Developing Estimates of Sampling Variability for the Planning Database's Low Response Score

Luke J. Larsen
U.S. Census Bureau

May 18-21, 2017
2017 AAPOR Conference in New Orleans, LA

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.

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Purpose

The purpose of today's discussion is two-fold:

1. Demonstrate how we can produce margin-of-error (MOE) estimates for the Low Response Score (LRS) metric on the Census Bureau's Planning Database (PDB)
2. Gain understanding into variance of LRS predictions and how the interpretability and usability of LRS may be affected.

Overview

1. Purpose and Introduction
   - *What is the PDB? What is the LRS?*
   - *Why should we care about LRS variability?*
2. Methods Review and Development
   - *Prediction variance for MLR models and variance estimation for ACS*
   - *LRS construction – and how to modify it to produce LRS MOEs*
3. Demonstration and Analysis
   - *Mock-up LRS MOEs using publicly available datasets*
   - *How to interpret LRS with MOE*
4. Conclusion and Next Steps

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 self-response rate)
- Based on 25 main-effect inputs: 17 from Census 2010 and 8 from ACS 5-year Summary Files

Why should we care about LRS variability?

- All ACS-based measures on the PDB have an MOE, **except LRS**
- LRS predictions are affected not only by model variance, but also by sampling error in ACS-based inputs
- Without estimates of LRS predicted variance, we:
  - Cannot accurately gauge reliability of individual LRS predictions
  - Cannot determine significance of LRS deltas between 2 geographies
- **Thus, development of LRS MOEs is vital to aid survey planners in making critical decisions on field operations.**

Methods Review and Development

Quick Primer 1: Variance Estimation for Predicted Values in Multivariate Linear Regression

For the multivariate regression model $\boldsymbol{Y} = \widehat{\boldsymbol{\beta}}\boldsymbol{X} + \epsilon$:

$$V\left(y_{Pred,i}\right) = \boldsymbol{x}_{Obs,i}^{T}\left[V\left(\widehat{\boldsymbol{\beta}}\right)\right]\boldsymbol{x}_{Obs,i} + MSE$$

Descriptive version:

Variance of predicted response for new case *i*

=

Observed values for predictor variables in new case *i*

X

Covariance matrix of estimated coefficients in regression model

+

Mean square error of regression model

Quick Primer 2: Variance Estimation by Replication for ACS-based Estimates

- The complex sample design of the ACS is built into a set of 80 replicate weights.
- Generate a base estimate $(\hat{Y})$ using the base weight and 80 replicate estimates $(\hat{Y}_i)$ using the replicate weights. Then, under the replication method, the estimated variance is:

$$\hat{V}(\hat{Y}) = \frac{4}{80}\sum_{i=1}^{80}(\hat{Y} - \hat{Y}_i)^2 \quad ; \quad MOE(\hat{Y}) = 1.645 * \sqrt{\hat{V}(\hat{Y})}$$

Quick Primer 3: Estimation of $V(\hat{\boldsymbol{\beta}})$ by Replication

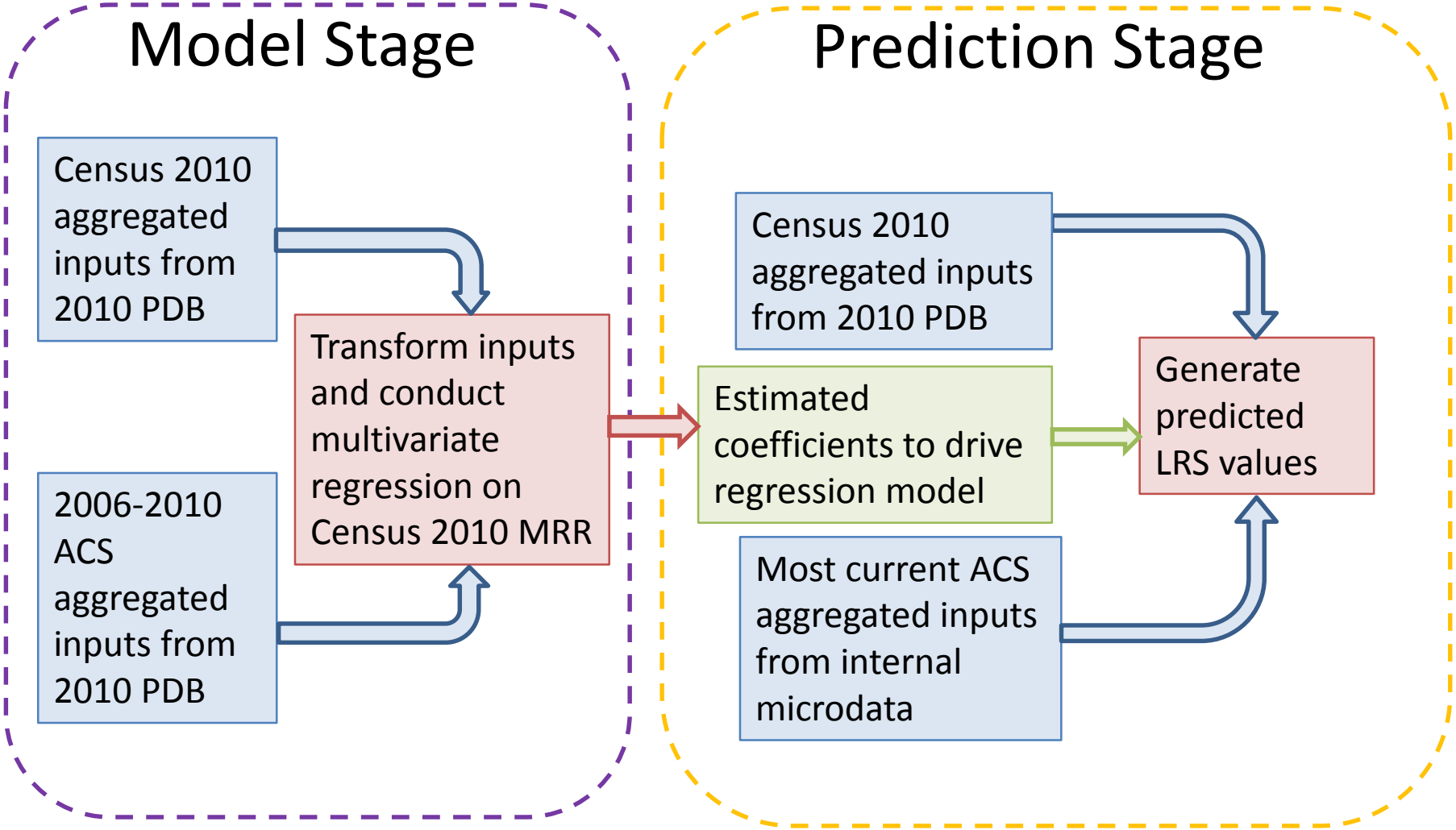
- When some of the $(p-1)$ predictor variables in the MLR model are ACS-based, the $p \times p$ covariance matrix $V(\hat{\boldsymbol{\beta}})$ from the model can be estimated with the replication method.

- The element in position (m, n) of $\hat{V}(\hat{\boldsymbol{\beta}})$ is:

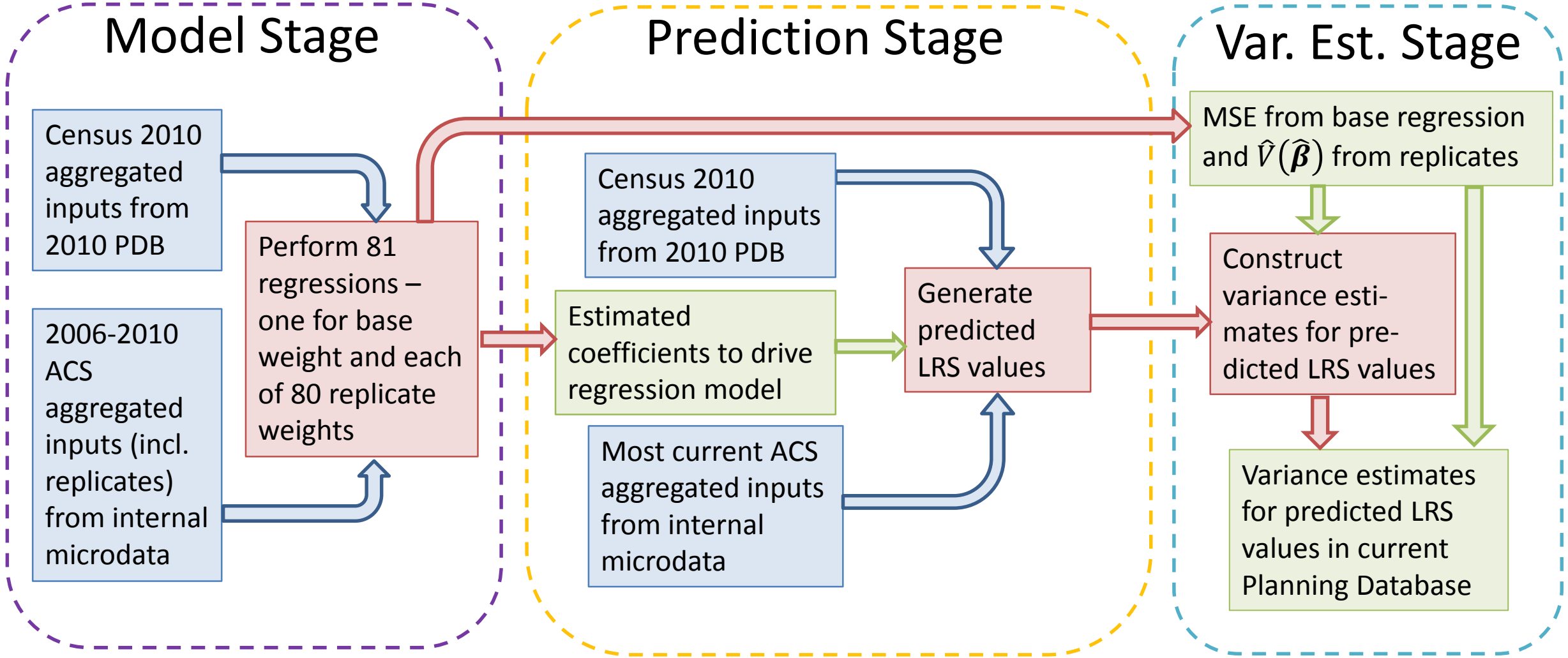
$$\frac{4}{80}\sum_{i=1}^{80}(\hat{\beta}_m - \hat{\beta}_{m,i})(\hat{\beta}_n - \hat{\beta}_{n,i})$$

- This accounts for sampling error in the ACS-based inputs.

LRS Process 1: Current Production Scheme



LRS Process 2: Expand to Construct MOEs (Ideal)



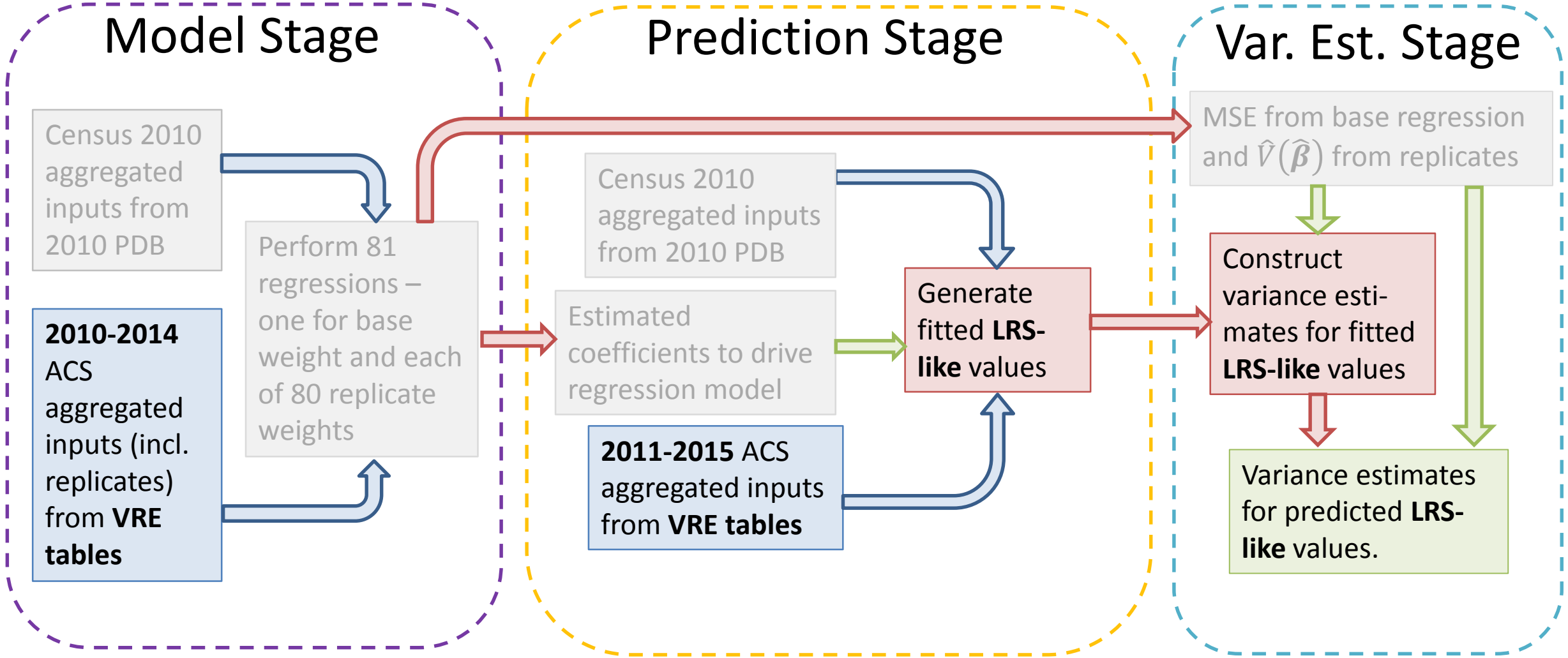
Unexpected Change of Plans

- Original aim was to obtain replicates from internal ACS microdata.
- Due to unforeseen circumstances and schedule conflicts, it was not possible to produce microdata-based output and clear Census internal review process in time for AAPOR.
- **Compromise**: Use publicly-available ACS tables with aggregated replicate estimates instead of microdata to create "modified LRS".
  - Con: Cannot exactly recreate official LRS model or predicted values
  - Pro: This modified process can be done by any external data user

What are the Variance Replicate Estimate Tables?

- Publicly available files that contain estimates, MOEs, and all 80 variance replicates for selected ACS 5-year Detailed Tables.
- Not all featured Detailed Tables are available at block group level, but most are available at tract level.
- Currently available for two most recent ACS 5-year periods: 2010-2014 and 2011-2015.
- Seven of eight ACS-based inputs to LRS tract model can be replicated with the VRE tables (lone exception is “% lived in different unit 1 year ago”).
- https://www.census.gov/programs-surveys/acs/data/variance-tables.html

LRS Process 3: Production for "LRS-like" (Compromise)



Limitations to Variance Estimates of Modified LRS

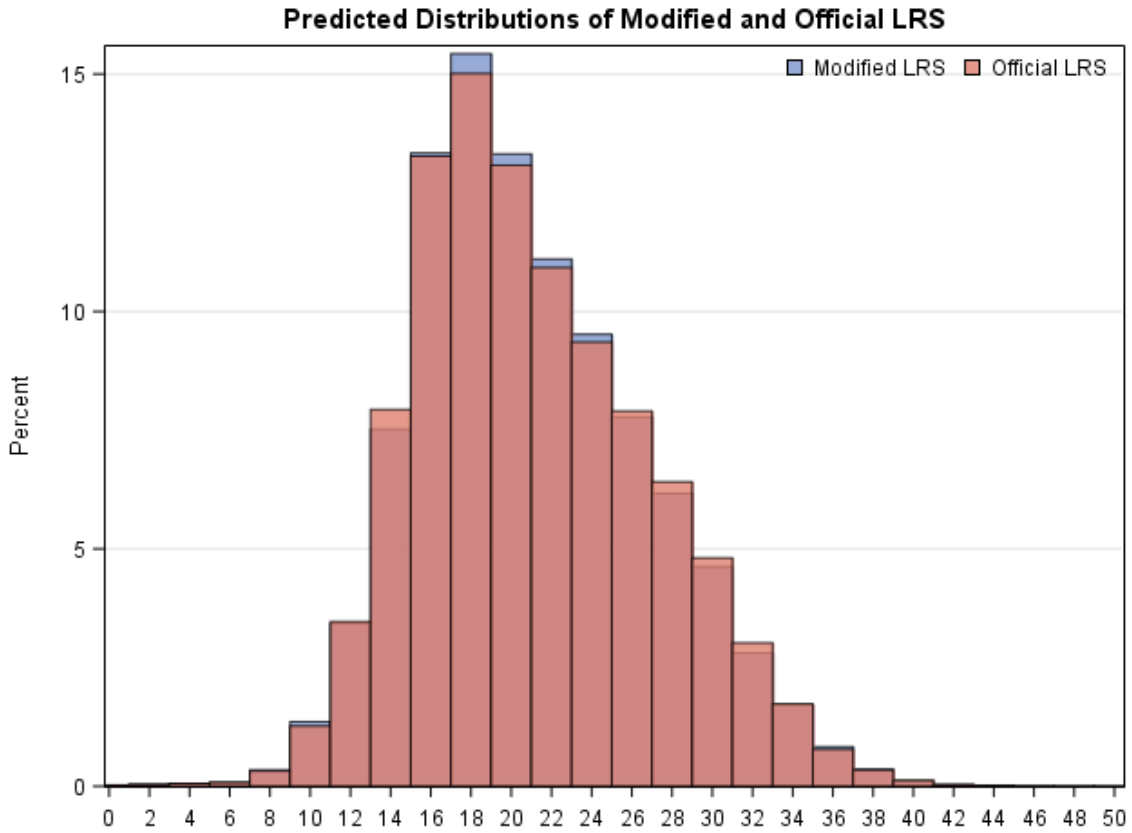
- Does not account for sampling variability among ACS-based inputs at the Prediction Stage → MOEs based on Modified LRS will be underestimated
- Modified LRS ≠ Official LRS → MOEs based on Modified LRS cannot be published in Planning Database
- Today's analysis: for demonstration purposes only!

Demonstration and Analysis

Comparing Modified LRS to Official LRS

- LRS_M: Census 2010 and 2011-2015 ACS
- LRS_O: Census 2010 and 2010-2014 ACS
- $N_{Pred} = 71{,}903$ tracts ; $\rho_{MO} = 0.9940$
- Distribution of differences ($|LRS_M - LRS_O|$):

$\lvert LRS_M - LRS_O \rvert$	Number	Percent	Cumulative
0.0 to 0.1	11598	16.1	16.1
0.1 to 0.5	37188	51.7	67.9
0.5 to 1.0	17583	24.5	92.3
1.0 to 2.0	4694	6.5	98.8
More than 2.0	840	1.2	100.0



Source: U.S. Census Bureau, 2010 Planning Database, 2010-2014 American Community Survey (ACS) 5-year Summary Files, 2010-2014 ACS 5-year Variance Replicate Estimate Tables, 2011-2015 ACS 5-year Variance Replicate Estimate Tables.

Modified LRS Margins of Error

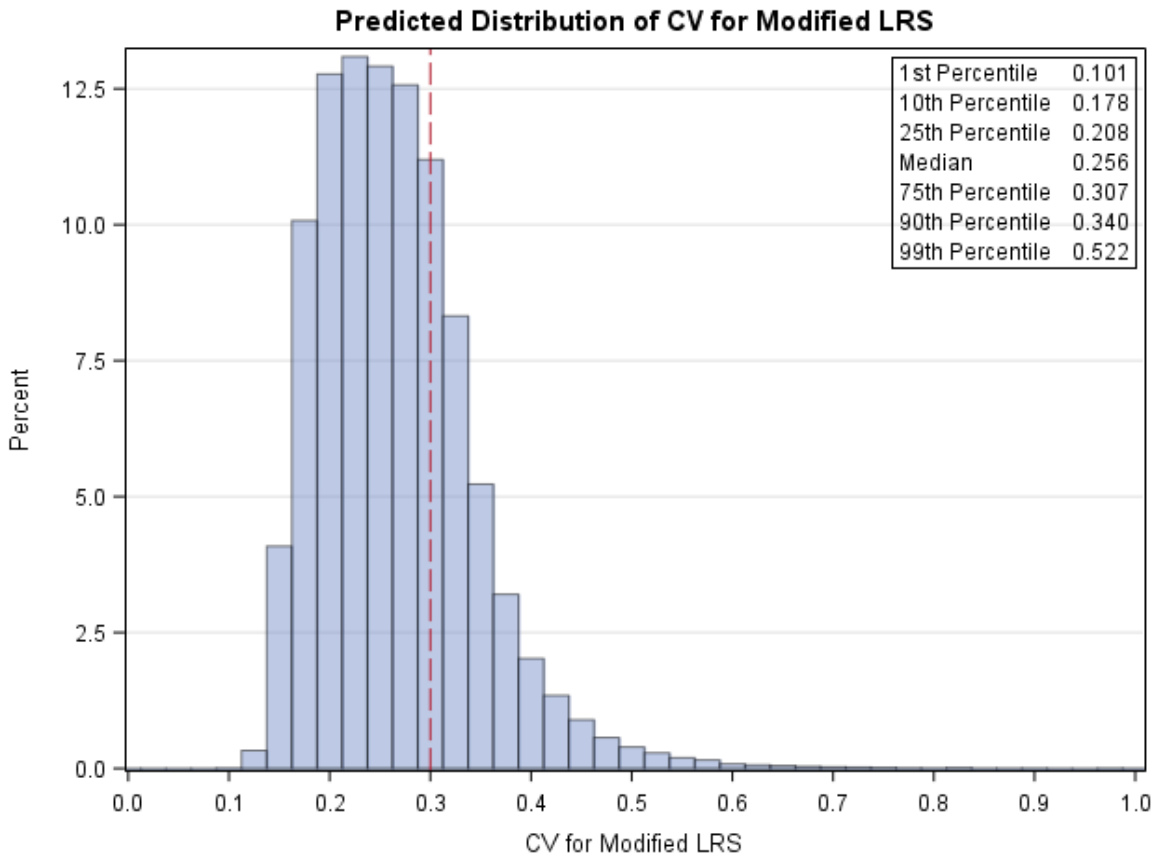
Recall: $\hat{V}(y_{Pred,i}) =$ Mean Response Variance $+ MSE$

For the tract-level LRS-like model, $MSE \cong 26.823$. Then, the lower bound on the 90% MOE of the predicted LRS_M is $1.645\sqrt{0 + 26.823} \cong 8.5$.

MOE Univariate Statistic	Value
Range	[8.5177, 10.9070]
Mean	8.5277
Median	8.5206
1st percentile	8.5181
95th percentile	8.5352
99th percentile	8.5953

- Greater variation among LRS MOEs is likely when sampling variability from ACS-based inputs is introduced.
- Clearly, variance of predicted LRS_M values is dominated by the mean square error.

Modified LRS Coefficients of Variation

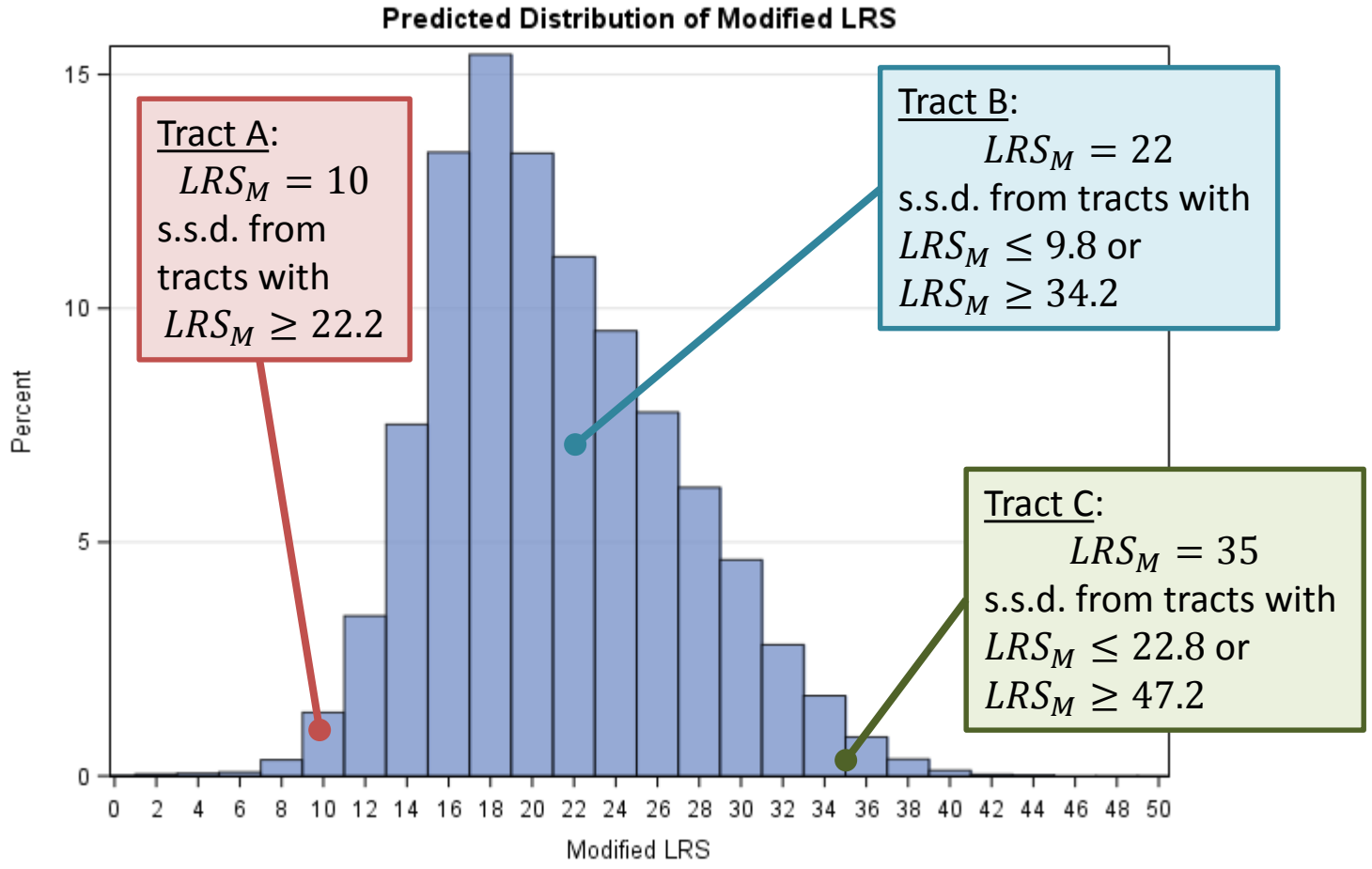


- $CV = SE(LRS_M)/LRS_M$
- Over half of tracts (about 52K tracts, or 72 percent) have $CV < 0.3$, satisfying Census Bureau data quality standards for the modified LRS (51 percent of CVs < 0.3).
- However, applying sampling error of ACS inputs to predictions may result in higher CVs → modified LRS could approach or exceed the standards threshold.

Source: U.S. Census Bureau, 2010 Planning Database, 2010-2014 ACS 5-year Variance Replicate Estimate Tables, 2011-2015 ACS 5-year Variance Replicate Estimate Tables.

Comparing LRS_M Values Between Tracts (1)

Nearly all MOEs for LRS_M are less than 8.6, so a reasonable estimate for the MOE of the difference in LRS_M between any two tracts (assuming independence) is $\sqrt{2(8.6)^2} \cong 12.2$.



Source: U.S. Census Bureau, 2010 Planning Database, 2010-2014 ACS 5-year Variance Replicate Estimate Tables, 2011-2015 ACS 5-year Variance Replicate Estimate Tables.

Comparing LRS_M Values Between Tracts (2)

- Tract A ($LRS_M = 10$) is not s.s.d. from about 62 percent of all tracts.
- Tract B ($LRS_M = 22$) is not s.s.d. from about 97 percent of all tracts.
- Tract C ($LRS_M = 35$) is not s.s.d. from about 35 percent of all tracts.

Takeaway: As a metric for determining relative propensity for self-response to a decennial census at an aggregated level, the modified LRS prediction is useful for very-low or very-high scores, but it is not very reliable for middle-ground scores.

Summary

- The proxy prediction, LRS_M, of the official Low Response Score does have some mean response error at the tract level, though it is dwarfed by the MSE of the regression model.
- The CV analysis of the tract-level predictions indicated sufficient reliability, though wide prediction intervals present a challenge to interpretability and usability.
- With substantially smaller sample sizes per geography, one can surmise that block-group predictions may be susceptible to more severe problems with variability.

Next Steps

- Use ACS microdata to produce accurate aggregated replicates with correct source data (2006-2010 ACS 5-year file).
- Use ACS microdata to generate aggregated replicates at the block-group level.
- Explore correct strategy for incorporating sampling error of ACS-based inputs to the predictive model.
- Once all of above are addressed: add LRS MOE to the Planning Database in a future release.

Questions, Comments, and Concerns?

Luke.J.Larsen@census.gov

Acknowledgements

Many thanks to Nancy Bates, Kathleen Kephart,
Fane Lineback, Ben Reist, Kevin Tolliver, and Brady West
for their advice and support.