# Variance Modeling Research for the Small Area Health Insurance Estimates Program

Mark Bauder, U.S. Census Bureau
Sam Szelepka, U.S. Census Bureau
Donald Luery, U.S. Census Bureau

Federal Committee on Statistical Methodology
Research Conference
Washington, DC
November 4, 2013

United States™ Census Bureau

# Outline

- SAHIE Background
- Variance Modeling
- Research Results
- Summary and Future Research

United States™ Census Bureau

# SAHIE Background

United States™ Census Bureau

# SAHIE

Small Area Health Insurance Estimates Program (SAHIE)

- Model-based estimates of numbers and proportions with and without health insurance for
  - states by age, race, sex and income groups
  - counties by age, sex and income groups
- August 2013: released estimates for 2011.
- Planned: release estimates for 2012 in Spring, 2014.

United States™ Census Bureau

# SAHIE domains

- Modeling is done at a base level of non-overlapping domains defined by the full cross-classification of state or county by
  - 5 age groups
  - 4 race/ethnicity groups (states only)
  - 2 sexes
  - 5 income groups, defined by the family income-to-poverty ratio (IPR)
- Total of 10,200 domains for states, 157,050 for counties.
- SAHIE publishes estimates for domains that are typically aggregates of the base-level domains.

United States™ Census Bureau

# SAHIE Model

United States™ Census Bureau

# SAHIE Model (1)

Estimate two proportions

- $p_{ai}^{\text{IPR}}$ = the proportion in income group $i$, among those in geographic by demographic group $a$
- $p_{ai}^{\text{IC}}$ = the proportion insured among those in geographic by demographic by income group $a, i$

The number insured, $N_{ai}^{\text{IC}}$ is given by

$$N_{ai}^{\text{IC}} = \text{POP}_a \cdot p_{ai}^{\text{IPR}} \cdot p_{ai}^{\text{IC}}$$

where $\text{POP}_a$ is a population estimate, treated as known.

United States™ Census Bureau

# Data Sources

Data that are modeled:

- American Community Survey (ACS) unpublished current year estimates
- ACS unpublished 5-year estimates for the 5 years prior to the current year
- Administrative data
  - Aggregate totals of IRS exemptions
  - Supplemental Nutrition Assistance Program participation
  - Medicaid participation
  - Children's Health Insurance Program participation

United States™ Census Bureau

# SAHIE Model (2)

Hierarchical model:

- Transformations of the $p_{ai}^{\text{IPR}}$ and $p_{ai}^{\text{IC}}$, conditional on coefficients and variances, follow normal linear models.
- ACS current-year estimates, ACS previous 5-year estimates, and administrative data are modeled, conditional on the $p_{ai}^{\text{IPR}}$ and $p_{ai}^{\text{IC}}$.

Model is fully Bayesian, estimated using Markov Chain Monte Carlo.

United States™ Census Bureau

# Model for the ACS estimate, $\hat{p}^{\text{IPR}}$

Let $\hat{p}_{ai}^{\text{IPR}}$ be the current year ACS estimate of $p_{ai}^{\text{IPR}}$ for $i = 1, \ldots, 5$. Then, conditional on all the $p_{ai}^{\text{IPR}}$ and parameters

- $(\hat{p}_{a1}^{\text{IPR}}, \ldots, \hat{p}_{a4}^{\text{IPR}})$ is multivariate normal.
- $\text{E}(\hat{p}_{ai}^{\text{IPR}}) = p_{ai}^{\text{IPR}}$.
- $\text{var}(\hat{p}_{ai}^{\text{IPR}}) = \lambda_0 p_{ai}^{\text{IPR}}(1 - p_{ai}^{\text{IPR}})/S_a^{\lambda_1}$.
- The correlation between $\hat{p}_{ai}^{\text{IPR}}$ and $\hat{p}_{aj}^{\text{IPR}}$ has the same form as a multinomial distribution.

$S_a$ is sample size and $\lambda_0$ and $\lambda_1$ are parameters to be estimated that may differ by demographic group.

United States™ Census Bureau

# Model for the ACS estimate, $\hat{p}^{\text{IC}}$

Estimates of the proportion insured are frequently 1, and sometimes 0. We assume a mixture model for $\hat{p}^{\text{IC}}$. Conditional on $p_{ai}^{\text{IC}}$ and parameters

$$\hat{p}_{ai}^{\text{IC}} \begin{cases} = 0 & \text{with probability } p_{ai}^{(0)} \\ = 1 & \text{with probability } p_{ai}^{(1)} \\ \sim \text{Beta}(\alpha_{ai}, \beta_{ai}) & \text{with probability } 1 - p_{ai}^{(0)} - p_{ai}^{(1)} \end{cases}$$

with

$$p_{ai}^{(0)} = (1 - p_{ai}^{IC})^{1+\zeta_0(S_{ai}-1)} \qquad p_{ai}^{(1)} = (p_{ai}^{IC})^{1+\zeta_1(S_{ai}-1)}$$

$$\text{var}(\hat{p}_{ai}^{\text{IC}}) = \lambda_0 p_{ai}^{\text{IC}}(1 - p_{ai}^{\text{IC}})/S_a^{\lambda_1} \ .$$

$\alpha_{ai}$ and $\beta_{ai}$ are functions of $p_{ai}^{\text{IC}}, p_{ai}^{(0)}, p_{ai}^{(1)}$ and $\text{var}(\hat{p}_{ai}^{\text{IC}})$.

United States™ Census Bureau

# Variance Modeling

United States™ Census Bureau

# Approaches to Modeling the Survey Variances

SAHIE domains are often too small for direct estimates of the survey variances to be reliable.

- Current approach: for $p$, either $p^{IC}$ or $p^{IPR}$, assume

$$\text{var}(\hat{p}) = \lambda_0 p(1-p)/S^{\lambda_1}$$

  where $\lambda_0$ and $\lambda_1$ may differ for $p^{IC}$ and $p^{IPR}$ and may differ by demographic group. Estimate the $\lambda$'s jointly with all of the other unknowns in the model.

- Possible alternative: plug in smoothed estimates of $\text{var}(\hat{p})$ based on generalized variance functions (GVFs) fitted to direct variance estimates.

United States™ Census Bureau

# Research questions

Questions investigated in this research:

- What do direct estimates of survey variances suggest about the dependence of survey variances on sample size? Note that parameter estimates from SAHIE modeling often suggest that the survey variances decrease with sample size at a rate less than the inverse of the sample size.
- What do they suggest about the dependence of survey variances on the quantity $p(1-p)$?
- What do they suggest about the dependence on the proportion of survey cases that were collected through Computer Assisted Personal Interviewing (CAPI) as opposed to mail or telephone? Note that CAPI cases are a sample from cases that did not respond through mail or telephone.

United States™ Census Bureau

# Design Effects

Define a *design effect* as the variance of a survey estimator divided by its variance under simple random sampling.

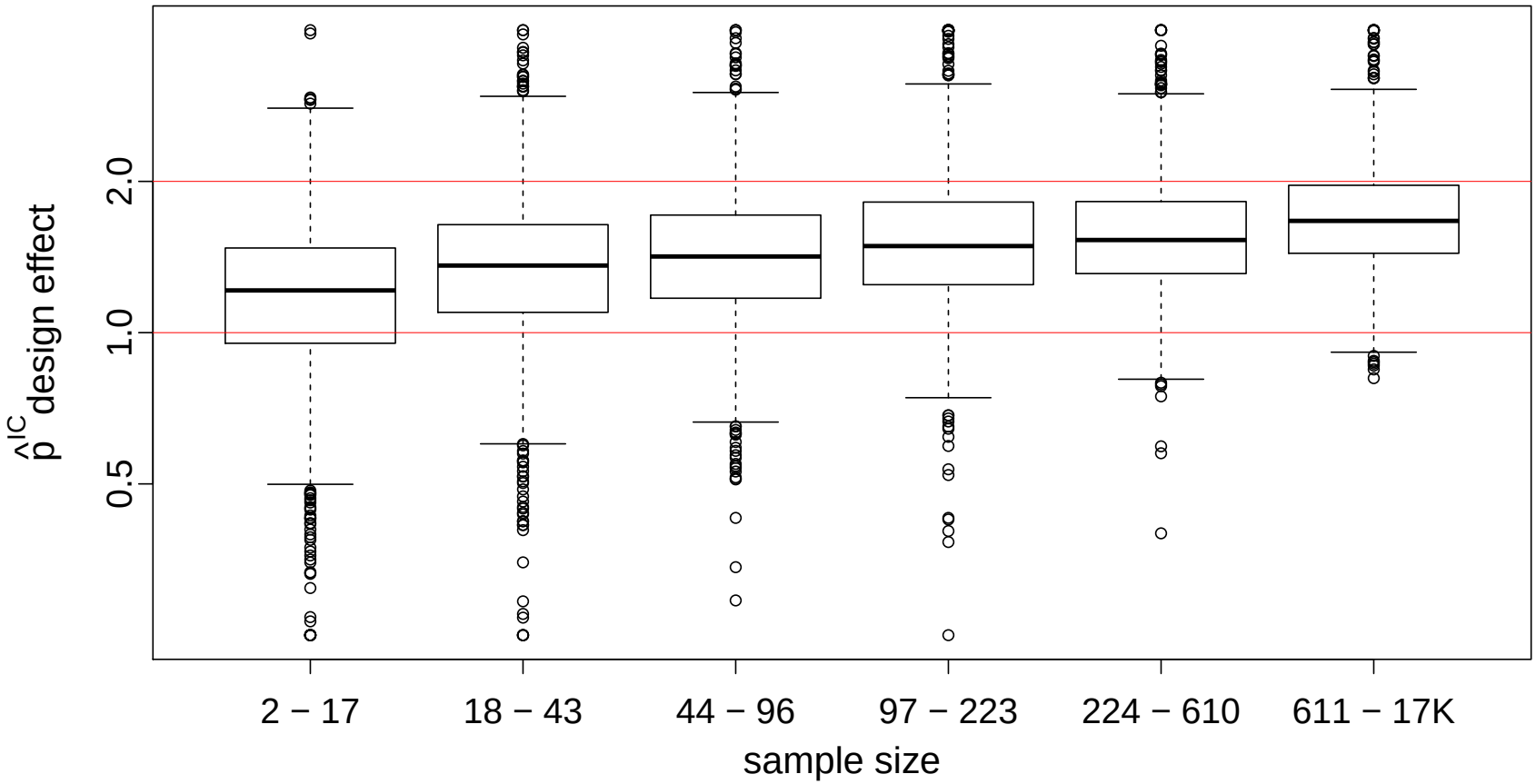
$$\text{deff} = \text{var}(\hat{p}) \cdot \frac{S}{p(1-p)}$$

where $S$ is sample size. For analyses and model fitting, we use an estimate of the design effect:

$$\widehat{\text{deff}} = \widehat{\text{var}}(\hat{p}) \cdot \frac{S}{\hat{p}(1-\hat{p})} \ .$$

United States™ Census Bureau

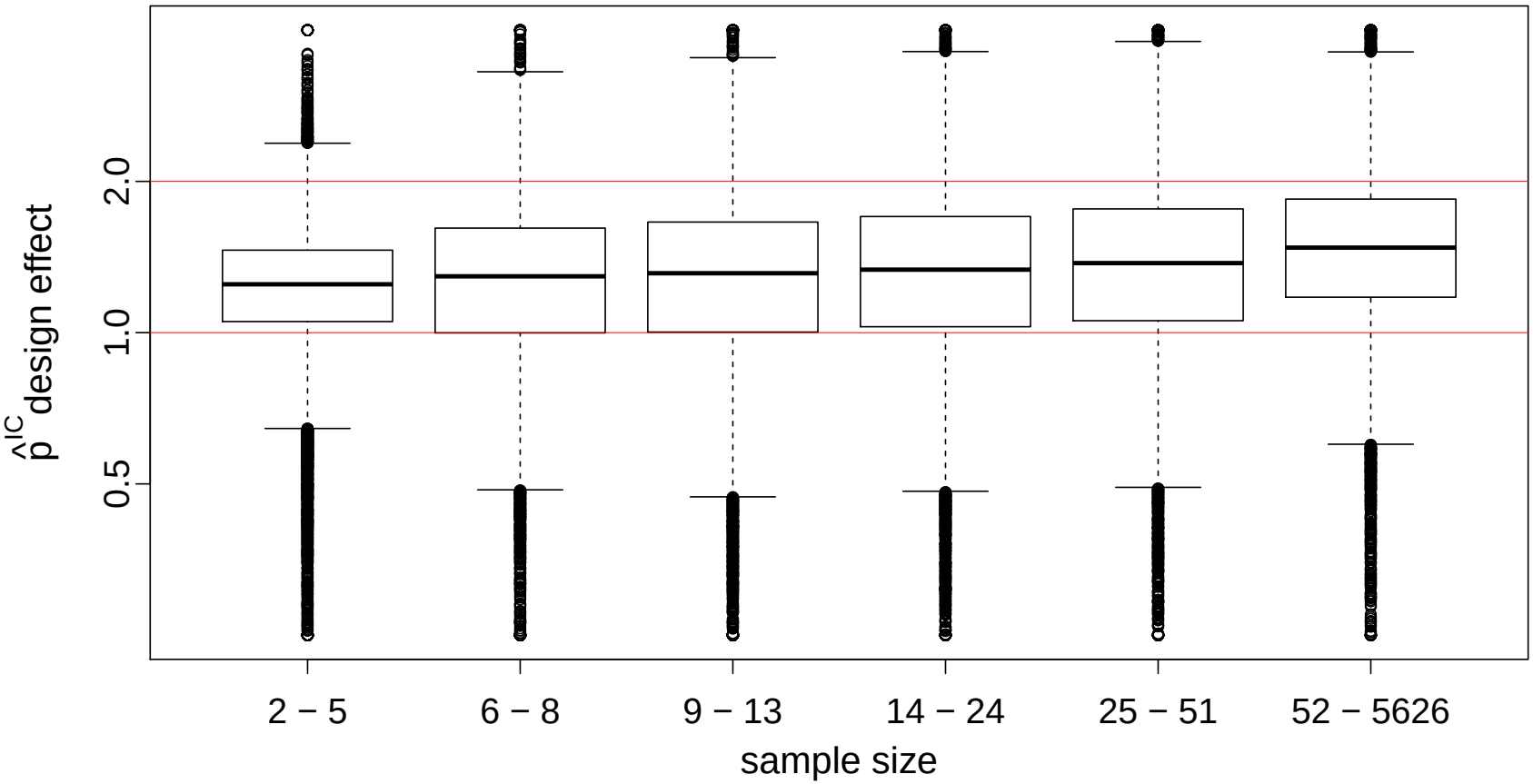
# $\hat{p}^{IC}$ Design Effect vs. Sample Size: States




$\hat{p}^{IC}$ estimated design effects vs. sextile of sample size. States. Sample size is the unweighted count in a state/age/race/sex/IPR group.

United States™ Census Bureau

# $\hat{p}^{IC}$ Design Effect vs. Sample Size: Counties




$\hat{p}^{IC}$ estimated design effects vs. sextile of sample size. Counties. Sample size is the unweighted count in a county/age/sex/IPR group.

United States™ Census Bureau

# Dependence on Sample Size

The previous plots suggest that the design effects increase with sample size, i.e. that the variances decrease with sample size at a rate slower than the inverse of the sample size.

United States™ Census Bureau

# Comparison to SAHIE Parameter Estimates

Fit a GVF model and compare parameter estimates to those from SAHIE for the variance function

$$\text{var}(\hat{p}) = \lambda_0 \, \hat{p} \, (1 - \hat{p}) \, \frac{1}{S^{\lambda_1}}$$

We estimate parameters in the model

$$\log(\widehat{\text{deff}}) = \alpha + \beta \log(S) + \epsilon$$

then transform to obtain GVF-based estimates of $\lambda_0$ and $\lambda_1$.

United States™ Census Bureau

# Model fitting methodology

- Separate models for different demographic groups, corresponding to SAHIE.
- Use weighted regression with weights equal to the inverse of the minimum of 1000 and the sample size.
- For models of the variance of $\hat{p}^{\text{IC}}$, exclude cases where the sample size is less than or equal to 50, or where the unweighted count uninsured is less than 5.
- For models of the variance of $\hat{p}^{\text{IPR}}$, exclude cases where the sample size is less than or equal to 50, or where the unweighted count in the IPR group is less than 5.

United States™ Census Bureau

# $\hat{p}^{IC}$ Variance Parameters: States

| par. | age | IPR | SAHIE mean | SAHIE st. dev. | GVF est. | GVF st. err. |
|---|---|---|---|---|---|---|
| $\lambda_0$ | 0-18 | 0-200 | 1.20 | 0.06 | 1.01 | 0.06 |
| | 0-18 | 200-400 | 1.35 | 0.06 | 1.07 | 0.07 |
| | 0-18 | 400+ | 1.14 | 0.08 | 0.93 | 0.11 |
| | 19-64 | 0-200 | 1.19 | 0.07 | 1.32 | 0.04 |
| | 19-64 | 200-400 | 1.21 | 0.06 | 1.21 | 0.04 |
| | 19-64 | 400+ | 1.36 | 0.16 | 1.14 | 0.05 |
| $\lambda_1$ | 0-18 | 0-200 | 0.87 | 0.02 | 0.89 | 0.01 |
| | 0-18 | 200-400 | 0.92 | 0.02 | 0.91 | 0.01 |
| | 0-18 | 400+ | 0.91 | 0.02 | 0.91 | 0.02 |
| | 19-64 | 0-200 | 0.91 | 0.02 | 0.98 | 0.01 |
| | 19-64 | 200-400 | 0.93 | 0.02 | 0.96 | 0.01 |
| | 19-64 | 400+ | 0.96 | 0.03 | 0.95 | 0.01 |

Note: $\text{var}(\hat{p}^{IC}) = \lambda_0 p^{IC}(1 - p^{IC})/S^{\lambda_1}$.

United States™ Census Bureau

# $\hat{p}^{IC}$ Variance Parameters: Counties

| par. | age | IPR | SAHIE mean | SAHIE st. dev. | GVF est. | GVF st. err. |
|---|---|---|---|---|---|---|
| $\lambda_0$ | 0-18 | 0-200 | 1.26 | 0.01 | 1.90 | 0.05 |
| | 0-18 | 200-400 | 1.27 | 0.01 | 1.55 | 0.05 |
| | 0-18 | 400+ | 1.33 | 0.02 | 2.08 | 0.06 |
| | 19-64 | 0-200 | 1.04 | 0.01 | 1.39 | 0.05 |
| | 19-64 | 200-400 | 1.08 | 0.01 | 1.25 | 0.05 |
| | 19-64 | 400+ | 1.17 | 0.02 | 1.67 | 0.05 |
| $\lambda_1$ | | 0-200 | 0.82 | 0.00 | 0.99 | 0.01 |
| | | 200-400 | 0.86 | 0.00 | 0.97 | 0.01 |
| | | 400+ | 0.90 | 0.01 | 1.02 | 0.01 |

Note: $\text{var}(\hat{p}^{IC}) = \lambda_0 p^{IC}(1 - p^{IC})/S^{\lambda_1}$.

United States™ Census Bureau

# $\hat{p}^{\text{IPR}}$ Variance Parameters: States

| | | SAHIE | | GVF | |
|---|---|---|---|---|---|
| par. | age | mean | st. dev. | est. | st. err. |
| $\lambda_0$ | 0-17 | 4.98 | 1.00 | 1.40 | 0.04 |
| | 18 | 1.63 | 0.18 | 1.08 | 0.05 |
| | 19-39 | 3.17 | 0.71 | 1.36 | 0.04 |
| | 40-49 | 1.64 | 0.26 | 1.34 | 0.04 |
| | 50-64 | 2.01 | 0.33 | 1.37 | 0.04 |
| $\lambda_1$ | 0-17 | 1.09 | 0.03 | 0.95 | 0.01 |
| | 18 | 1.02 | 0.03 | 0.95 | 0.01 |
| | 19-39 | 1.10 | 0.04 | 0.98 | 0.01 |
| | 40-49 | 1.00 | 0.03 | 0.99 | 0.01 |
| | 50-64 | 1.06 | 0.03 | 1.00 | 0.01 |

Note: $\text{var}(\hat{p}^{\text{IPR}}) = \lambda_0 p^{\text{IPR}}(1 - p^{\text{IPR}})/S^{\lambda_1}$.

United States™ Census Bureau

# $\hat{p}^{\text{IPR}}$ Variance Parameters: Counties

| | | SAHIE | | GVF | |
|---|---|---|---|---|---|
| par. | age | mean | st. dev. | est. | st. err. |
| $\lambda_0$ | 0-17 | 1.96 | 0.06 | 1.63 | 0.02 |
| | 18 | 1.26 | 0.02 | 1.11 | 0.02 |
| | 19-39 | 1.38 | 0.03 | 1.43 | 0.01 |
| | 40-49 | 1.24 | 0.02 | 1.37 | 0.01 |
| | 50-64 | 1.21 | 0.03 | 1.35 | 0.01 |
| $\lambda_1$ | 0-18 | 0.95 | 0.01 | 0.95 | 0.01 |
| | 19-64 | 0.94 | 0.01 | 0.99 | 0.00 |

Note: $\text{var}(\hat{p}^{\text{IPR}}) = \lambda_0 p^{\text{IPR}}(1 - p^{\text{IPR}})/S^{\lambda_1}$.

United States™ Census Bureau

# Summary of parameter comparison

Dependence of survey variances on sample size:

- Variance of $\hat{p}^{IC}$:
  - For states, the estimates from fitting the GVF are generally close to those from SAHIE, and agree in being significantly less than 1.
  - For counties, the estimates from fitting the GVF are not as close to those from SAHIE. However, over 90% of the county observations are excluded from the GVF modeling, because of small sample sizes or small unweighted counts uninsured.
- Variance of $\hat{p}^{IPR}$:
  - For states, the SAHIE estimates are not as nicely behaved. They have large standard deviations and have large differences by age.
  - For counties, estimates are closer.

United States™ Census Bureau

# Dependence on $p(1-p)$

To assess: is the survey variance proportional to $p(1-p)$?

Fit model

$$\log(\widehat{\text{deff}}) = \alpha + \beta \log(S) + \eta \log(p(1-p)) + \epsilon.$$

The implied variance is proportional to $[\, p\,(1-p)\,]^{1+\eta}$.

United States™ Census Bureau

# Dependence on $p(1-p)$

| type | geo. | age | $\eta$ est. | st. err. | Pr $> \lvert t \rvert$ |
|---|---|---|---|---|---|
| $\hat{p}^{\text{IC}}$ | state | 0-18 | 0.156 | 0.015 | $< .0001$ |
| | | 19-64 | -0.005 | 0.010 | 0.591 |
| | county | 0-18 | 0.169 | 0.021 | $< .0001$ |
| | | 19-64 | 0.017 | 0.010 | 0.079 |
| $\hat{p}^{\text{IPR}}$ | state | 0-18 | 0.070 | 0.015 | $< .0001$ |
| | | 19-64 | 0.031 | 0.011 | 0.007 |
| | county | 0-18 | 0.059 | 0.010 | $< .0001$ |
| | | 19-64 | 0.024 | 0.005 | $< .0001$ |

Note: model is $\quad \text{var}(\hat{p}) = \lambda_0 \left[ p(1-p) \right]^{1+\eta} / S^{\lambda_1}$.

United States™ Census Bureau

# CAPI

- The ACS uses three modes of data collection: mail, telephone, and Computer Assisted Personal Interviewing (CAPI).
- CAPI: a subsample of cases from which no mail questionaire been received and no telephone interview completed.
- Proportion of CAPI cases may affect survey variances.

Fit model

$$\log(\widehat{\text{deff}}) = \alpha + \beta \log(S) + \eta \log(p(1-p)) + \gamma PCAPI + \epsilon$$

where $PCAPI$ is the proportion of CAPI cases.

United States™ Census Bureau

# Effect of CAPI rate

| geography | age | $\gamma$ est. | st. err. | Pr $> \lvert t \rvert$ |
|---|---|---|---|---|
| state | 0-18 | -0.066 | 0.069 | 0.3409 |
| | 19-64 | -0.297 | 0.051 | $<$ .0001 |
| county | 0-18 | -0.533 | 0.097 | $<$ .0001 |
| | 19-64 | -0.276 | 0.075 | 0.0002 |

United States™ Census Bureau

# Summary

- Fitting GVF models largely confirms the SAHIE results that the survey variances decrease more slowly than with the inverse of sample size.
- There is evidence that the survey variances are not proportional to $p(1-p)$, but instead to $p(1-p)$ raised to a power other than 1.
- There is some evidence for the dependence of the survey variances on collection mode.

United States™ Census Bureau

# Possible Further Research

- Investigate properties of the direct variance estimates:
  - Are apparent dependencies of design effects on sample size and proportion artifacts of the estimation procedure?
  - What is a reasonable set of observations to use for variance modeling?
- How best to incorporate direct variance estimates into SAHIE modeling:
  - Plug in GVF estimates?
  - Use direct variance estimates to suggest other variance models to use within SAHIE modeling?

United States™ Census Bureau

# Disclaimer

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

United States™ Census Bureau

# Contact Information

Mark Bauder: donald.m.bauder@census.gov

Sam Szelepka: samuel.szelepka@census.gov

Donald Luery: donald.m.luery@census.gov

United States™ Census Bureau