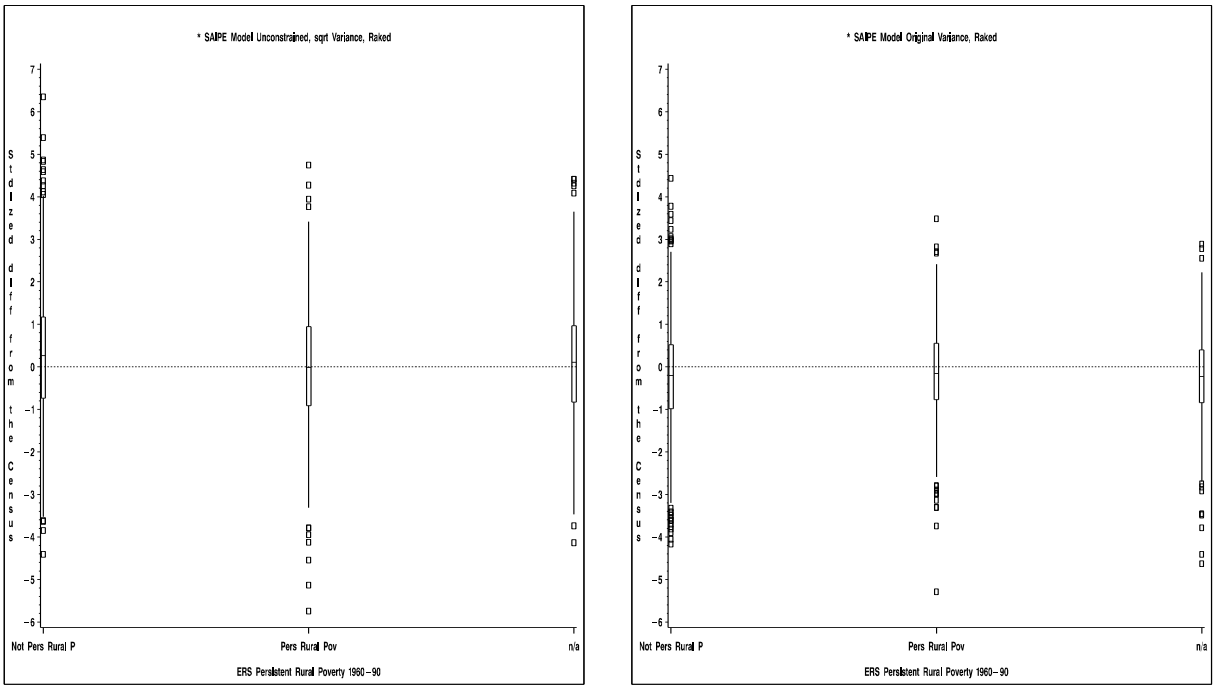


Figure 26 - Relative Difference from the Census, Compared by ERS Persistent Rural Poverty 1960-1990

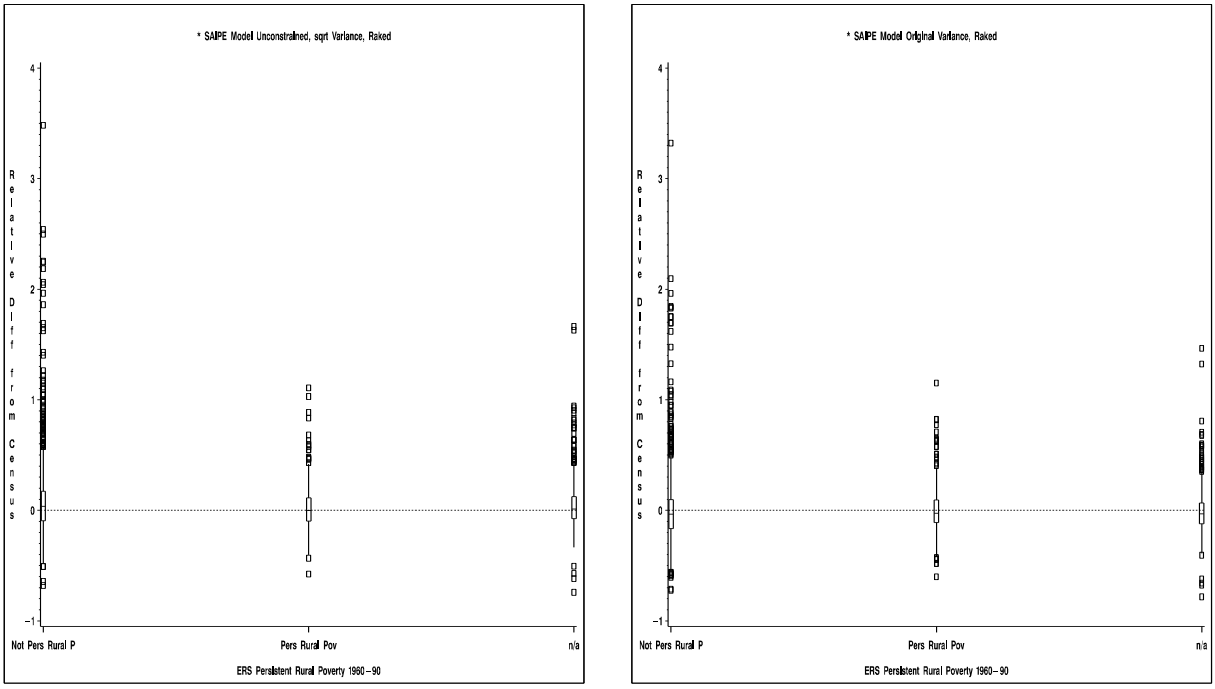


Figure 27 - Absolute Relative Difference from the Census, Compared by ERS Persistent Rural Poverty 1960-1990

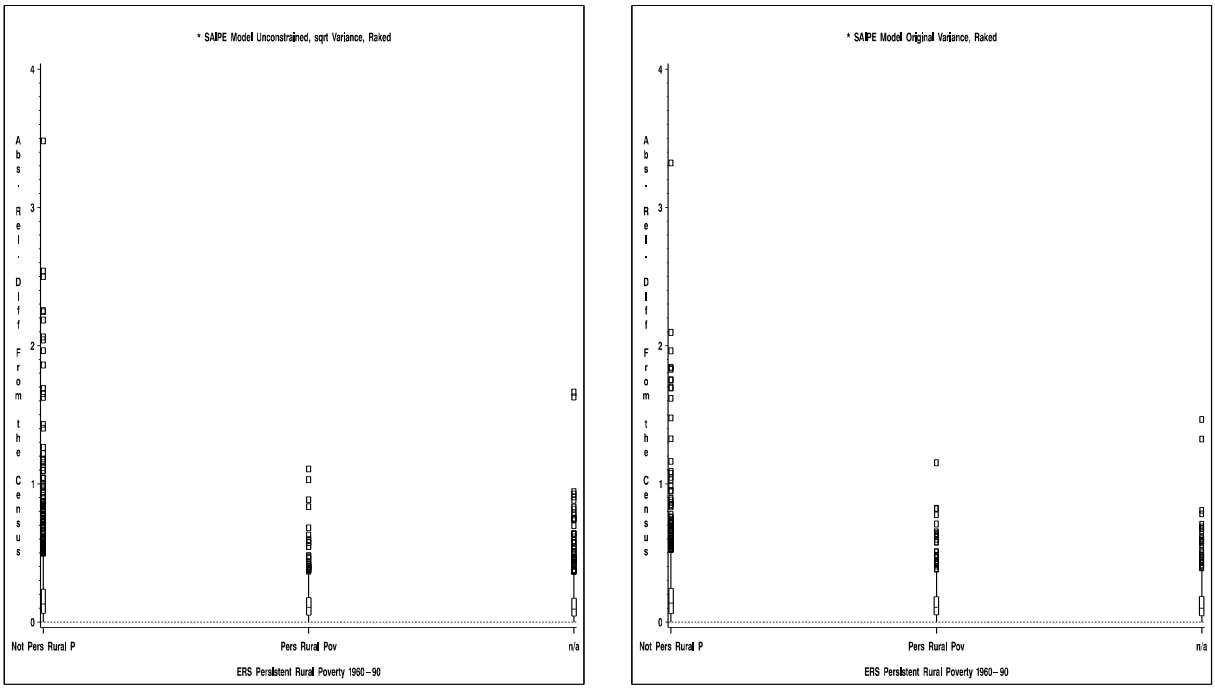


Figure 28 - Standardized Difference from the Census, Compared by ERS Economic Type for Rural Counties

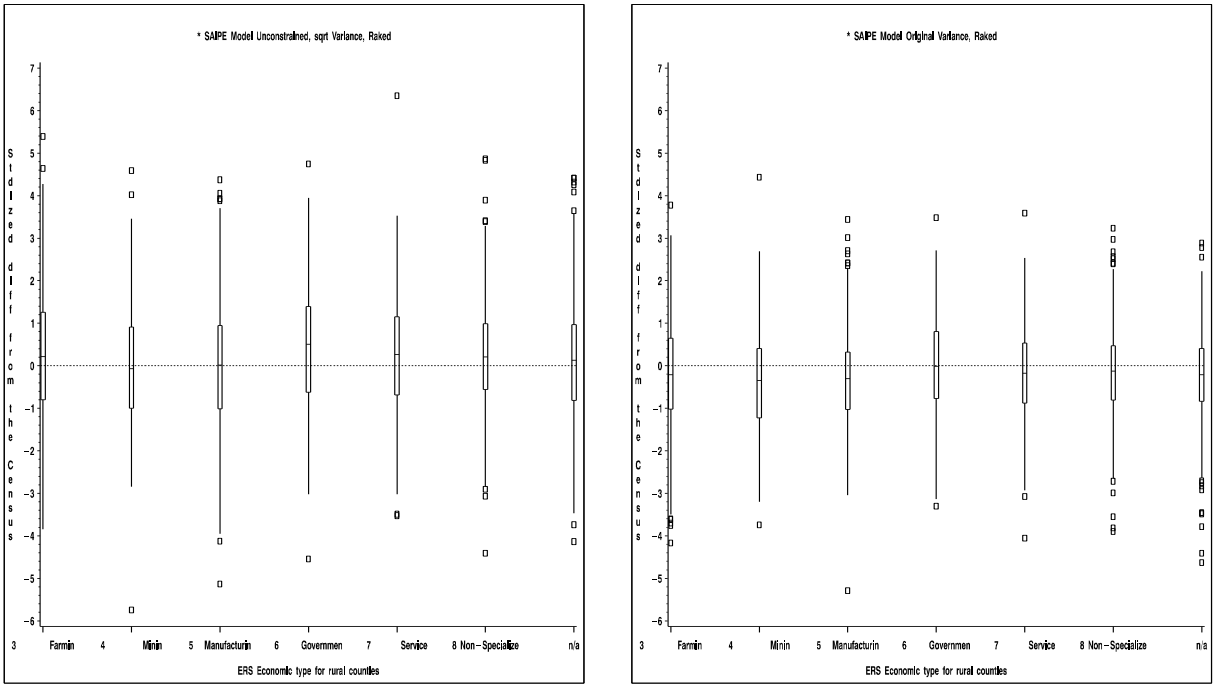


Figure 29 - Relative Difference from the Census, Compared by ERS Economic Type for Rural Counties

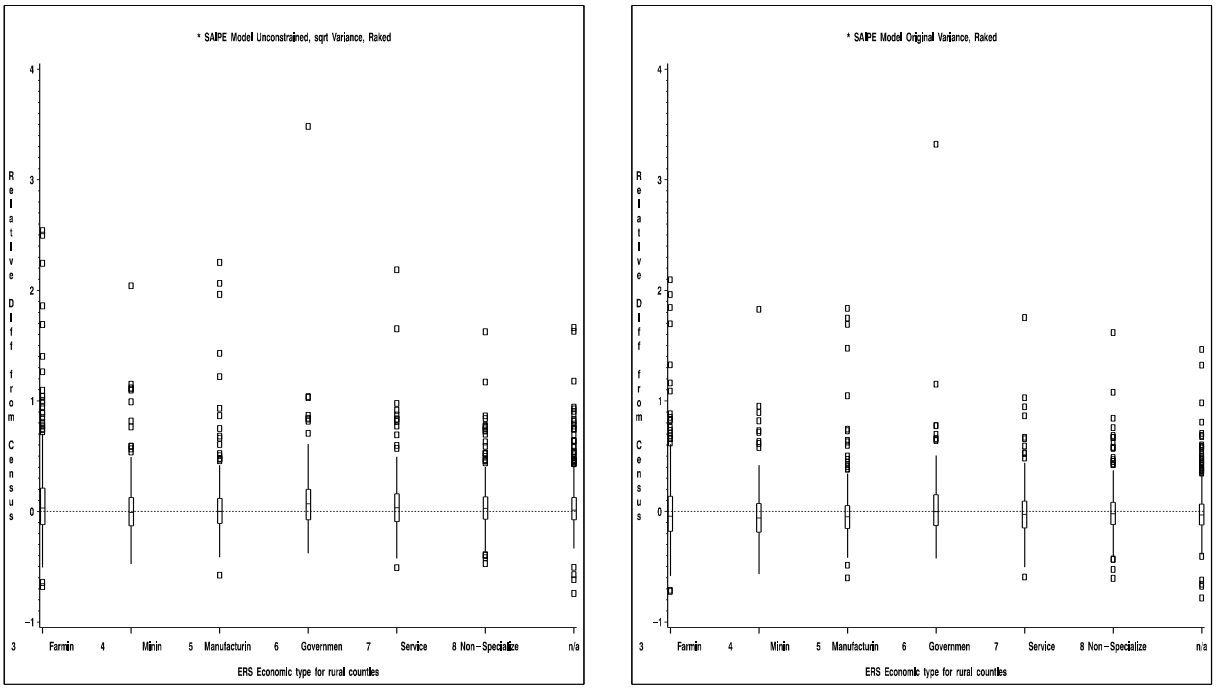


Figure 30 - Absolute Relative Difference from the Census, Compared by ERS Economic Type for Rural Counties

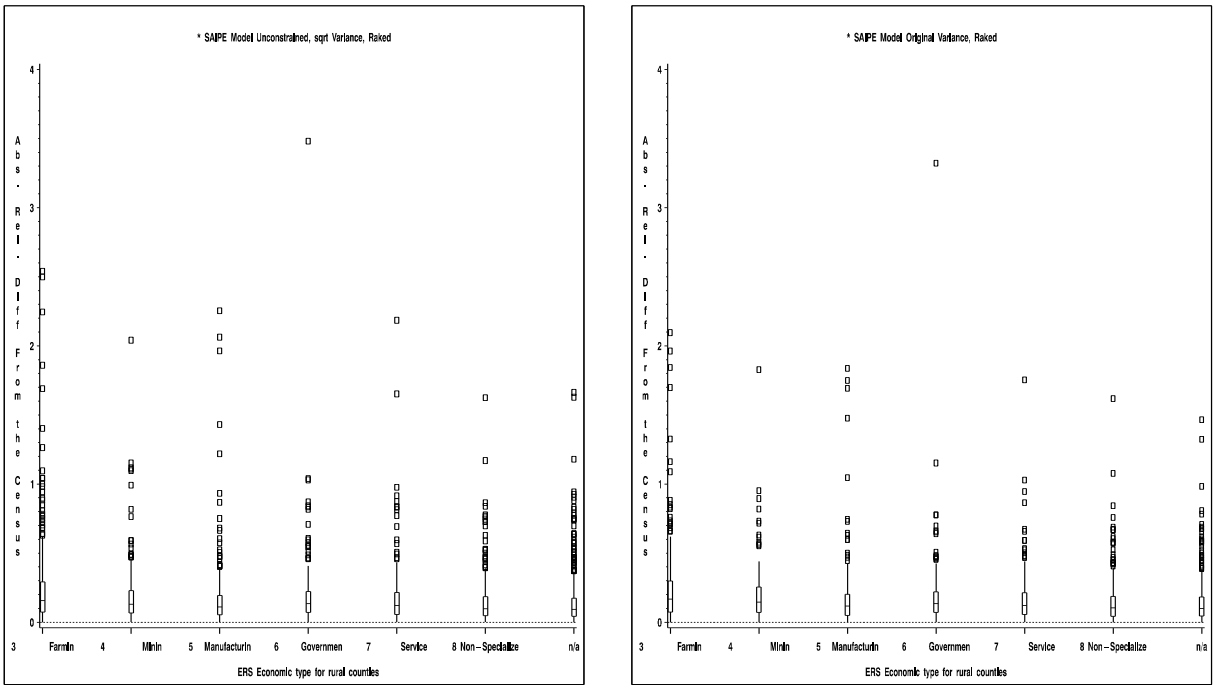


Figure 31 - Standardized Difference from the Census, Compared by Percent of Population in Group Quarters (1990 Census)

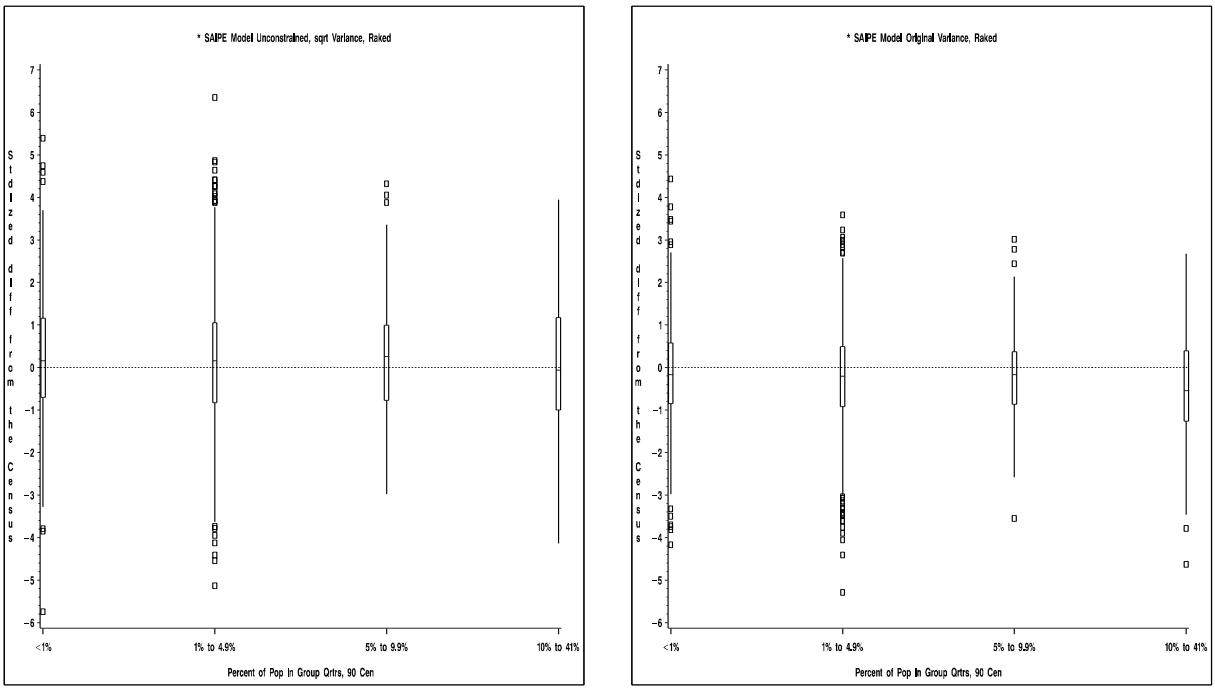


Figure 32 - Relative Difference from the Census, Compared by Percent of Population in Group Quarters (1990 Census)

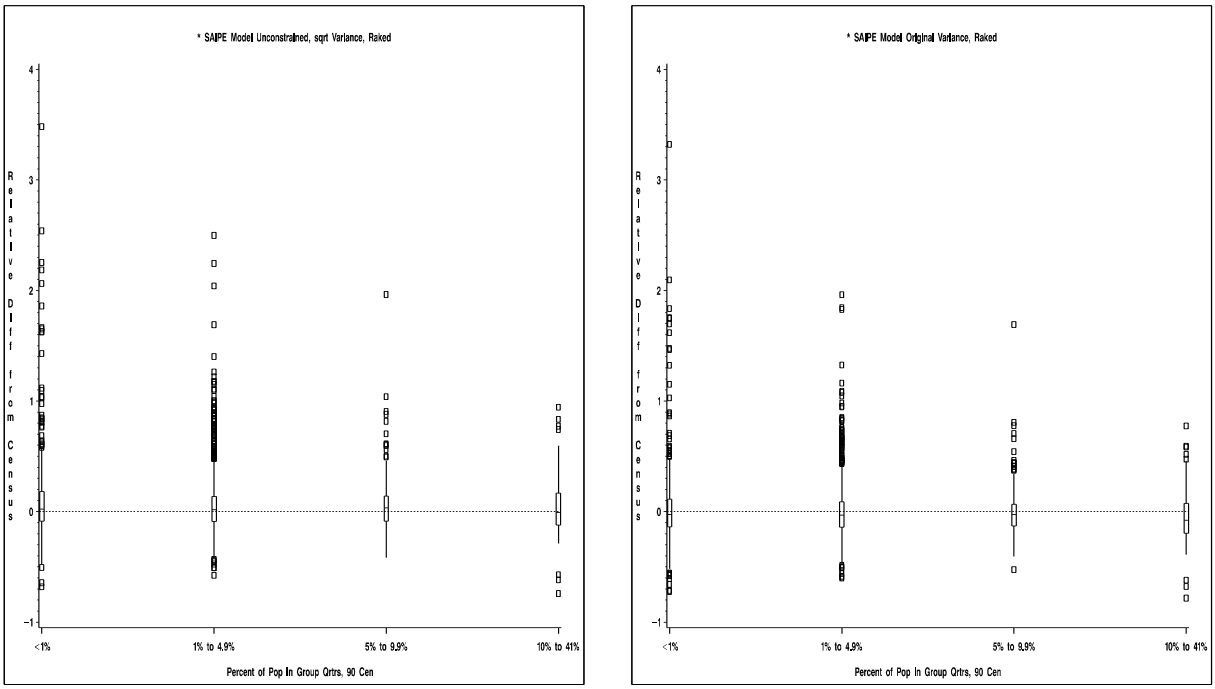


Figure 33 - Absolute Relative Difference from the Census, Compared by Percent of Population in Group Quarters (1990 Census)

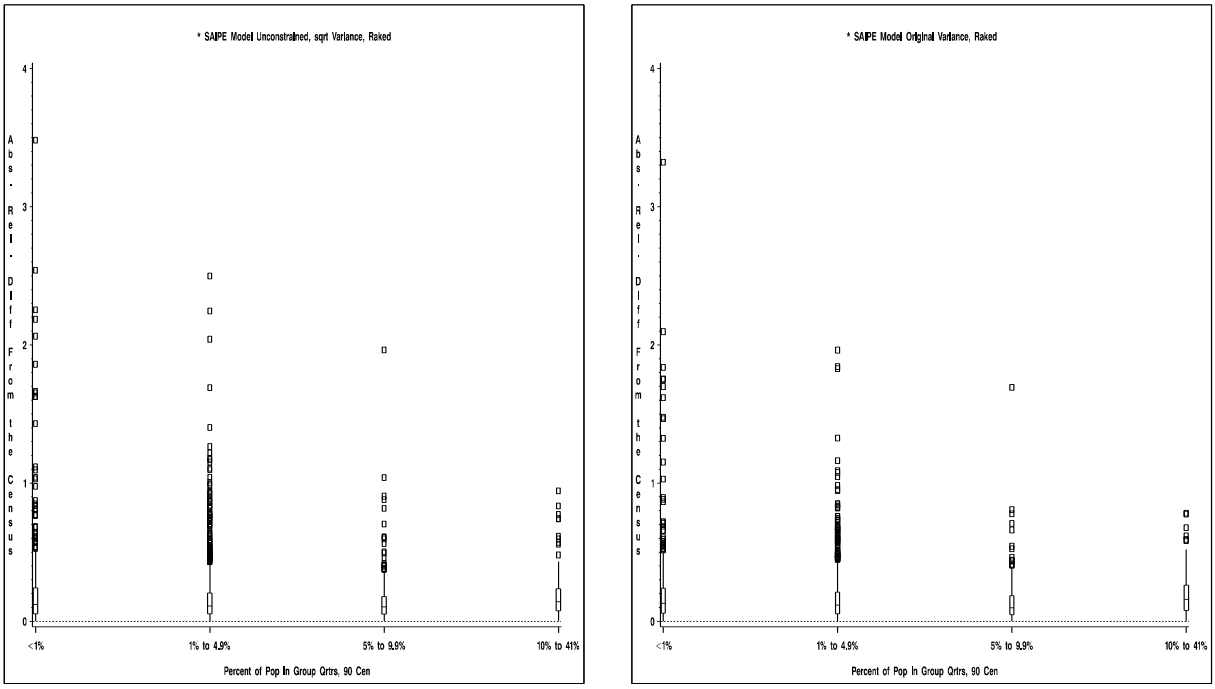


Figure 34 - Standardized Difference from the Census, for Virginia Independent Cities

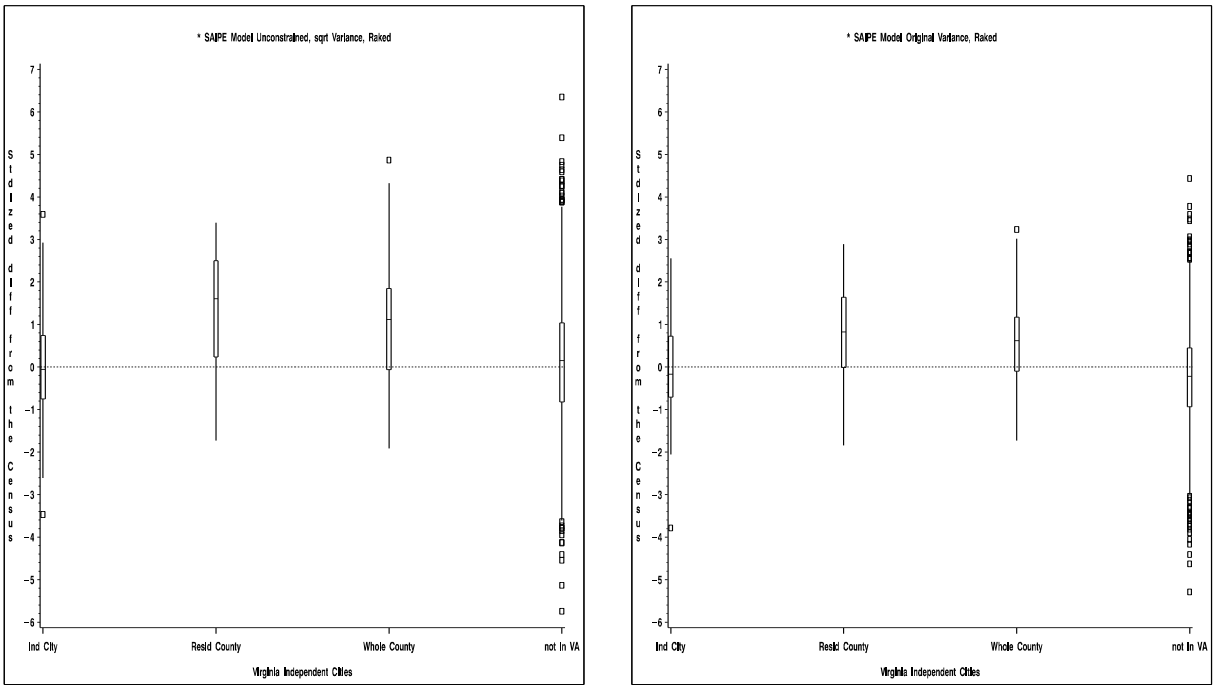


Figure 35 - Relative Difference from the Census, for Virginia Independent Cities

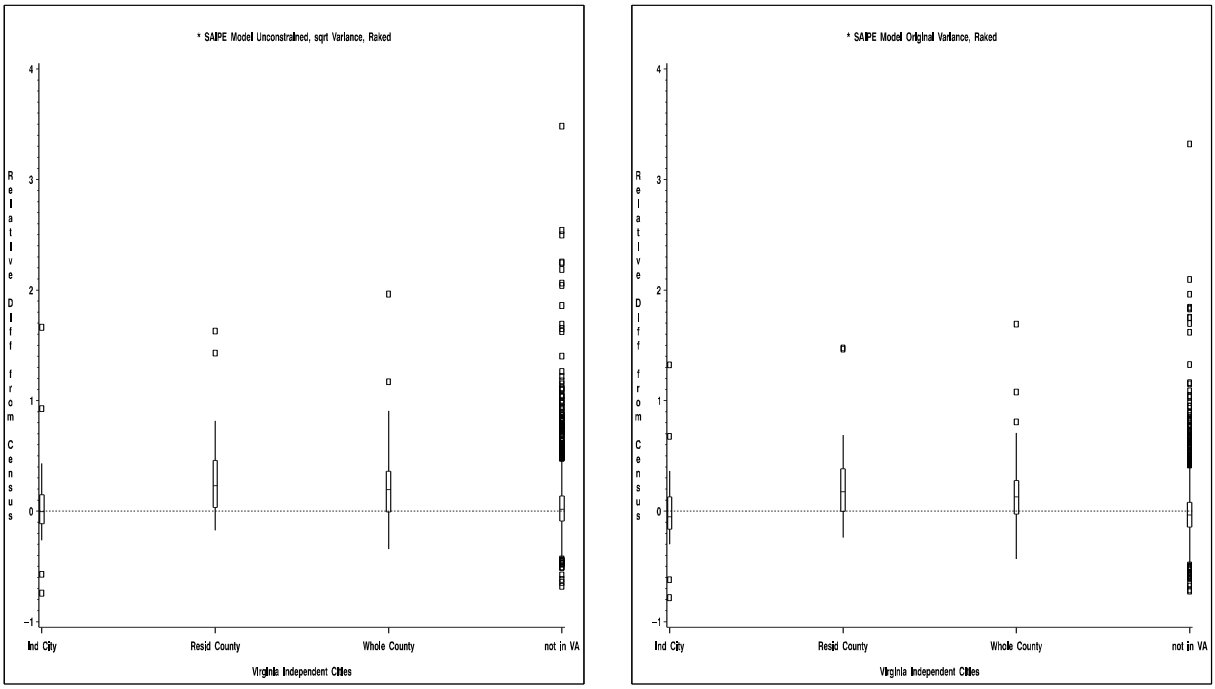


Figure 36 - Absolute Relative Difference from the Census, for Virginia Independent Cities

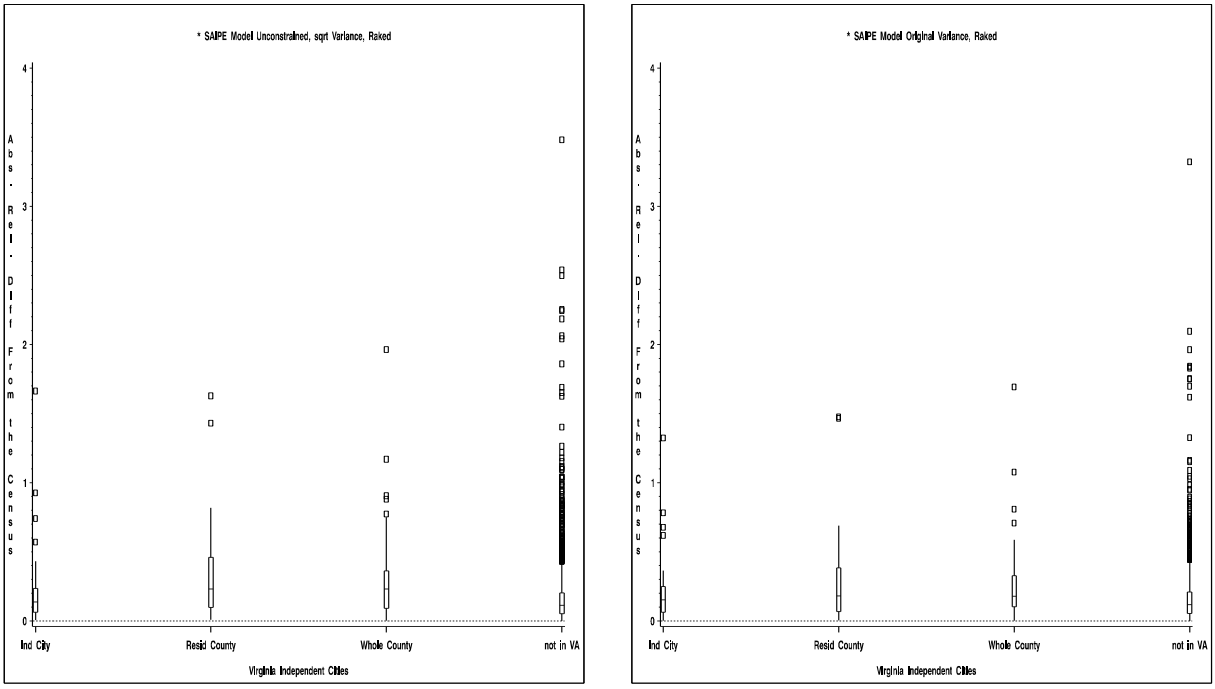


Figure 37 - Standardized Difference from the Census, Compared by Status in Current Population Survey, 1989-1991

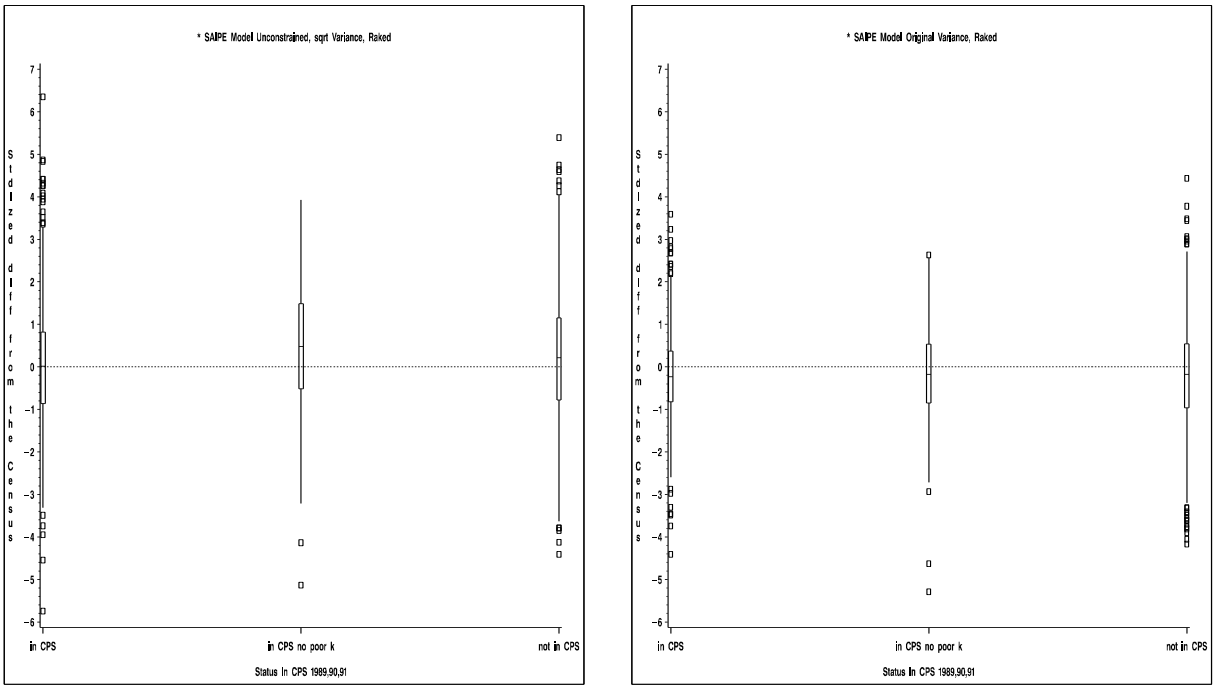


Figure 38 - Relative Difference from the Census, Compared by Status in Current Population Survey, 1989-1991

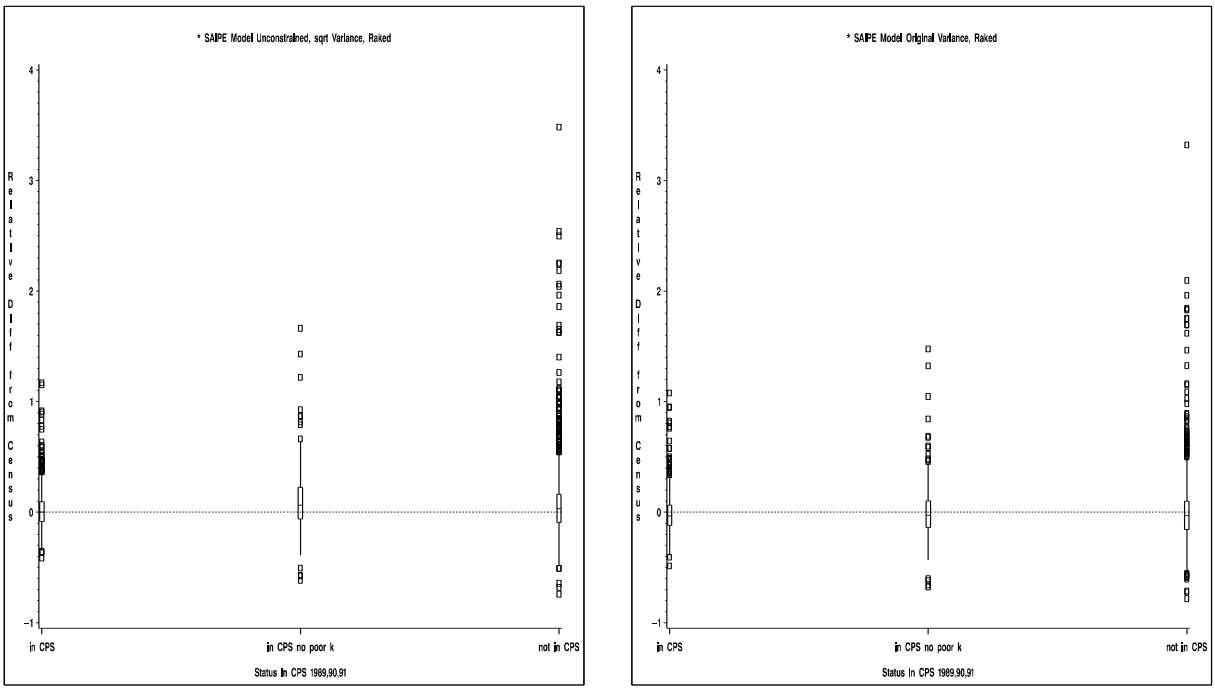


Figure 39 - Absolute Relative Difference from the Census, Compared by Status in Current Population Survey, 1989-1991

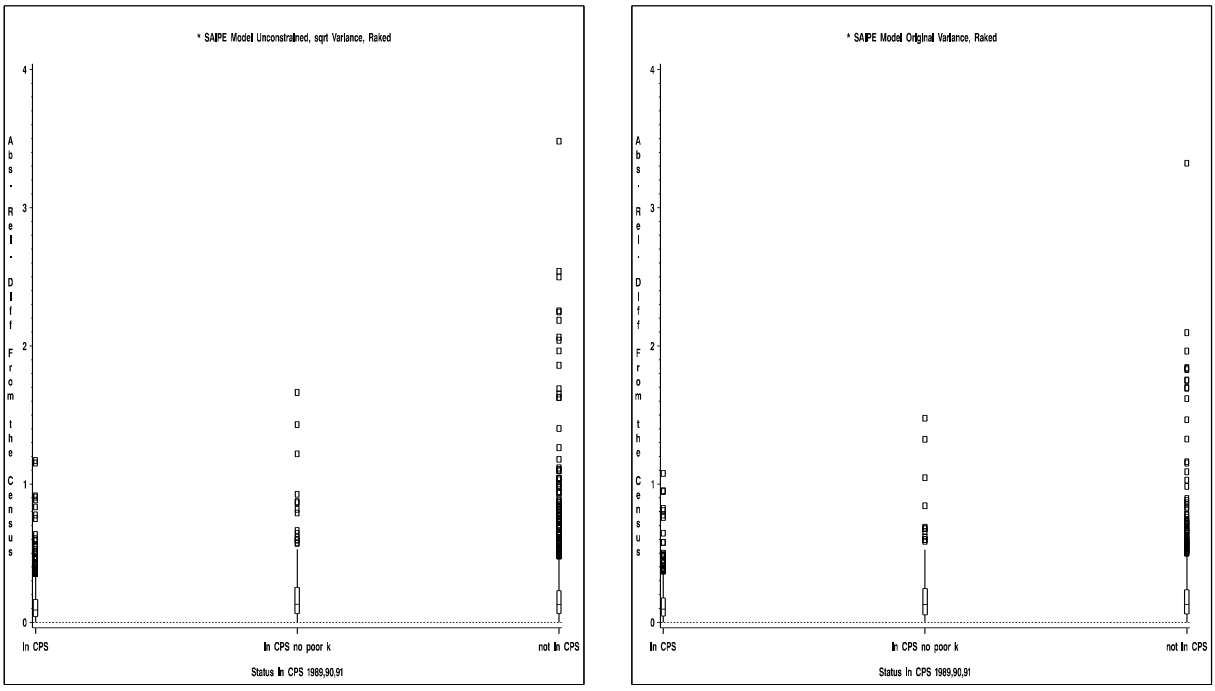


Figure 40 - Standardized Residuals from Original Model versus rank of CPS Sample Size

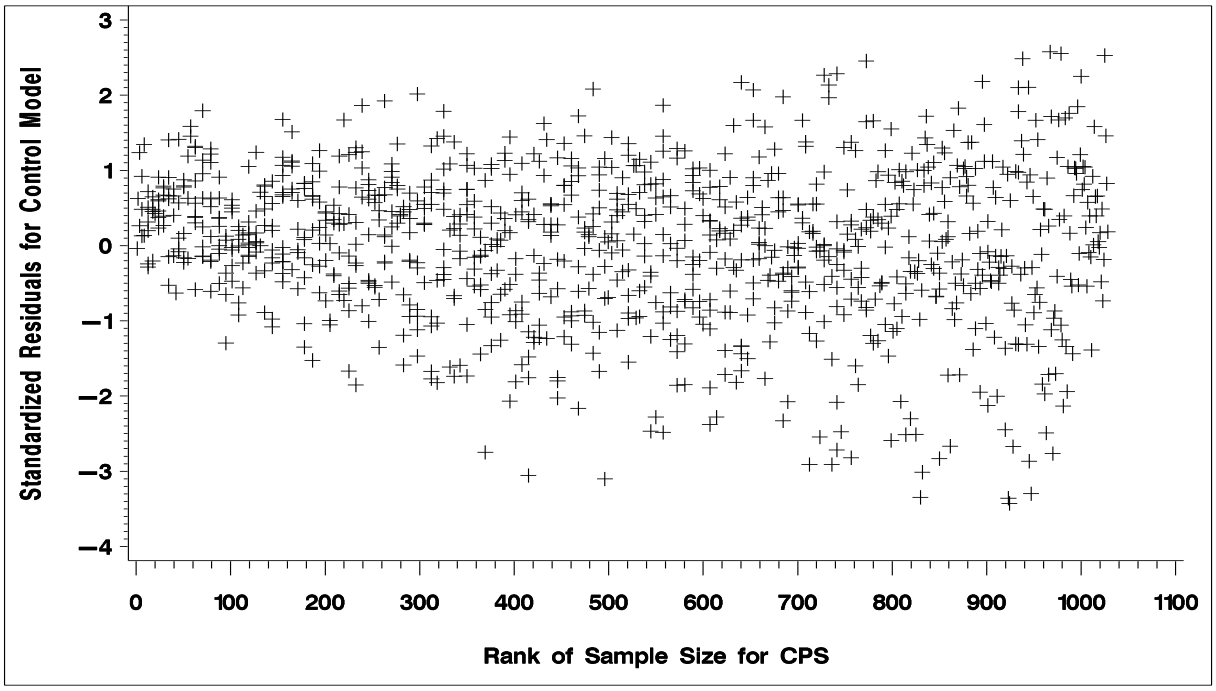


Figure 41 - Standardized Residuals from Proposed Model versus rank of CPS Sample Size