Review of Experimental Estimates from the U.S. Census Bureau's Small Area Health Insurance Estimates (SAHIE) Program

Prepared by

Gestur Davidson, Ph.D. Senior Research Associate
Michael Davern, Ph.D., Assistant Professor
Pamela Jo Johnson, Ph.D., Post Doctoral Associate

University of Minnesota's State Health Access Data Assistance Center (SHADAC)

http://www.shadac.org

June, 2005

SHADAC's review of the Census Bureau's Small Area Health Insurance Estimates:

We would like to thank the Small Area Estimates branch of the Housing and Household Economic Statistics Division of the US Census Bureau for giving us the opportunity to review their experimental small area health insurance estimates (SAHIE). We appreciate the difficulty associated with developing, producing, and improving small area estimation techniques as we have produced estimates that likely could be improved if we had more time and resources to do so. We also understand the importance of getting the best possible numbers into the public domain in order to better inform the public about health insurance coverage dynamics and to give health policy analysts another tool to use to monitor health insurance coverage.

We have three general conclusions. First, we think the Small Area Estimates branch of the Census Bureau has put together an excellent first set of small area estimates for counties in the United States. Second, we have some concerns with the current model that may have already been addressed (but not included in the short paper) or should be investigated as the program continues to progress and the estimates move from an experimental release to a more standard set of numbers that may someday be used in something as important as the State Children's Health Insurance Program (SCHIP) funding formula. And, third we think these estimates should be released to the public soon and that SHADAC should sponsor a conference call between the Census Bureau and other interested parties shortly before the estimates are released.

The following review is comprised of three sections. The first section outlines our concerns with the current SAHIE model and areas that may be appropriate for future research, including both comments on the existing model as well as questions we would like to see addressed. We realize that some of this work may have already been done, but making reference to it in the methods report would benefit readers' understanding. The second section of this review presents results from our empirical investigation comparing the Census Bureau results with those we obtained using our state survey data from three states: Missouri, Oklahoma, and Alabama. Finally, we conclude our review with recommendations.

I. County and State Health Insurance Coverage Estimation

We reviewed the SAHIE methods document prepared by Dr. Fisher (2005) and found this paper well written and methodologically rigorous. Indeed, it foreshadowed many of our concerns. In academia, we are accustomed to manuscripts that attempt to hide potential weaknesses but Dr. Fisher's paper made them more transparent. We appreciate his candor in presenting the limitations of the SAHIE model as it helped focus our review.

<u>Model specification issues</u>

1. Concerning the explanatory variable, "mean Log AGI / FPL":
    - Although empirically it is measuring the average of the log of the individual Income to Poverty Ratios (IPR's) of families in the county, the definition provided in the paper is ambiguous. We recommend "mean log of Income to Poverty ratio".

- As such, and as is seen below, this represents the log of the *geometric mean* of the individual family Income to Poverty ratios. For the j^{th} county with N_j families,

  "Mean Log (IPR)" $= 1/ N_j \times [\Sigma_i \ln (IPR_i)]$.

  The geometric mean of (IPR_1 , …, IPR_{Nj}) is defined as:

  $[IPR1 \times IPR_2 \times \cdots \times IPR_{Nj}]^{(1/Nj)}$

  thus log geometric mean = $1/ N_j \times [\Sigma_i \log (IPRi)]$.

  Why was the log of the geometric mean used as opposed to the log of the arithmetic mean?

- As a summary measure of the distribution of income in a county, this measure confounds the impact of a full ensemble of %POV indicator variables (see below), which are a more direct way to measure the distribution of an area's population by income. That is, one might choose to include either this log geometric mean—and perhaps also the variance of it—*or* the full ensemble of %POV indicator variables, but not both.

- In place of the "mean Log AGI / FPL" variate we would like to see an evaluation of whether the full ensemble of %POV variables would work better. For example you could try:

  - % of families < 50% FPL
  - % of families between 51% and 100% FPL
  - % of families between 101% and 150% FPL
  - % of families between 151% and 200% FPL
  - % of families between 201% and 250% FPL
  - % of families between 251% and 300% FPL
  - % of families > 300% FPL (reference category)

In empirical analyses of individual survey data relating the probability of being uninsured to this (individually-defined) set of income-to-poverty range variables, we generally find a strong non-linear trend. Specifically, the lowest income is associated with the highest likelihood of being uninsured and each successive step up the income range (IPR indicator level) has a lower probability of being uninsured up to about 250% FPL, after which there are no significant differences. This non-linear response pattern would suggest the desirability of including this set of IPR indicators. The SAHIE model includes just one of these measures—"% families 200-300% FPL". As it stands, this variable does not have a clear analytical reference group, since all lower income families and all higher income families are excluded. Their effects are also being expressed in the "mean Log AGI / FPL" variate and its variance. That is, it is interpreted as the *extra* impact of being in the 200%-300% FPL, *given* the values of the log of the geometric mean of FPL's and its variance.

2. We would also suggest using the tax return data (if this is possible) to compute interesting income information. For example, is it possible to compute:

    - % total AGI from agriculture
    - % total AGI from self-employment

3. Lacking these—or even if they are available—we would also insert a group of indicator variables to reflect the degree of urbanicity of the county, made from the so-called urban-influence or urban-rural continuum codes for counties.

Measurement Issues

It may be useful to think about what is being measured by "coverage" in this model. We would raise the following points concerning measurement:

1. There is great uncertainty in the literature—and in most researchers' minds—about what exactly is being measured in the CPS by coverage or no-coverage. It's useful to consider first taking the CPS at face value. That is,

    coverage = private full year coverage + public full year coverage
    + private part-year coverage + public part-year coverage

    Concentrating on the full-year versus part-year coverage difference, we would expect a different set of predictors for these two constructs. Full-year coverage will be associated with reasonably high-paying private jobs and all public jobs. The proportion of a county's total employment represented by these sources would not be expected to change rapidly, or one might label them as more 'structural' in nature. Part-year coverage, on the other hand, is on the face of it likely to reflect more dynamic events like the loss of a job, a change in job type, or the end or beginning of welfare benefits. These things suggest some likely covariates for the model:

    - % of employment from manufacturing
    - % of employment from the public sector
    - the unemployment rate
    - the change in the unemployment rate

2. Given the household rotation schedule of the CPS, is your count of housing units (ki) "unique" housing units or a count of housing units per year?

3. As an aside, we are curious as to how variance estimation with the ASEC has been impacted by the SCHIP sample enhancement with more of the ASEC interviews coming from outgoing/incoming November, February and April CPS households who have minority group members and/or kids in the household. We have also noted that even with this change in sample design the generalized variance parameters, that you mention you use for

comparative purposes (on the top of page 8), have not changed (e.g., the year to year covariance is still set at .35).

4. On page 4, middle section, you mention state and federal employees. Is it possible to get an estimate of the workforce that is employed in the public sector (from either the decennial census or the Wage data from BLS)?

5. The "Log Child Medicaid Rate" variable. How is eligibility determined? By eligibility, do you mean enrollees? We are aware that among state administrative data users these words (eligibles and enrollees) are used synonymously. We wanted to know if you are using eligibility in this same sense, or if not, what your criteria for eligibility are?

6. Why only West and South regional dummies? For modeling insurance coverage you may want to consider a Mexican border-state dummy (TX, AZ, NM and CA) instead of either of these. We have found in past analyses that these states have higher rates of coverage.

7. Top of page 8 and the use of the generalized variance parameters for comparative purposes. We have been doing some work that indicates that the generalized variance parameters for health insurance are biased considerably (the variances are too small). We would be happy to share this work with you.

8. Finally, we think the concluding section would benefit from a more thorough literature review of the empirical findings concerning health insurance coverage in the United States. We note that you acknowledge this by saying you would like to collaborate with an expert on health insurance coverage in general. We agree and hope you continue to include us as one of your experts.

II. Empirical Findings

To empirically assess the SAHIE small area estimates, we compared county-level SAHIE estimates of uninsurance to a set of county-level small area estimates that we generated independently using survey data from three states: Missouri, Alabama, and Oklahoma. Each state was examined separately. We compared the Census Bureau estimates to three different uninsurance estimates derived with our model based on differing measures of health insurance coverage: (1) uninsured at the point in time of the survey (point in time), (2) uninsured at any point during the prior 12 months (any time), and (3) uninsured for the entire past twelve months (all year). This latter measure—uninsured all year—is what the CPS-ASEC is purportedly measuring, although as noted previously few analysts actually think that the CPS-ASEC is *only* measuring this. [The Census Bureau's other demographic survey the SIPP produces full year uninsured estimates that are roughly half of the CPS estimates]. While we produced small area estimates using three distinct models, we only report the results from our "raked mixture model". The correlations presented below were weighted by county population size.

Correlation Coefficients of SHADAC estimates with SAHIE estimates by state

Oklahoma

Point in time	.52
Any time	.52
All Year	.52

Alabama

Point in time	.52
Any time	.60
All Year	.56

Missouri

Point in time	.56
Any time	.61
All Year	.56

- These correlations seem reasonable.

- We examined whether the residuals between the Census estimates and the SHADAC estimates varied by whether the county was sampled in the CPS and they did not.

- The Census Bureau estimates are almost universally higher than SHADAC estimates of full-year uninsured. The Census estimates also tend to be higher than SHADAC point-in-time and SHADAC any-time-during-the-year uninsured. Especially our first finding suggests that the CPS-ASEC estimates are capturing more respondents than the uninsured all year, its face value measure.

See the Appendix for more details.

III. Summary and Recommendations

We believe that the Census Bureau Small Area Estimate's branch has produced a reasonable set of county-level estimates of the uninsured for the U.S. This type of work is difficult for two reasons. First, small area estimation techniques rely on statistical models to produce estimates. Model-based estimates are open to additional covariates and alternative parameterizations that can impact the final estimates. We presented several issues to consider with respect to the model specification in our review. Second, many states collect their own survey data on health insurance coverage, and it varies a great deal from the Census Bureau's CPS estimates. As a result, any model that uses the CPS as a basis will be viewed as suspect by many states. Moreover, many states may be upset by the release of additional estimates of uninsurance, especially if they differ from what they already have. Our recommendations follow from both of these difficulties associated with this work.

1. The Census Bureau should continue to research alternative model specifications. This includes altering the covariates used in the model to see whether they make a difference. We suggested several specific variables to consider for further investigation.

2. We would encourage the Census Bureau to perform a more thorough literature review for the SAHIE paper (section 8.4) in order to critically review and systematically document potential variables for model inclusion and what the expected sign of those variables might be. We are happy to participate or help in this endeavor.

3. We also recommend that you contact staff at the Agency for Healthcare Research and Quality (AHRQ). They have been building a small area model for the Medical Expenditure Panel Survey-Insurance Component (MEPS-IC) and Household Component (MEPS-HC). The IC is an employer survey and AHRQ staff know the covariates of employer offer rates and the strengths/limitations of economic census, Bureau of Labor Statistics data and the Internal Revenue Service data for predicting employer offer rates. Drawing on this knowledge could be useful to your model.

4. Our final recommendation is that the Census Bureau and SHADAC continue to work together in releasing these numbers. SHADAC has contacts in almost every state with individuals who are responsible for state survey data and/or who need to know the number of uninsured people in their state. Of course, these will be the very people who will be interested in your data release. They are also likely to be the people who may give you any critical feedback as many of them are sophisticated data analysts who know the strengths and weaknesses of the CPS data for deriving state health insurance estimates. Working with SHADAC to set up a couple of conference calls that bring the state analysts together with the Census Bureau's Small Area Estimation experts will help make this process run more smoothly. We recommend releasing a technical paper (similar to Dr. Fisher's) to the telephone call participants before the actual estimates are released to provide the states with a "heads up" and an opportunity to think through and formulate questions that will likely benefit all.

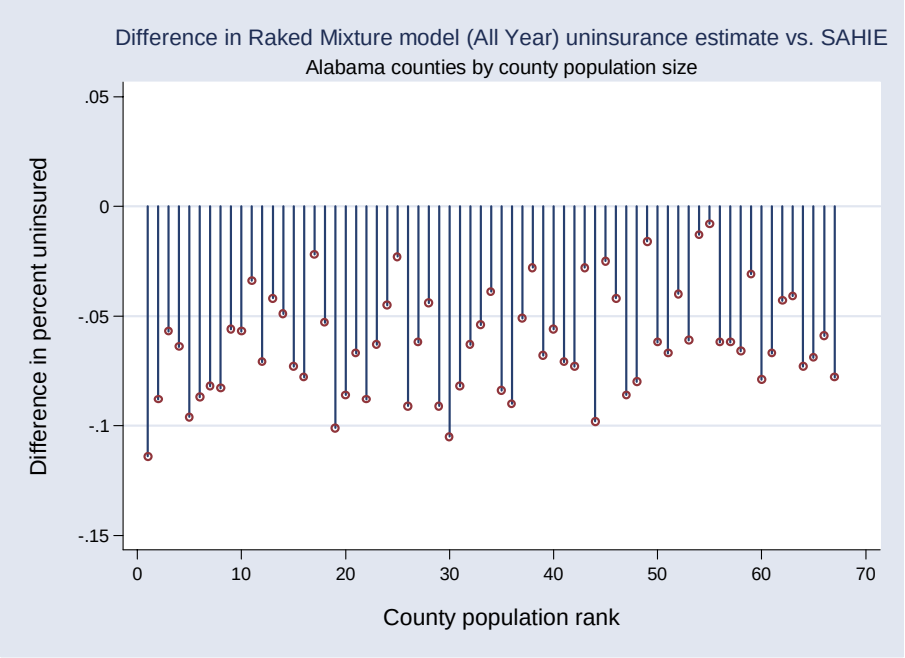
We would like to thank the Census Bureau Small Area Estimates branch for the opportunity to comment on this innovative, informative, and important research project. We have learned a great deal from the experience and have benefited from your frank write up of the current small area estimates strengths and weaknesses. We hope you have benefited from our review as well.

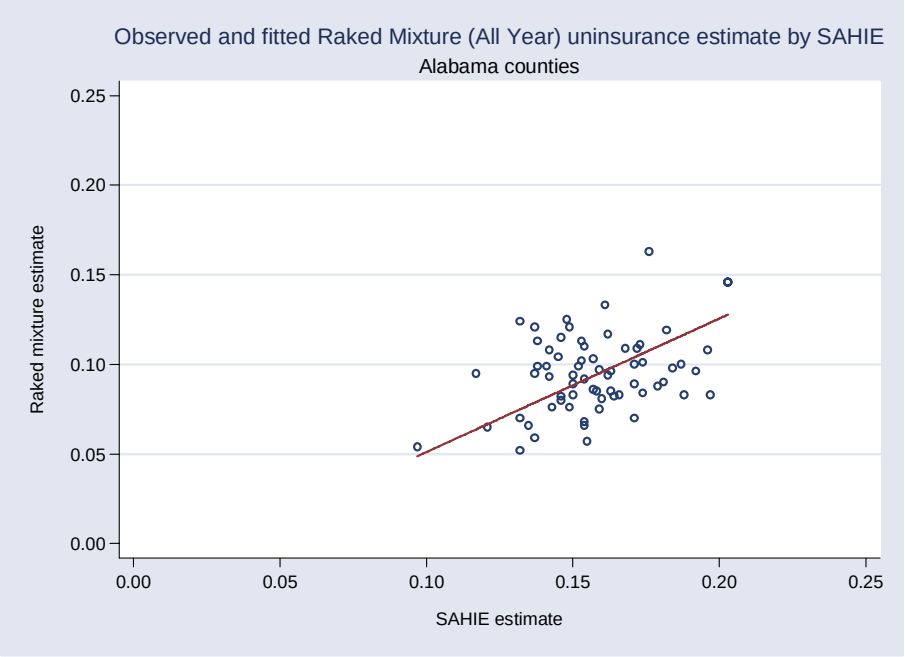
Appendix

Alabama

Differences in uninsurance estimates using raked mixture (All Year) model compared with SAHIE estimate by county population rank, Alabama.

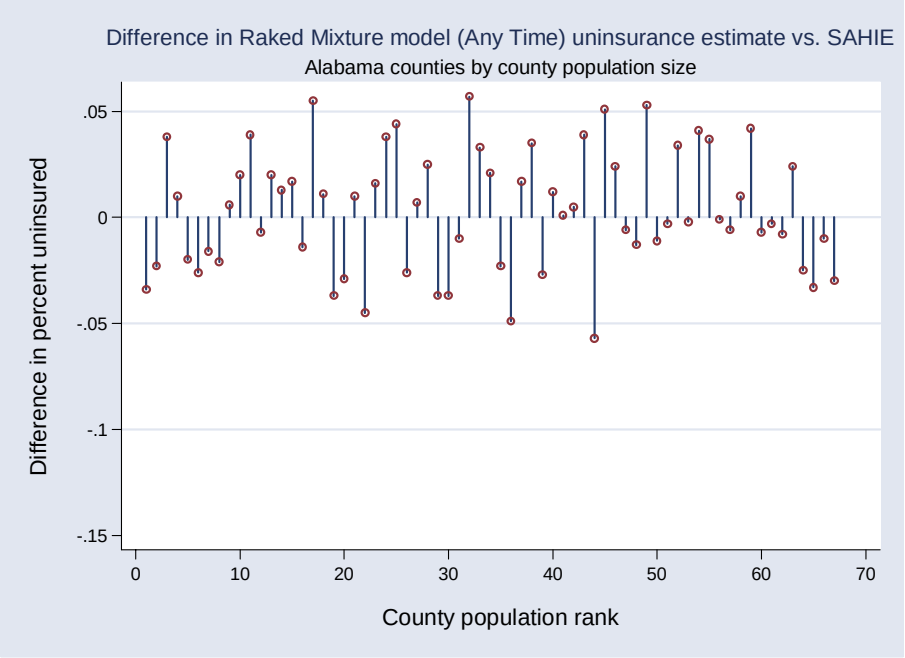
	Obs	Mean difference	Std. Dev.	Min	Max
All counties	67	-0.062	0.021	-0.114	-0.008
CPS counties	24	-0.061	0.020	-0.098	-0.008
Non-CPS counties	43	-0.062	0.023	-0.114	-0.016

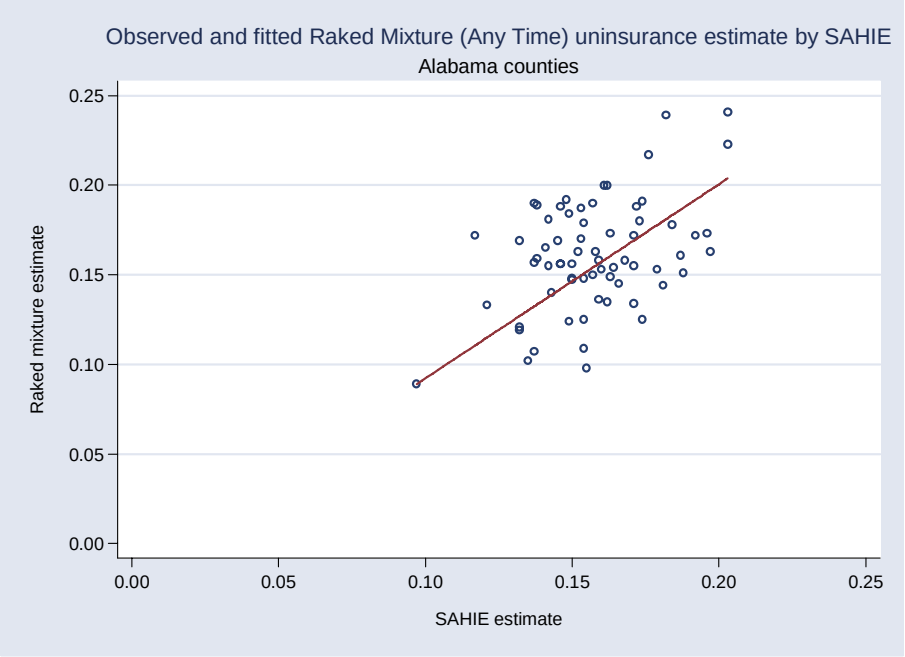




Differences in uninsurance estimates using raked mixture (Any Time) model compared with SAHIE estimate by county population rank, Alabama.

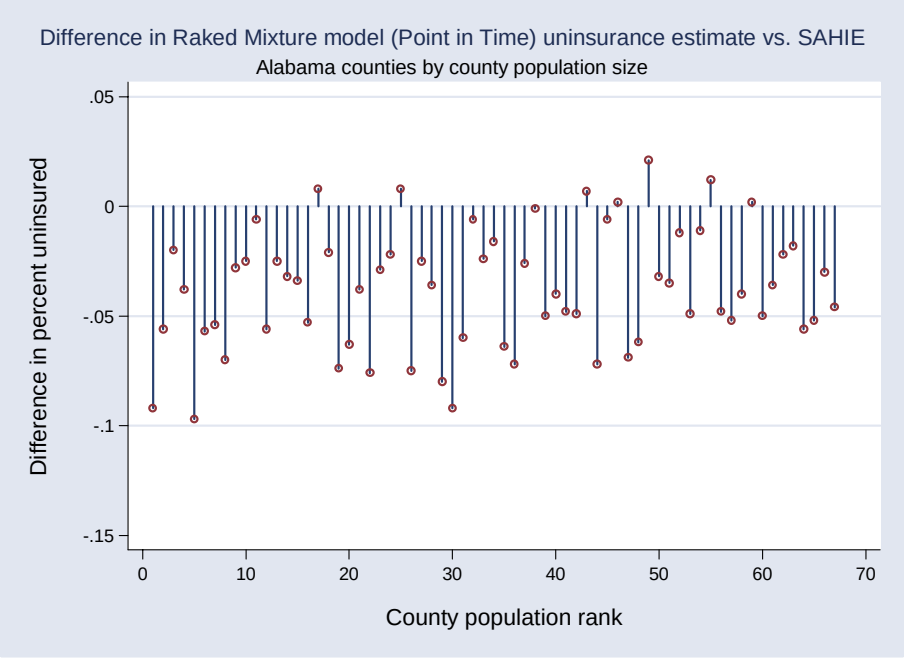
	Obs	Mean difference	Std. Dev.	Min	Max
All counties	67	-0.004	0.027	-0.057	0.057
CPS counties	24	-0.008	0.025	-0.057	0.055
Non-CPS counties	43	0.005	0.028	-0.049	0.057

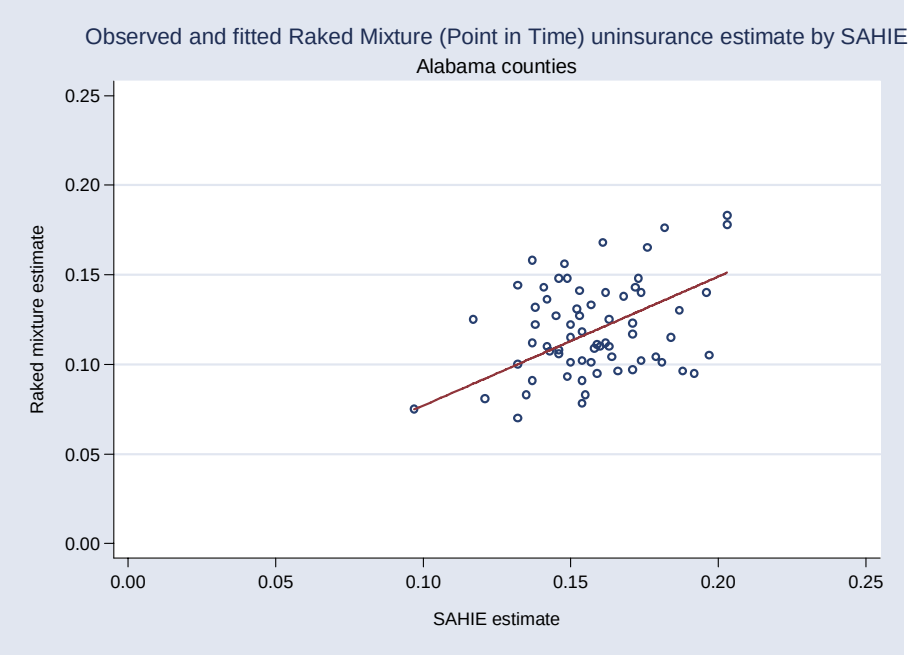




Differences in uninsurance estimates using raked mixture (Point in Time) model compared with SAHIE estimate by county population rank, Alabama.

	Obs	Mean difference	Std. Dev.	Min	Max
All counties	67	-0.037	0.022	-0.097	0.021
CPS counties	24	-0.036	0.019	-0.076	0.012
Non-CPS counties	43	-0.039	0.028	-0.097	0.021

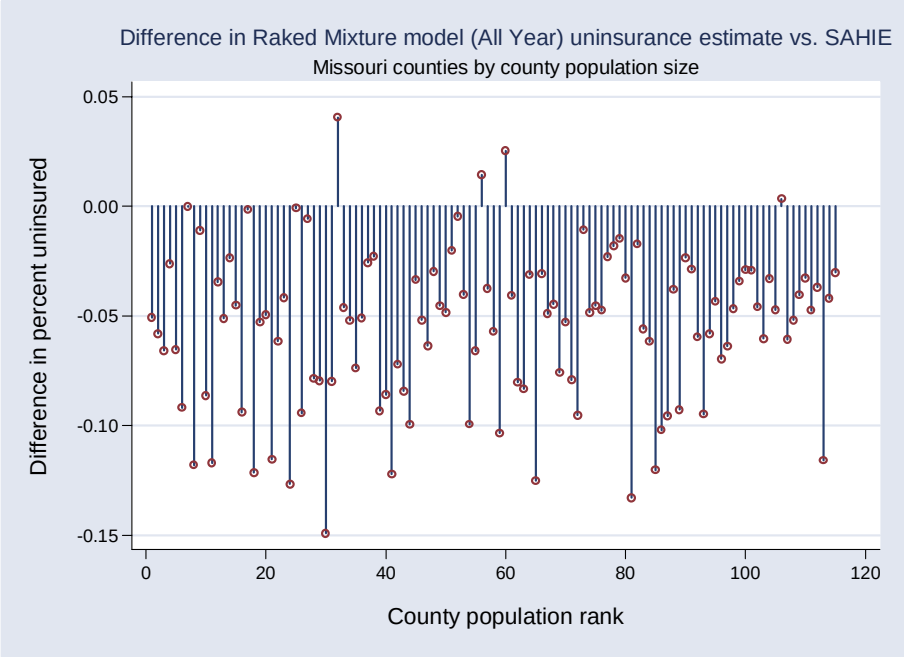


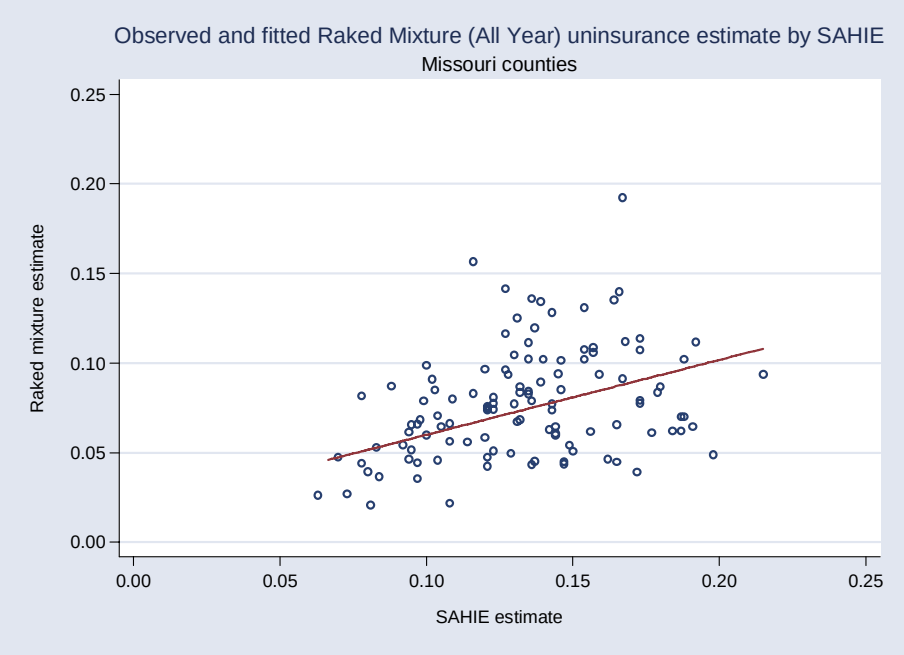


Missouri

Differences in uninsurance estimates using raked mixture (All Year) model compared with SAHIE estimate by county population rank, Missouri.

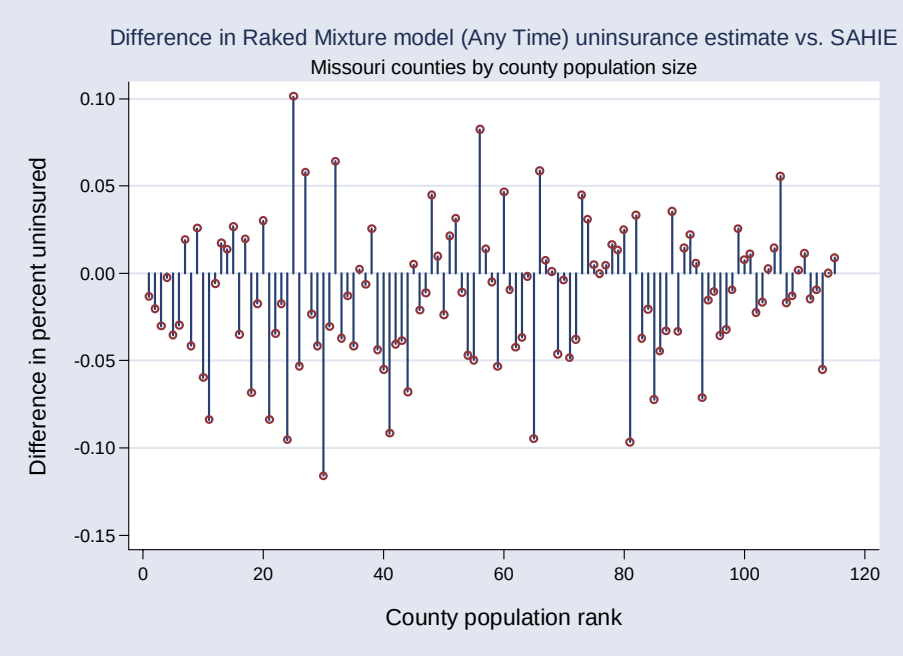
	Obs	Mean difference	Std. Dev.	Min	Max
All counties	115	-0.049	0.029	-0.149	0.041
CPS counties	24	-0.043	0.026	-0.116	0.014
Non-CPS counties	91	-0.060	0.033	-0.149	0.041

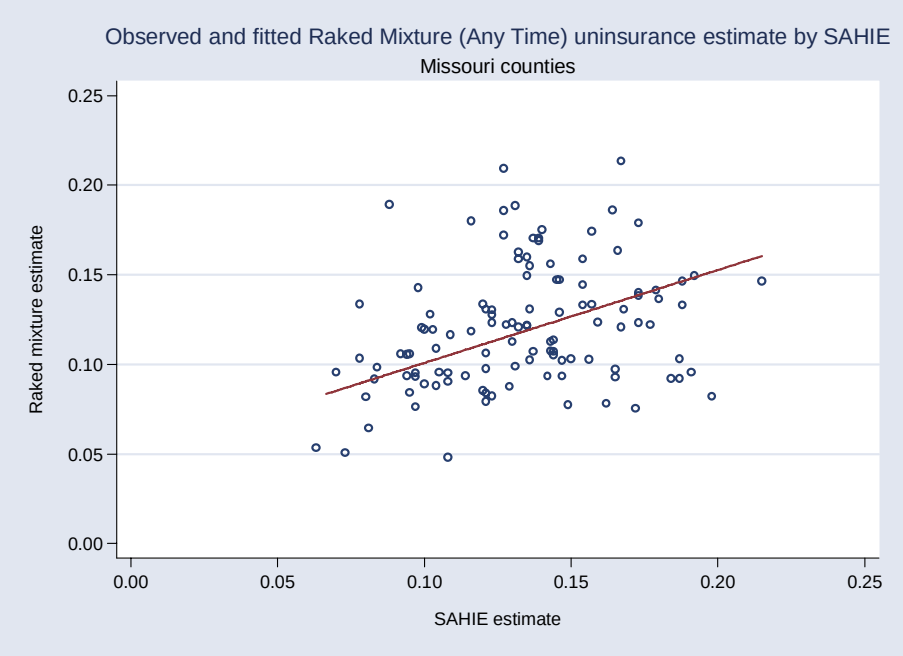




Differences in uninsurance estimates using raked mixture (Any Time) model compared with SAHIE estimate by county population rank, Missouri.

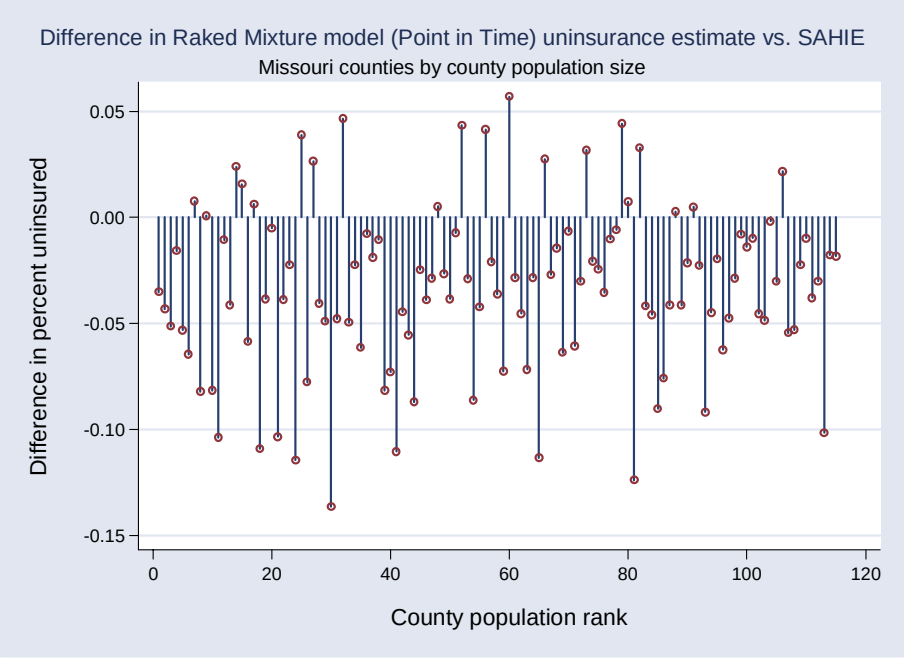
	Obs	Mean difference	Std. Dev.	Min	Max
All counties	115	-0.006	0.029	-0.116	0.101
CPS counties	24	-0.002	0.022	-0.055	0.082
Non-CPS counties	91	-0.016	0.037	-0.116	0.101

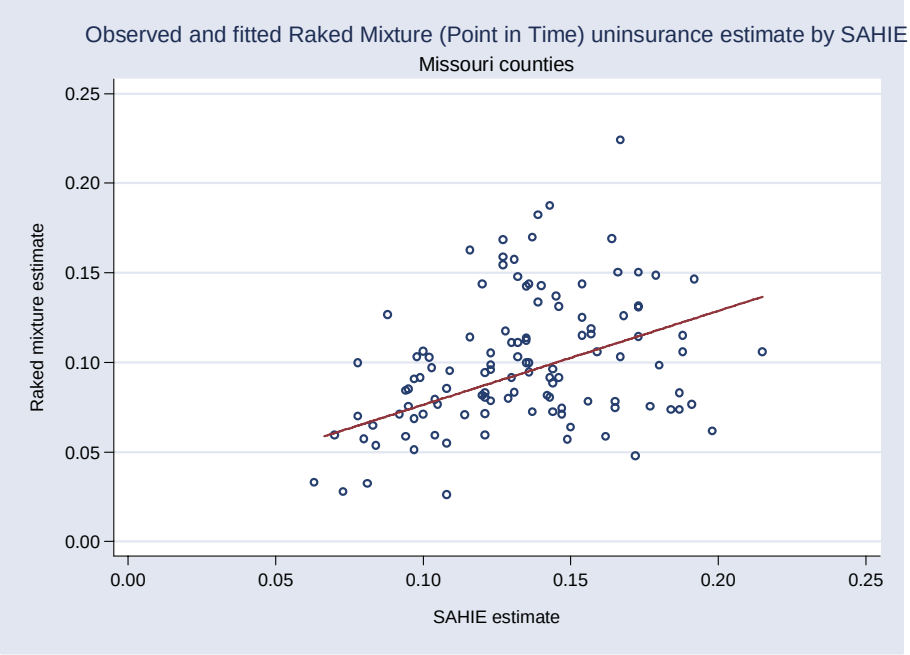




Differences in uninsurance estimates using raked mixture (Point in Time) model compared with SAHIE estimate by county population rank, Missouri.

	Obs	Mean difference	Std. Dev.	Min	Max
All counties	115	-0.031	0.032	-0.136	0.057
CPS counties	24	-0.028	0.028	-0.102	0.043
Non-CPS counties	91	-0.038	0.038	-0.136	0.057

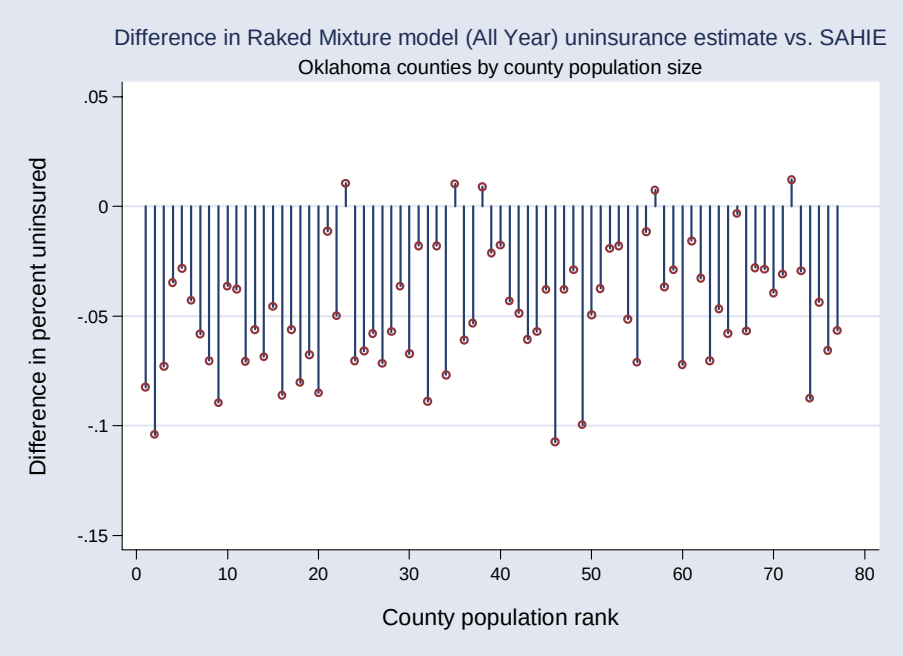


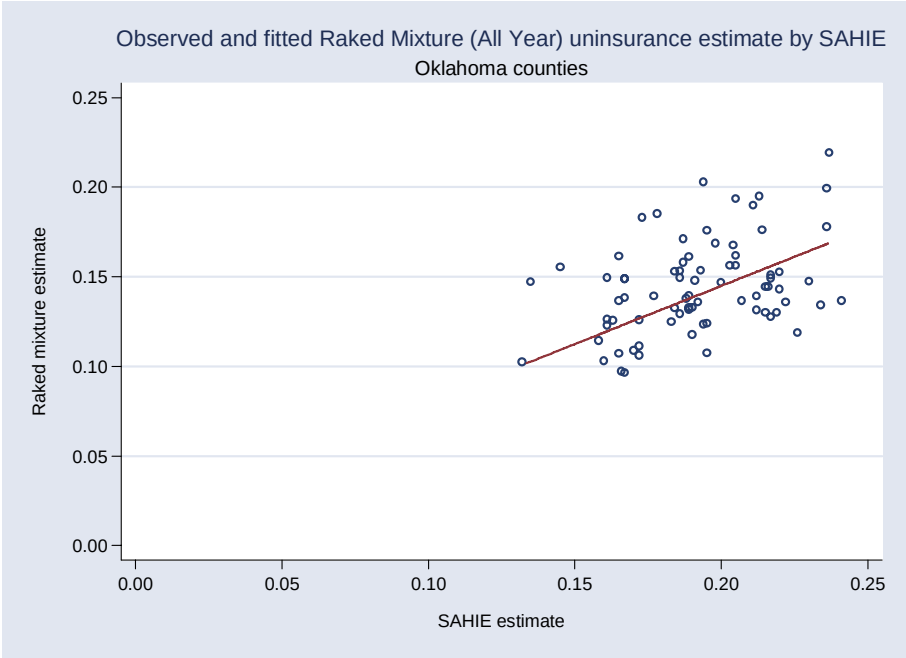


Oklahoma

Differences in uninsurance estimates using raked mixture (All Year) model compared with SAHIE estimate by county population rank, Oklahoma.

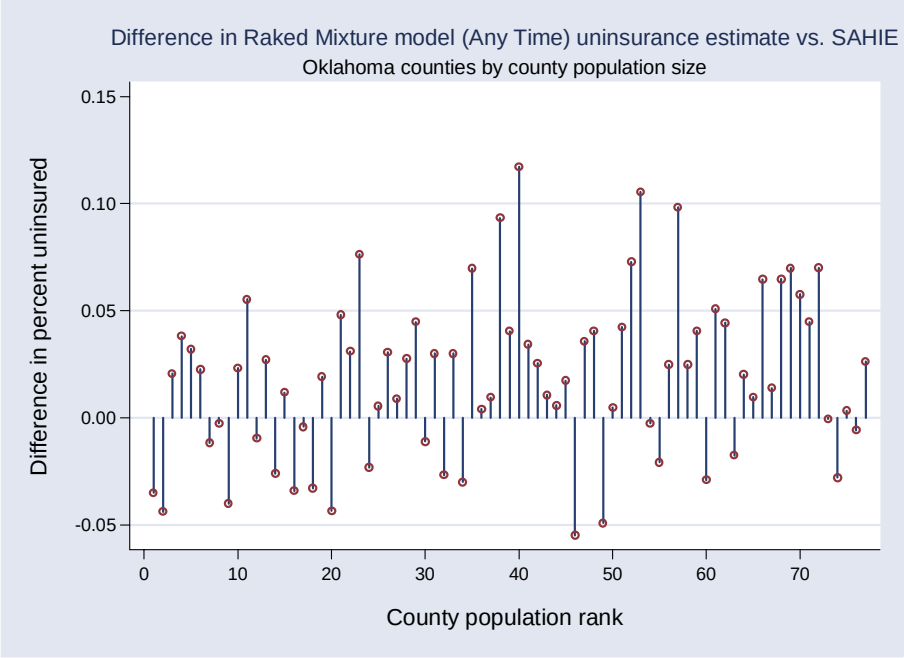
	Obs	Mean difference	Std. Dev.	Min	Max
All counties	77	-0.049	0.024	-0.107	0.012
CPS counties	24	-0.052	0.022	-0.100	0.012
Non-CPS counties	53	-0.042	0.026	-0.107	0.010

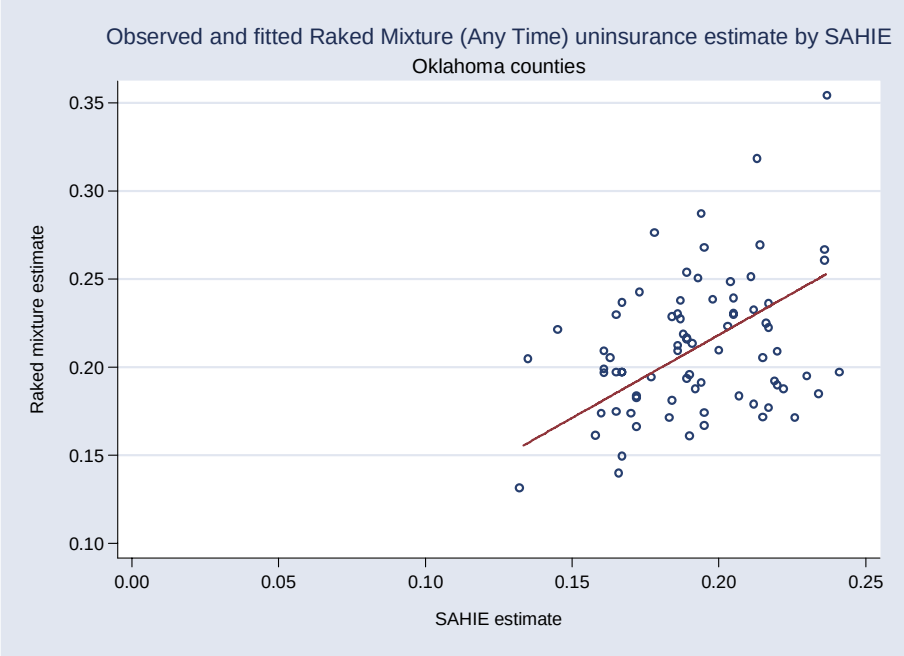




Differences in uninsurance estimates using raked mixture (Any Time) model compared with SAHIE estimate by county population rank, Oklahoma.

	Obs	Mean difference	Std. Dev.	Min	Max
All counties	77	0.019	0.032	-0.055	0.117
CPS counties	24	0.016	0.030	-0.049	0.105
Non-CPS counties	53	0.028	0.037	-0.055	0.117





Differences in uninsurance estimates using raked mixture (Point in Time) model compared with SAHIE estimate by county population rank, Oklahoma.

	Obs	Mean difference	Std. Dev.	Min	Max
All counties	77	-0.009	0.028	-0.070	0.058
CPS counties	24	-0.012	0.027	-0.067	0.046
Non-CPS counties	53	-0.001	0.030	-0.070	0.058