Accounting for Group Classification Error in Variance Estimates Using the American Community Survey

Matthew W. Brault

For the American Community Survey Data Users Conference

May 29, 2014

United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily that of the U.S. Census Bureau.

Motivation

- Wanted to classify occupations as “high” or “low” wage jobs.
  - Great! ACS can do that!
  - Calculate median earnings for individual occupations. Set a criteria. Define occupations.
  - But how certain am I about that classification?
  - How do I account for that uncertainty in my final estimate?

Outline

- General case
  - Naïve method
  - Sophisticated method
- Example using areas of concentrated poverty
- Simulation

General Case

- Individuals are arranged in K groups
- For each group $K = k$, a statistic θ_k is calculated with standard error σ_k.
- θ_k is compared against some critical value τ and an indicator y_k is set to 1 or 0.

$$y_k \begin{cases} = 0, & if \theta_k \leq \tau \\ = 1, & if \theta_k > \tau \end{cases}$$

General Case
Naïve Variance

- Final estimate is $Y = \sum(y_k W_i) / \sum W_i$
- Naïve variance

$$\tilde{Y}_r = \sum(y_k W_{i,r}) / \sum W_{i,r}$$

$$\sigma_{\hat{y}}^2 = \widetilde{var}(Y) = \frac{\sum\left(\tilde{Y}_0 - \tilde{Y}_r\right)^2}{R(1-\varepsilon)^2}$$

General Case
Measurement Error

- The assignment of y_k is a measurement error problem

$$y = \hat{y} + \eta, \qquad \eta \sim N(0, \sigma_\eta)$$

$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_\eta^2 + 2\sigma_{\hat{y},\eta}$$

  - σ_η^2 is related to the variances of θ_k

General Case
Sophisticated Variance

$$y_{k,r}\begin{cases} = 0, & if\,\theta_{k,r} \leq \tau \\ = 1, & if\,\theta_{k,r} > \tau \end{cases}$$

$$Y_r = \sum \left(y_{k,r} W_{i,r}\right) / \sum W_{i,r}$$

$$\sigma_y^2 = var(Y) = \frac{\sum (Y_0 - Y_r)^2}{R(1-\varepsilon)^2}$$

General Case
Measurement Error (Cont.)

- *Variance attributed to indicator alone*

$$\delta_{k,r} = y_{k,0} - y_{k,r}$$

$$\Delta_r = \frac{\sum\left(\delta_{k,r} W_{i,r}\right)}{\sum W_{i,r}} = \tilde{Y}_r - Y_r$$

$$\sigma_\eta^2 = var(\Delta) = \frac{\sum(\Delta_0 - \Delta_r)^2}{R(1-\varepsilon)^2}$$

How they relate

$$\sigma_y^2 = var(Y) = \frac{\sum_R (Y_0 - Y_r)^2}{R(1-\varepsilon)^2} = \frac{\sum_R \left(Y_0 - \widetilde{Y}_r + \widetilde{Y}_r - Y_r\right)^2}{R(1-\varepsilon)^2}$$

$$= \frac{\sum_R \left[\left(Y_0 - \widetilde{Y}_r\right)^2 + \left(\widetilde{Y}_r - Y_r\right)^2 + 2\left(Y_0 - \widetilde{Y}_r\right)\left(\widetilde{Y}_r - Y_r\right)\right]}{R(1-\varepsilon)^2}$$

$$= \frac{\sum_R \left(Y_0 - \widetilde{Y}_r\right)^2}{R(1-\varepsilon)^2} + \frac{\sum_R \left(\widetilde{Y}_r - Y_r\right)^2}{R(1-\varepsilon)^2} + 2\frac{\sum_R \left(Y_0 - \widetilde{Y}_r\right)\left(\widetilde{Y}_r - Y_r\right)}{R(1-\varepsilon)^2}$$

$$= \sigma_{\hat{y}}^2 + \sigma_\eta^2 + 2\sigma_{\hat{y},\eta}$$

Applications

- Industries as generous providers of health insurance
- Foreign born groups (by country of birth) as “new/emerging” immigrants
- Neighborhoods as impoverished
  - Bishaw, 2011
- Etc…

Poverty Areas Example

- *Areas With Concentrated Poverty: 2006-2010*
  - ACS Brief that examines census tracts by poverty rate:

Category I (0-13.7%)

Category II (13.8%-19.9%)

Category III (20.0%-39.9%)

Category IV (40.0%-100.0%)

Areas With Concentrated Poverty: 2006–2010

American Community Survey Briefs

Issued December 2011

ACSBR/10-17

By
Alemayehu Bishaw

People living in poverty tend to be clustered in certain neighborhoods rather than being evenly distributed across geographic areas. Measuring this concentration of poverty is important because researchers have found that living in areas with many other poor people places burdens on low-income families beyond what the families' own individual circumstances would dictate. Many argue that this concentration of poverty results in higher crime rates, underperforming public schools, poor housing and health conditions, as well as limited access to private services and job opportunities.[1] In recognition of these burdens, some government programs target resources to communities with concentrated poverty. Many of these programs use the Census Bureau's definition of "poverty areas" (census tracts with poverty rates of 20 percent or more).[2]

Using the 5-year poverty rate estimates from the American Community Survey (ACS), this report analyzes demographic and socioeconomic characteristics of census tracts by categorizing the tracts into four categories based on their poverty rate levels.[3] Category I includes census tracts with poverty rates less than 13.8 percent.[4] Category II includes those with poverty rates of 13.8 percent to 19.9 percent. Category III includes those with poverty rates of 20.0 percent to 39.9 percent, and Category IV includes those tracts with poverty rates of 40.0 percent or more. Like previous census publications, in this report census tracts with poverty rates of 20 percent or more (tracts in category III and category IV) are referred to as "poverty areas."

This report is an update of the special reports on areas with concentrated poverty, prepared to show the distributions and characteristics of the U.S. population by the levels of poverty of the census tracts in which they live. The previous two reports used data collected using the long forms of the 1990 and 2000 decennial censuses.

Table 1 shows the distribution of people in census tracts by poverty levels for the nation, regions, and states and the District of Columbia. Figure 1 displays the proportion of people living in poverty areas (tracts with poverty rates of 20 percent or higher) by state.

[1] See for example, *The Enduring Challenge of Concentrated Poverty in America: Case Studies from Communities across the U.S.A.*, A Joint Project of the Community Affairs Offices of the Federal Reserve System and the Metropolitan Policy Program at the Brookings Institution. (Federal Reserve Bank of Cleveland, 2008).

[2] For example, the Developmental Disabilities Assistance and Bill of Rights Act in Section 103(a) allows Basic State Grant projects whose activities or products target poverty areas to receive as much as 90 percent in federal support and the New Market Tax Credit (NMTC) program, enacted in December 2000 as part of the Community Renewal Tax Relief Act, defines eligibility as projects in census tracts with a poverty rate of at least 20 percent.

[3] The ACS collects and releases data by calendar year for geographic areas that meet specific population thresholds. One-year estimates are published for areas with populations of 65,000 or more, 3-year estimates for populations of 20,000 or more, and 5-year estimates for populations of almost any size. ACS 1-, 3-, and 5-year estimates are period estimates, which mean they represent the characteristics of the population and housing over a specific data collection period. Data are combined to produce 12 months, 36 months, or 60 months of data.

[4] According to the 5-year ACS, the poverty rate for all people in the United States was 13.8 percent.

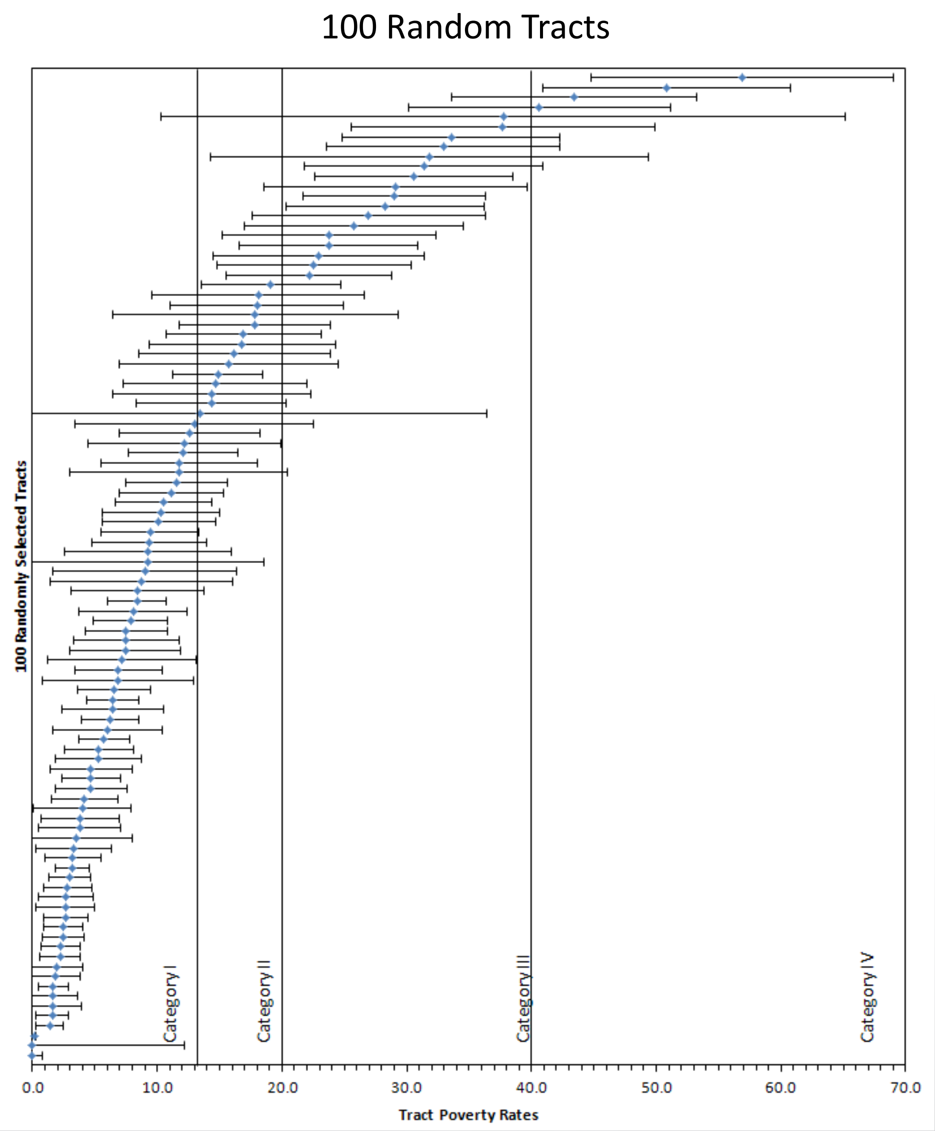
United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

Methods

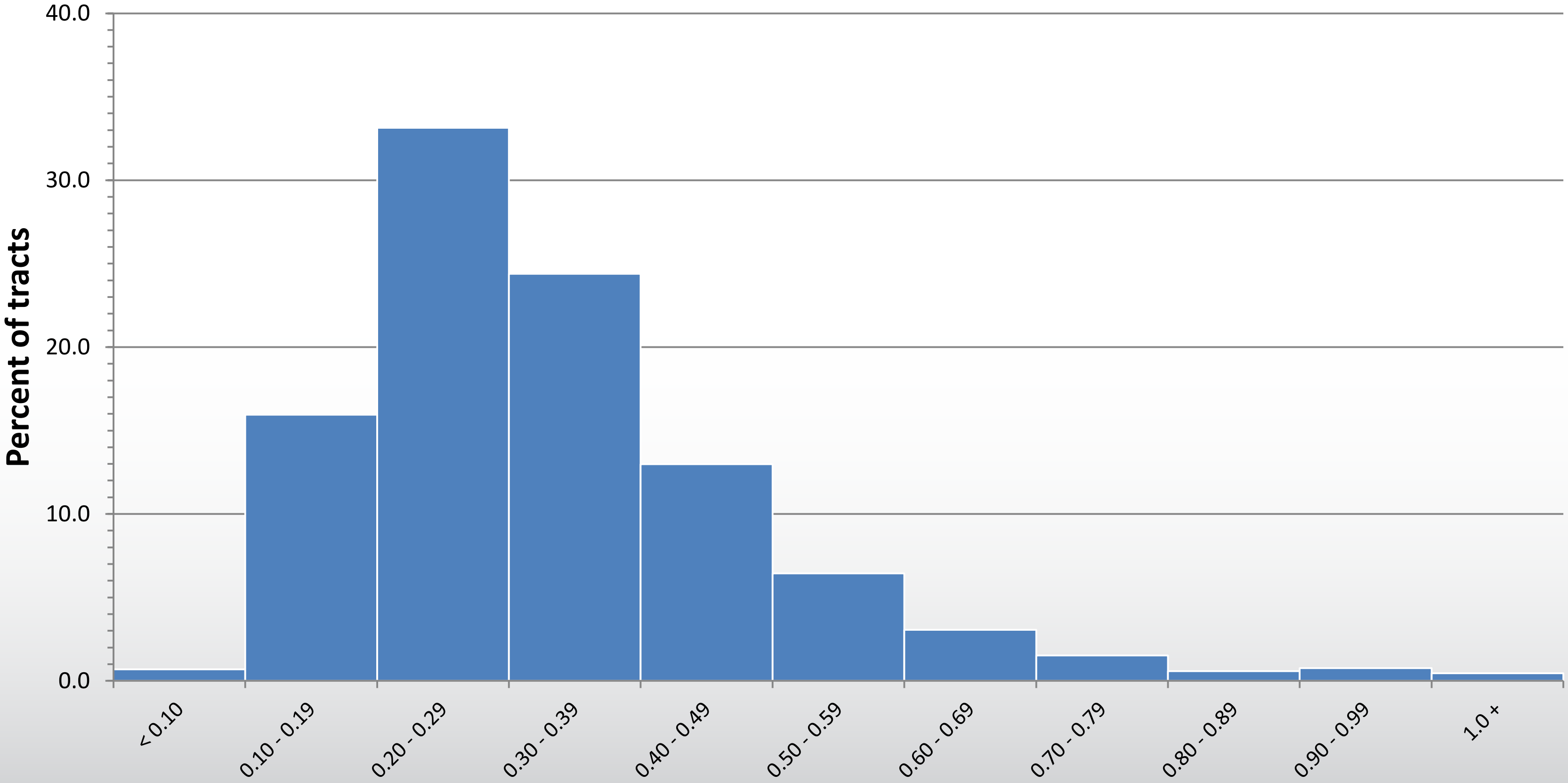
- 72,254 Census tracts in U.S.
- Used 2006-2010 ACS 5-year data to calculate poverty rates for tracts
    - Standard errors calculated using replicate weights.
    - 517 tracts had rates of 0 percent and 18 had rates of 100 percent
        - Standard errors calculated using ACS Production method (based on tract size and average weight in the state)
        - Replicate poverty rates simulated from SE

Tract Poverty Rates



Tract Group	Number of Tracts	Percentage of Population
Category I	42,383	61.4
Category II	11,574	16.0
Category III	14,823	19.1
Category IV	3,474	3.5

Tract CVs

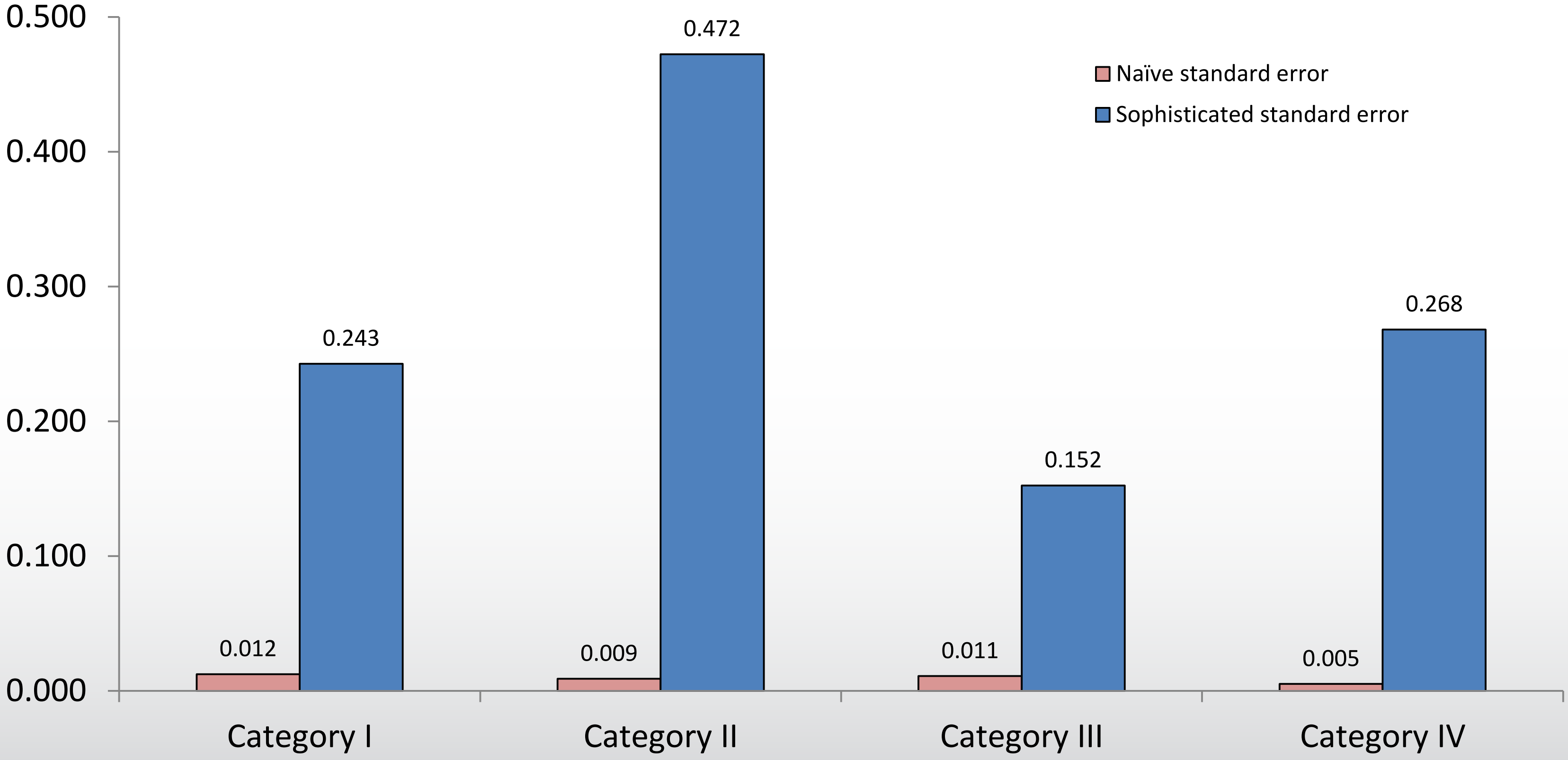


40.0
30.0
20.0
10.0
0.0
Percent of tracts
< 0.10
0.10 - 0.19
0.20 - 0.29
0.30 - 0.39
0.40 - 0.49
0.50 - 0.59
0.60 - 0.69
0.70 - 0.79
0.80 - 0.89
0.90 - 0.99
1.0 +

Results

	Category I	Category II	Category III	Category IV
Naïve variance ($\widetilde{var}(Y)$)	0.000153	0.000079	0.000118	0.000025
Naïve standard error	0.012355	0.008887	0.010850	0.005017
Sophisticated variance ($var(Y)$)	0.058874	0.223126	0.023230	0.071863
Sophisticated standard error	0.242639	0.472362	0.152413	0.268073
Measurement error variance ($var(\Delta)$)	0.062034	0.225113	0.024111	0.071109
Covariance ($cov(Y,\Delta)$)	-0.001656	-0.001033	-0.000499	0.000364
Ratio of standard errors	19.64	53.15	14.05	53.43

Size of Standard errors



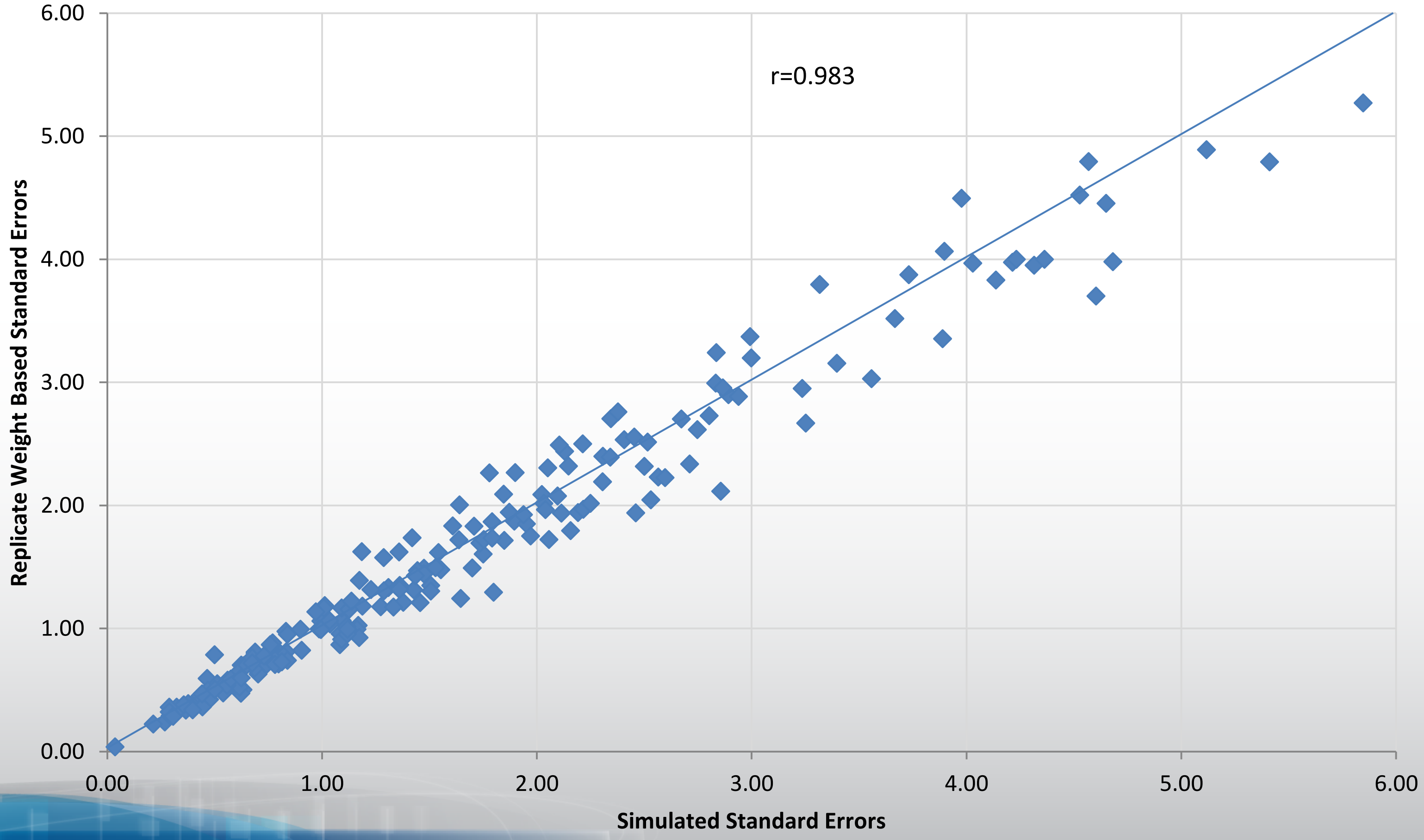
State Estimates

	Category I	Category II	Category III	Category IV
Median CV (Naïve)	0.001	0.004	0.004	0.014
Median CV (sophisticated)	0.029	0.118	0.085	0.190
Smallest Ratio of Standard Errors	12.2	18.7	12.1	7.1
Largest Ratio	37.5	55.0	34.6	30.2
Median Ratio	19.9	29.4	17.4	13.7

Simulation

- Attaching to other datasets
  - Different number of replicates
- Can’t get replicate estimates from public use data
- Use FactFinder Estimates/Standard Errors
- Simulate the Replicate Distribution
  - Normal distribution $\sim N\left(\theta_k, \gamma\sigma_k^2\right),\ \gamma = \frac{R(1-\varepsilon)^2}{(R-1)}$

Simulated and Replicate Based Standard Errors - States



Conclusion

- Error can be quite large!!
- Provide greater utility to working with estimates for small domains as an aggregate
- Properly reflect the level of uncertainty associated with estimates

SAS code available in an appendix to the paper

Thank You!

Matthew W. Brault

Health and Disability Statistics Branch

Social Economic and Housing Statistics Division

matthew.w.brault@census.gov

301-763-9112