Assessment of Computer Availability and Internet Access Statistics to Improve the Planning Database's Low Response Score

JSM Annual Conference, Denver CO
July 27 – August 1, 2019

Luke J Larsen &
Kathleen Kephart
Center for Behavioral Science Methods
U.S. Census Bureau

United States Census Bureau

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

Disclaimer: This presentation is intended to inform people about research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Today's Discussion

1. **Introduction**
   - What is the PDB / What is the LRS / New content for 2019
   - Purpose and Research Questions
2. **Univariate Assessment**
   - Assess ACS Self-Response Rate (RQ1)
   - Assess and select computer/Internet predictor candidates
3. **Modeling**
   - Model selection process and assessment (RQ2)
4. **Prediction**
   - Sample design and Experiment/Control comparison process
   - Assess prediction means and residuals (RQ3)
5. **Conclusion and Next Steps**

Introduction (1): What is the Planning Database?

- Contains most popular American Community Survey (ACS) 5-year tract and block group aggregated estimates
- These estimates are matched to corresponding 2010 Census counts and operational metrics for each geography
- Easier to download than full ACS Summary Files
  - Available in CSV format as well as API
  - Select PDB content available on the Census ROAM application
- Primary source for the Census Bureau's Low Response Score (LRS)
- 2019 PDB released to the public in June 2019
  - Latest updates based on 2013-2017 ACS 5-year Summary Files

Statistics downloaded from American FactFinder or data.census.gov receive clearances from the Disclosure Review Board before being posted. No clearance number is required at this time for such information products. When presenting packages to the DRB that include such data, their provenance must be documented to the satisfaction of the DRB. This statement covers all PDB-based statistics used in this presentation.

Introduction (2): What is the Low Response Score?

- In 1990s, Census Bureau developed a Hard to Count Score (HTC)
  - The higher the score, the harder to count
- For 2020 Census, a new hard-to-survey metric had been developed: the Low Response Score (LRS)
  - Based on OLS model of 25 PDB variables regressed on 2010 Mail Return Rate (MRR). Predicted level of Census self non-response
  - LRS is updated yearly using latest 5-year ACS inputs
- Key limitation: LRS only considers mail self-response – 2020 Census will offer internet, phone, AND mail
- Methodology: see Erdman and Bates (2017)

Introduction (3) New to 2019 PDB: Computer/Internet Variables & ACS Self-Response Rates

- 2013-2017 ACS 5-Year Census Tract Self-Response Rate
  - Never before made public, it is only available on the 2019 PDB and the ROAM app

- ACS 5-year Internet and Technology variables at the census tract level

Households with:
- … smartphone-only access
- … no computing devices
- … a desktop or laptop computer
- … no Internet access
- … broadband Internet access

Population in households with:
- … broadband Internet access and a computing device
- … no computing devices

The ACS 2013-2017 5-year Self-Response Rates have been approved for public release by the U.S. Census Bureau's Disclosure Review Board (approval no.: CBDRB-FY19-244)

Purpose and Research Questions

To determine whether the tract-level ACS self-response rate and the computer-availability & Internet-access metrics on the 2019 PDB could be used to determine whether the LRS model might be improved.

- (**RQ1**) Is ACS Self Response Rate an acceptable proxy for Census 2010 Mail Return Rate?

- (**RQ2**) Does an LRS model with one or more of the new computer/Internet variables yield better model fit than the original model construction?

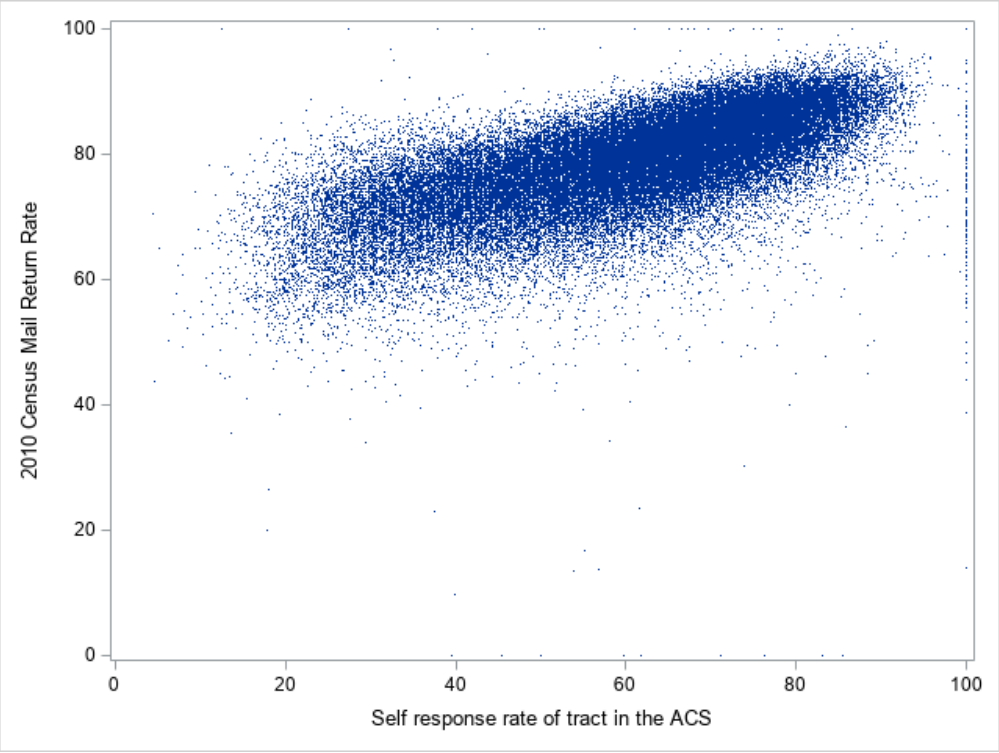
- (**RQ3**) Do LRS predictions differ between the new and original models?

Stage 1

Univariate Assessment

ACS Self-Response Rate Assessment

Mean of Census 2010 Mail Return Rate	78.7%
Mean of 2013-2017 ACS Self Response Rate	60.9%
Correlation between 2010 MRR and ACS SRR	0.68
Correlation between 2019 LRS and ACS SRR	-0.80



Source: U.S. Census Bureau, 2019 Planning Database

Correlation Coefficients between Selected Independent and Dependent Variables

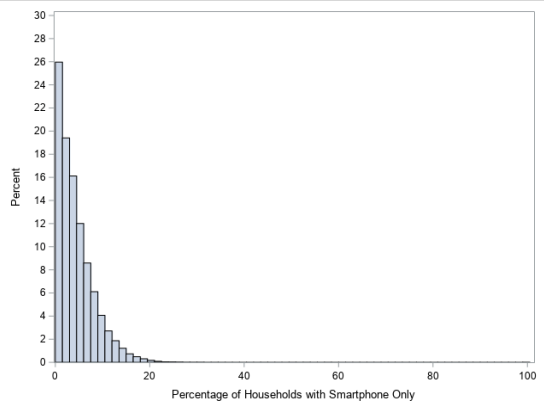
	Broadband access	No computing device	Only smartphone	ACS SRR
Broadband access	1.00			
No computing device	-0.81	1.00		
Only smartphone	-0.62	0.48	1.00	
ACS SRR	0.47	-0.46	-0.58	1.00

Source: U.S. Census Bureau, 2019 Planning Database

Of the seven candidates variables, these three had lowest magnitudes of correlation with each other and with the core 25 LRS predictors.

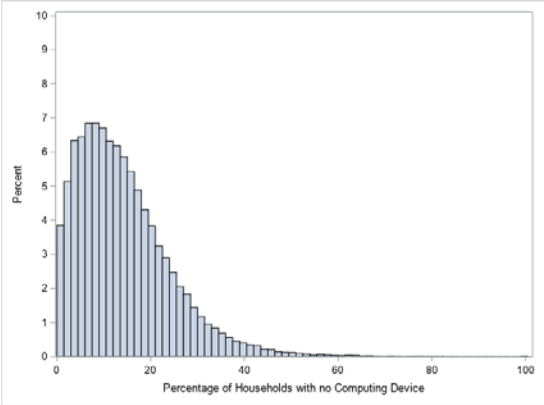
Histograms and Univariate Statistics for Predictor Candidates

Tracts w/ only smartphone



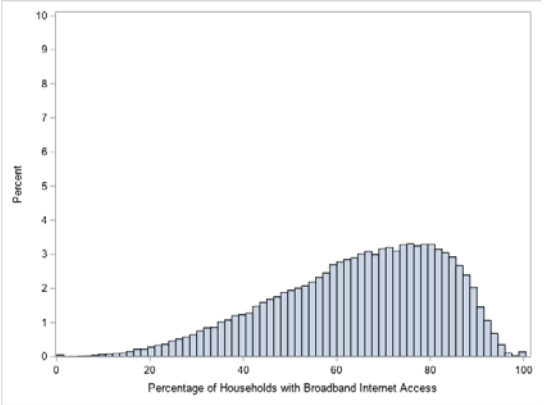
Median	3.4
Mean	4.3
Minimum	0
Maximum	100

Tracts w/ no computing device



Median	12.3
Mean	14.2
Minimum	0
Maximum	100

Tracts w/ broadband access



Median	67.0
Mean	64.6
Minimum	0
Maximum	100

Source: U.S. Census Bureau, 2019 Planning Database

Stage 2

Modeling

LRS Model Selection Process

- Three predictor candidates → Eight regression models to assess

Tier 0 (Control)	Tier 1 (Add one variable)			Tier 2 (Add two variables)			Tier 3 (Add all three)
M0	M1	M2	M3	M4	M5	M6	M7
Core only (25 orig. variables)	Core + nocomp	Core + sphone	Core + broad	Core + nocomp + sphone	Core + nocomp + broad	Core + sphone + broad	Core + all three variables

- Fit models using all tracts in 2013-2017 ACS 5-year Summary Files (N=71,694 excluding PR)

- Used adjusted R^2 to identify best performing model at each tier and partial F-tests to compare each tier's best model against M0

Results (1) Model Fit Statistics

Model	Composition	MSE	Adjusted R^2	Best of Tier?
M0	**Core**	**576913.5**	**0.7824**	**---**
M1	Core + Nocomp	555651.3	0.7837	No
M2	**Core + Sphone**	**556667.3**	**0.7851**	**Yes**
M3	Core + Broad	556608.1	0.7850	No
M4	Core + Nocomp + Sphone	537053.1	0.7866	No
M5	Core + Nocomp + Broad	536163.6	0.7853	No
M6	**Core + Sphone + Broad**	**537124.3**	**0.7867**	**Yes**
M7	**Core + Nocomp + Sphone + Broad**	**518288.9**	**0.7872**	**---**

Source: U.S. Census Bureau, 2019 Planning Database

Results (2) Compare Test Models to Control

Comparison	F-statistic	DF1/DF2	P-value
M2 to M0	0.0907	25/26	0.2344
M6 to M0	0.1480	25/27	0.1369
M7 to M0	**0.1722**	**25/28**	**0.0857**

Source: U.S. Census Bureau, 2019 Planning Database

Conclusion: Only the M7 model (core variables + all three predictor candidates) has significantly better model fit than the control model (α=0.10).

Next, we'll do some comparative analysis of predictions generated by M0 (control) and M7 (experimental).

Stage 3

Prediction

Sample Design for M0-M7 Predictive Comparison

1. Stratified tract pool by two variables (150 strata in total)
   - Geographic location (50 states, excluded PR and DC)
   - Population density (3 groups: Low/Middle/High)

2. Split tract pool in roughly half by drawing a 50% stratified sample

3. From each half-pool, draw a 20% stratified sample
   - Sample A receives the M0 treatment (n_A = 7243 tracts)
   - Sample B receives the M7 treatment (n_B = 7233 tracts)

4. In this way, we can compare predicted scores under the two models from representative samples without having "shared" tracts.

Sample Design Limitations

- Original sample design involved splitting the tract pool into representative sub-groups (70/15/15), modeling on the 70% group and applying the control/experimental models to sub-samples from either of the 15% groups to generate LRS predictions as a cross-validation measure.
- For several reasons (programming errors, time crunch, etc.), the original plan did not work out correctly, so it was replaced with the design outlined in the previous slide.
- This study is a work-in-progress; we expect to return to the original plan for the paper when these problems have been resolved.
- Meanwhile, we stand behind the following findings but place less emphasis upon their significance.

Predictive Comparison Process

- SAS PROC SURVEYMEANS to estimate means and standard errors:
  - Applied weights from both sample stages
  - Fay's BRR for variance estimation
  - FPC ≈ 0.89 applied to standard errors

- Two-sample t-tests (unequal sample size, unequal variances) used to compare differences between the two models.
  - Assume 90% confidence level for all inferences

- <u>Key metrics</u>:

Predicted LRS	$(\hat{Y})$
Residual	$(Y - \hat{Y})$
Absolute Error	$(\lvert Y - \hat{Y} \rvert)$

Results (3) Comparison of Predicted LRS under Different Models – Overall

	Control (M0)		Experimental (M7)		Difference (M0 – M7)		
Means	**Estimate**	**Std. Error**	**Estimate**	**Std. Error**	**Delta**	**Std. Error**	**P-value**
Prediction $(\hat{Y})$	39.109	0.166	39.127	0.131	-0.017	0.211	0.9353
Residual $(Y - \hat{Y})$	-0.184	0.075	0.071	0.070	-0.255	0.103	0.0152
Absolute Error $\lvert Y - \hat{Y} \rvert$	**5.910**	**0.042**	**5.770**	**0.054**	**0.140**	**0.069**	**0.0501**

Source: U.S. Census Bureau, 2019 Planning Database

Conclusion: The experimental model (M7) has a significantly smaller MAE than the control model (M0), indicative of better performance.

Conclusions

- The ACS Self-Response Rate is a reasonable proxy for the Census 2010 Mail Return Rate for conducting this LRS model assessment.
- Adding the three computer/Internet regressors improved the fit of the LRS model by a small, yet significant degree.
- Evidence suggests that LRS predictions had significantly better performance under the experimental model than the control.
- On the basis of these findings, we recommend that the computer/Internet variables should be considered for addition to the official LRS model in future iterations (after 2020 Census).

Next Steps

- Continue analysis (address cross-validation, revisit sample design, domain analysis by selected tract characteristics)
- Await Census 2020 returns and assess the new Census 2020 Self Return Rates
- Construct new LRS model based upon Census 2020 SRR and incorporate the computer/Internet variables into the new model
- Publish the post-Census 2020 LRS in the 2022 Planning Database

Contact Info

Luke J Larsen

Center for Behavioral Science Methods

Luke.j.larsen@census.gov

Kathleen Kephart

Center for Behavioral Science Methods

Kathleen.m.kephart@census.gov