A Simulation-Based Approach to Refining Estimates of Sampling Variability for the Planning Database's Low Response Score

Luke J. Larsen

U.S. Census Bureau

July 29 – August 3, 2018

2018 JSM Conference

Vancouver, BC

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

Presentation Agenda

- Introduction and context (PDB, LRS, and research questions)
- Method (simulation-based variance estimation)
- Data source and sample design
- Analysis
- Conclusion

What is the Planning Database?

- Publicly available collection of popular measures
  - Ex: # of HUs, % Pop under 5 yrs, Median Hhld Income, Pop Density
- Data comes from Census 2010 and ACS 5-year Summary Files
- Aggregated counts and percents at **tract** & **block group** levels.
- Many uses – primary function is to aid in planning field operations for Census 2020 and other survey projects
- https://www.census.gov/research/data/planning_database/

What is the Low Response Score?

- Metric created for PDB as predictor of self-response propensity
- Derived from multivariate linear regression (MLR) model with Census 2010 mail non-response rate as dependent variable
- Ranges from 0 to 100 (low LRS = higher predicted self-response rate); Example: when LRS = 25, we predict that 25% of households in that tract will not self-respond to the Census.
- Based on 25 main-effect inputs from ACS 5-year Summary Files
- Methodology: see Erdman and Bates (2017)

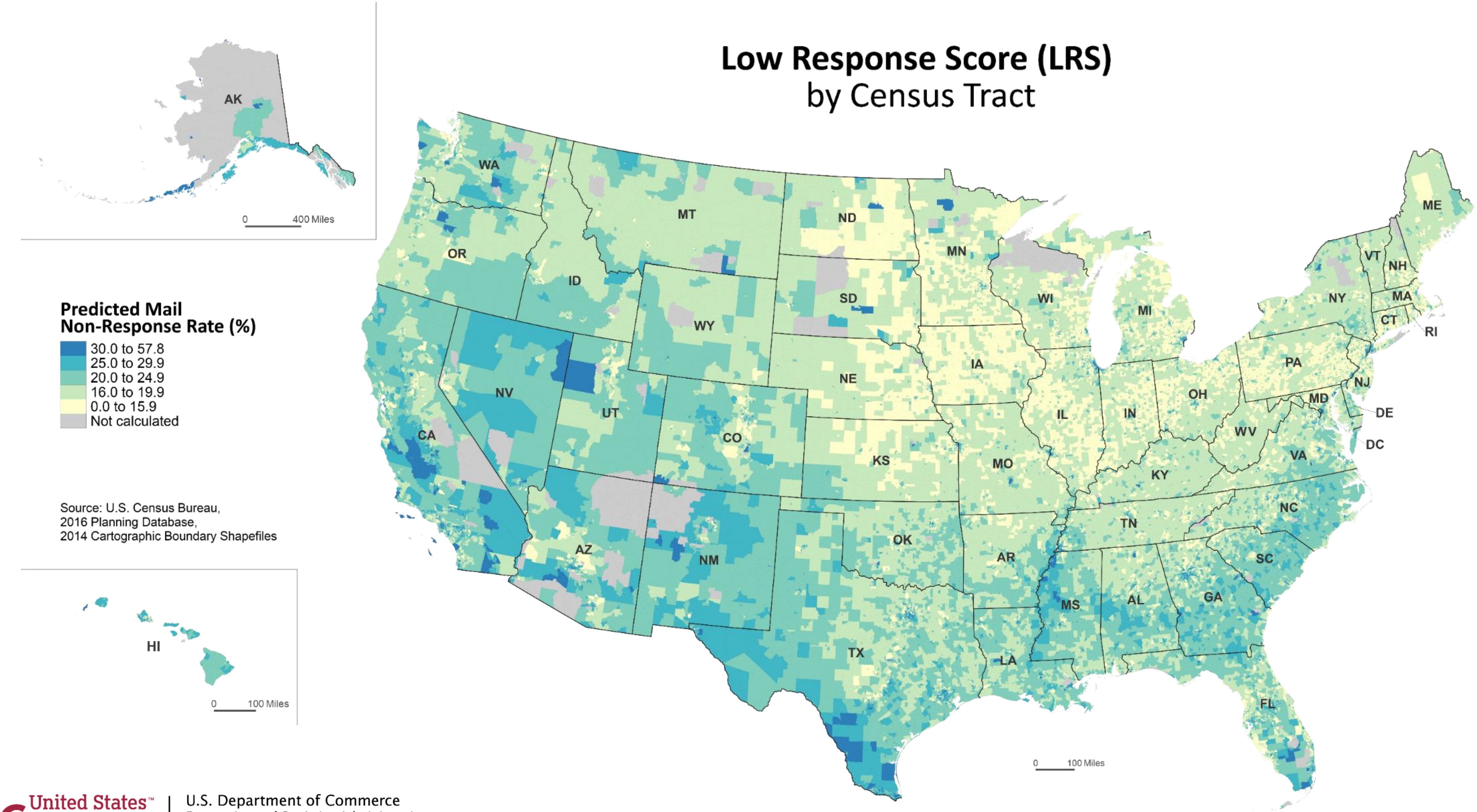
Low Response Linear Regression Model (Block Group)

	Coef	Sig		Coef	Sig
(Intercept)	10.29	***	Renter occupied units	1.08	***
Ages 18-24	0.64	***	Female head, no husband	0.58	***
Non-Hispanic White	-0.77	***	Ages 65+	-1.21	***
Related child <6	0.46	***	Males	0.09	***
Married family households	-0.12	***	Ages 25-44	-0.06	
Vacant units	1.08	***	College graduates	-0.32	***
Median household income	0.24	***	Ages 45-64	-0.08	*
Persons per household	3.44	***	Moved in 2005-2009	0.09	***
Hispanic Any Race	0.41	***	Single unit structures	-0.52	***
Population Density	-0.40	***	Below poverty	0.11	***
Different HU 1 year ago	-0.12	***	Ages 5-17	0.17	***
Non- Hispanic Black	-0.04	**	Single person households	-0.24	***
Not high school grad	-0.06	***	Median house value	0.71	***

Sig: * * * $p < .001$; * * $.001 \leq p < .01$; * $.01 \leq p < .05$ R-squared: 56.10%, $n =$ 217,417
Main Effects only, no interaction terms

Source: Erdman and Bates, 2017.

Low Response Score (LRS)
by Census Tract



Source: U.S. Census Bureau,
2016 Planning Database,
2014 Cartographic Boundary Shapefiles

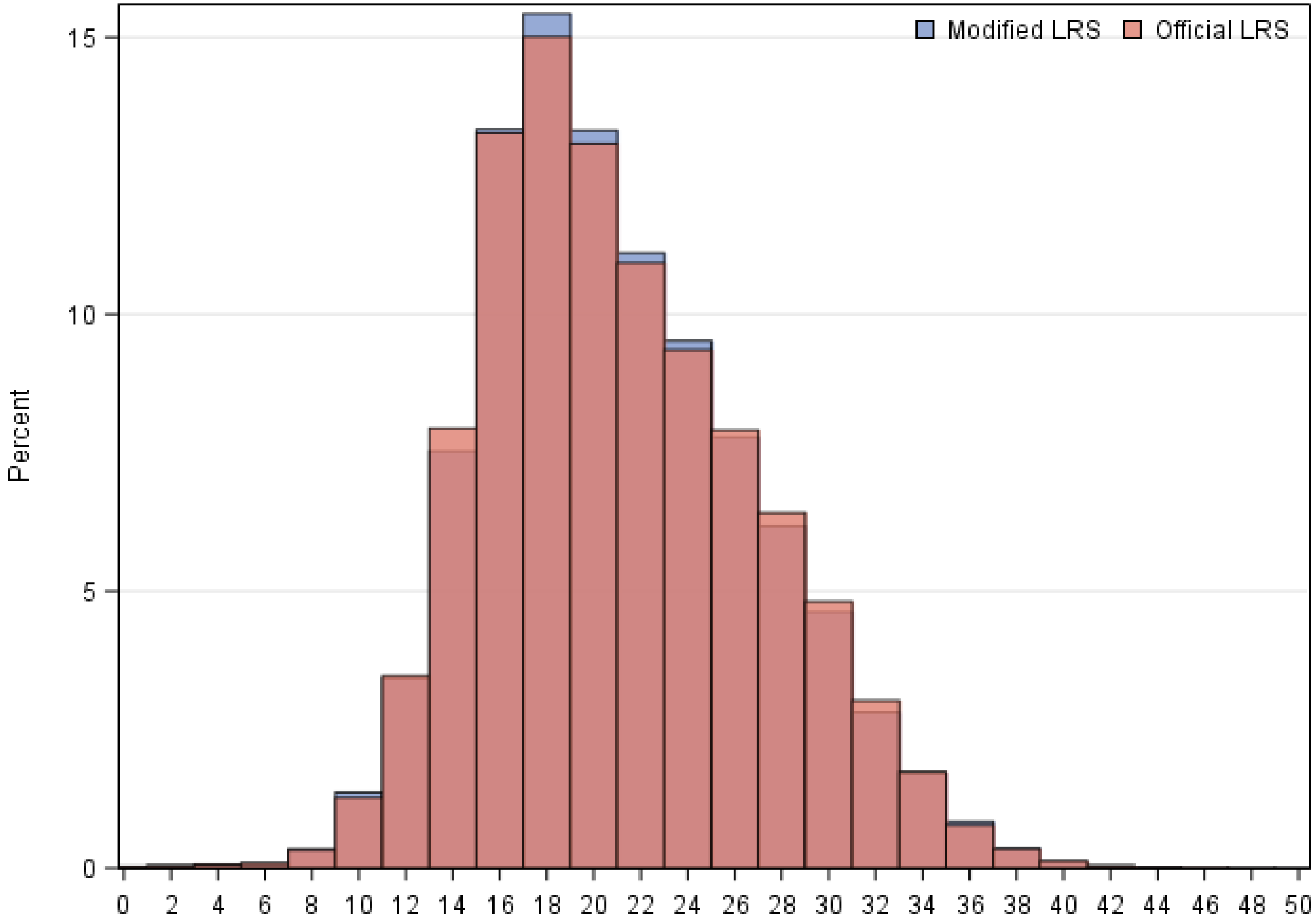
United States Census Bureau

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

Why should we care about LRS variability?

- Need to be able to discern significant differences between LRS predictions for field planning purposes.
- Ex: Tract A has LRS = 15, Tract B has LRS = 22. Are these significantly different?

Predicted Distributions of Modified and Official LRS



Source: Larsen, 2017.

Statement of Purpose

- Ongoing research into variability of the Low Response Score.
- Last time (Larsen, 2017), I used ACS replicate weights to generate approximate MOEs for LRS predictions at tract level.
  - Did not account for sampling variability in regressor inputs.
  - Currently do not have method that addresses sampling error from both the coefficient estimates and the regressor inputs.
- Can we use simulation techniques to determine whether the MOEs would significantly change under a “full” strategy?

Research Question

Consider two strategies for estimating the variance of LRS predictions using a Monte Carlo simulation approach:

- ***"Partial":*** LRS predictions are simulated by allowing only the coefficients to vary while fixing the inputs in place.
- ***"Full":*** LRS predictions are simulated by allowing both the coefficients and the inputs to vary.

RQ: Are the Full variance estimates significantly different than the Partial variance estimates for individual tracts?

Method: Monte Carlo variance estimate

1. Obtain the tract-level LRS model coefficients (Erdman and Bates, 2017).
2. For a given tract in the current PDB, generate 50 simulations of the LRS (either Full or Partial strategy)
3. Calculate the sample variance of the 50 simulations.
4. Repeat steps 2 and 3 over a large number (4000) of iterations.
5. The mean of these simulated variances is the Monte Carlo variance estimate.
6. Predicted LRS variance = MC variance of fitted LRS + MSE of model fit (27.8)
7. Repeat steps 2 through 6 for all tracts in the sample (n=1000)

LRS simulation example (1)

Miami-Dade County, Florida
Tract 0083.05, 1st iteration

X = 50 LRS simulations

Mean LRS_{Part} = 23.34
Variance LRS_{Part} = 7.11

Mean LRS_{Full} = 22.75
Variance LRS_{Full} = 4.44

Miami-Dade County, Florida
Tract 0083.05, all iterations

N = 4000 iterations

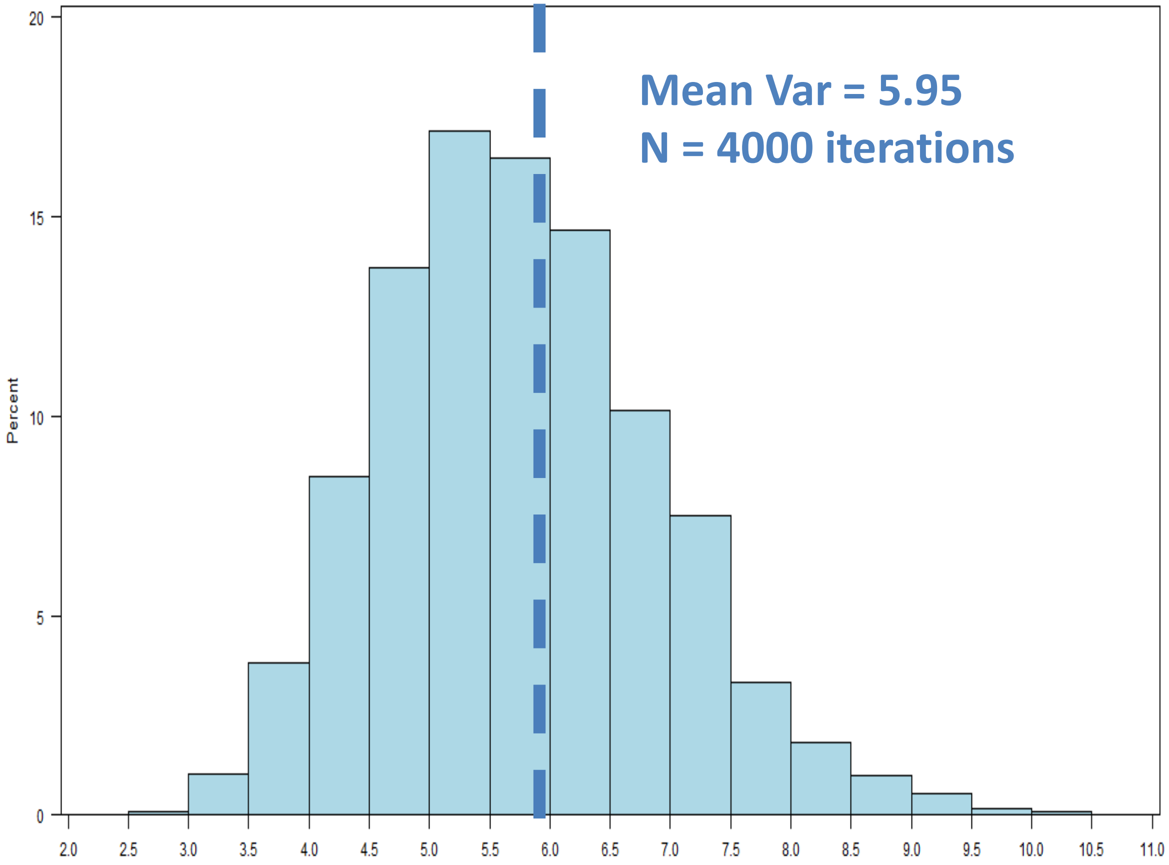
Mean Var(LRS_{Part}) = 5.95
Variance Var(LRS_{Part}) = 1.41

Mean Var(LRS_{Full}) = 5.76
Variance Var(LRS_{Full}) = 1.34

Source: U.S. Census Bureau, 2012-2016 American Community Survey 5-year Summary Files

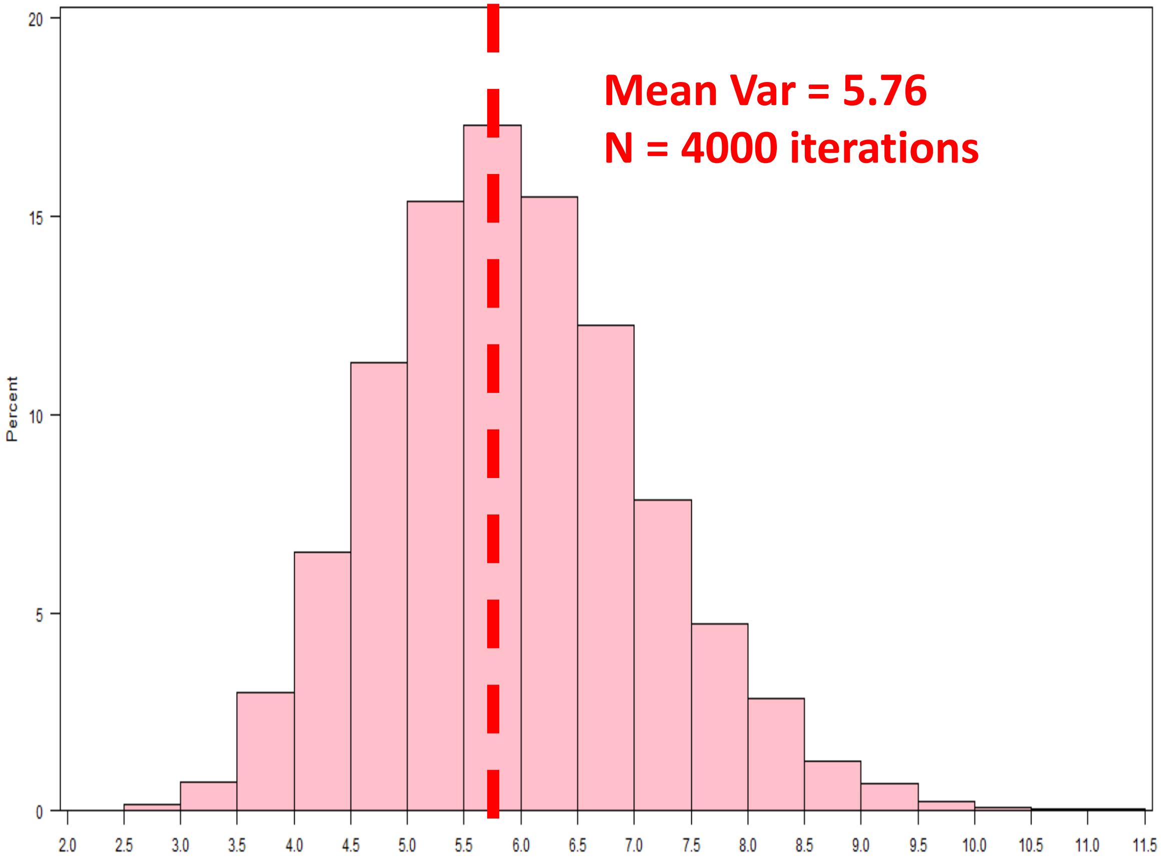
LRS simulation example (2)

Histogram of Var(partial)



Mean Var = 5.95
N = 4000 iterations

Histogram of Var(full)



Mean Var = 5.76
N = 4000 iterations

Source: U.S. Census Bureau, 2012-2016 American Community Survey 5-year Summary Files

Data sources

- As usual, the LRS model was fit with inputs from the 2010 Tract PDB (Census 2010 and 2006-2010 ACS 5-year aggregated data)
- For this study, simulated LRS predictions utilized estimates from the 2018 Tract PDB (Census 2010 and 2012-2016 ACS data)
  - Over 74,000 tracts in the 2018 PDB
  - Of these, about 71,000 tracts were eligible to receive an LRS
  - For simplicity, tracts with missing data on any regressor were excluded

Sample design

To ensure a reasonable degree of representativeness across the U.S., the sample pool of tracts was stratified by two variables:

Census Region

- Northeast
- Midwest
- South
- West

Population Size*

- Less than 3000 people
- 3000 to 4999 people
- 5000 people or more

In total, the sample pool was split into 12 strata. Two samples of 1,000 cases were independently drawn using a proportionally allocated stratified sample design. **Two-sample approach is for research not presented today; the samples were combined for this RQ (n = 1990).**

* Based on total population estimates from the 2012-2016 ACS 5-year file.

Composition of tract universe and samples by Census Region and estimated population

Universe distribution	Less than 3000	3000-4999	5000 or more
Northeast	3626	5516	3848
Midwest	5388	7176	4112
South	6358	9806	9344
West	2776	6647	6032

Sample 1&2 distribution	Less than 3000	3000-4999	5000 or more
Northeast	51	78	55
Midwest	76	102	58
South	90	139	132
West	39	94	86

Shared tracts between 1&2	Less than 3000	3000-4999	5000 or more
Northeast	1	0	0
Midwest	1	0	2
South	1	2	0
West	1	0	2

Source: U.S. Census Bureau, 2012-2016 American Community Survey 5-year Summary Files

Process for tract-level assessment

- For each tract in the combined sample, find $Var_{MC(fit)}$ under both Full and Partial strategies.
- Find $Var_{MC(pred)} = Var_{MC(fit)} + MSE$ under both strategies.
- Conduct F-tests for equality of variance at the tract level using full-to-partial variance estimate ratios.

Examples of Tract-Level MC Variances and Ratios

			F-test (fitted)		F-test (predicted)	
County, State Tract #	MC Var (LRS_{Full})	MC Var (LRS_{Full})	Full/Partial Ratio	P-value	Full/Partial Ratio	P-value
Miami-Dade Cty, FL Tract 0083.05	5.946	5.756	1.033	p = 0.3003	1.006	p = 0.4644
Los Angeles Cty, CA Tract 1352.02	6.353	5.879	1.081	p = 0.1101	1.014	p = 0.4125
Prince George’s Cty, MD Tract 8012.16	7.515	5.820	1.291	p < 0.0001	1.050	p = 0.2182
Collier Cty, FL Tract 0102.15	8.867	5.814	1.525	p < 0.0001	1.091	p = 0.0845

Source: U.S. Census Bureau, 2012-2016 American Community Survey 5-year Summary Files

Tract-level assessment

Combined Sample Sub-group	Number of Sampled Tracts	F-test Summary $(\alpha = 0.1,; df_1 = df_2 = 4000)$*			
		Fitted LRS		Predicted LRS	
		Number Sig	Percent Sig	Number Sig	Percent Sig
All tracts	1989**	177	8.9	2	0.1
Region					
Northeast	367	37	10.1	2	0.5
Midwest	469	30	6.4	0	0.0
South	718	62	8.6	0	0.0
West	435	50	11.5	0	0.0
Pop. Size					
< 3000	507	73	14.4	1	0.2
3000 – 5000	824	65	7.9	0	0.0
> 5000	658	41	6.2	1	0.2

* Family-wise error rate; multiple comparisons controlled with Holm-Bonferroni.

** One tract in the sample was shown to present unusually large outlier characteristics, so it was omitted from this analysis.

Source: U.S. Census Bureau, 2012-2016 American Community Survey 5-year Summary Files

Tract-level assessment summary

- For most tracts, Var(LRS_{Full}) is not sig. different from Var(LRS_{Part}).
- This appears especially so for the predicted LRS values.
- It is reasonable to assume that variance estimates derived under the Partial strategy in a practical application will sufficiently account for the true sampling variability in the Low Response Score.

Source: U.S. Census Bureau, 2012-2016 American Community Survey 5-year Summary Files

Conclusion

- The evidence suggests that an actual (not simulation) variance estimation process using the Full strategy might not yield LRS MOEs that are significantly different from the current process (Larsen, 2017) that uses the Partial strategy.
- **Recommendation: Continue investigation, but favor the Partial strategy over the Full strategy.**

Next Steps

- Expand the simulation parameters
- Explore regional and population size differences
- Consider the block-group LRS
- Publication (Census Bureau Report Series)
- Approval to publish LRS MOEs in the Planning Database

Questions and Comments?

luke.j.larsen@census.gov

Citations

- **Erdman, C. and N. Bates (2017). "The Low Response Score (LRS): A Metric to Locate, Predict, and Manage Hard-to-Survey Populations",** *Public Opinion Quarterly,* Volume 81, Issue 1, 1 March 2017, pp. 144—156.
- **Larsen, L. (2017). "Developing Estimates of Sampling Variability for the Planning Database's Low Response Score",** slide presentation, American Association for Public Opinion Research annual conference, May 18-21, 2017, New Orleans, LA.