Geographic Area Prevalence Estimates for Food Insecurity

Katherine Li[1], Yajuan Si[12], Brady T. West[2], John A. Kirlin[3], and Xingyou Zhang[4]

[1]Dept. of Biostatistics, University of Michigan—Ann Arbor

[2]Institute for Social Research

[3]Kirlin Analytic Services

[4]Bureau of Labor Statistics

November 3, 2021

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Motivation

- Small area estimation (SAE) estimates the population quantity of interest (e.g. food insecurity prevalence) for an "area" (e.g. demographic groups or geographic tract) that has sparse data where direct estimates could be unstable.
- Developing techniques that can improve the SAE precision is crucial for data-driven decision making in health and policy research.
- Model-based SAE approaches fit models to survey data and use external data from auxiliary or population records for prediction.
- However, accounting for multi-stage complex sampling design features, such as strata, clusters, and weights of the survey data, poses a challenge for generalizable estimates.

Motivation

Goal: Produce valid prevalence estimates of food insecurity for tracts, counties, and states ("small areas") in the contiguous U.S. with a sample survey

Setting: The survey was obtained under a complex sampling design and did not include observations from all areas of interest.

We consider two approaches: *weighted analysis* and *synthetic population generation analysis.*

Outline

Datasets

- **National Household Food Acquisition and Purchase Survey (FoodAPS):** Collected person-level data (n=9,894) from 4,826 non-institutionalized households April 2012–January 2013.
  - Multi-stage stratified cluster sampling design. Sampling weights calibrated to totals in the 2013 Current Population Survey (CPS).
  - Included pseudo-strata and pseudo-PSUs for variance estimation.
  - Collected variables expected to be predictive of food security.
- **Enriched sample:** FoodAPS linked with ACS 2010 and SNAP 2012 (n=9,894). Used to fit the outcome model.
- **External Prediction dataset:** ACS 2010 linked with SNAP 2012 (N=7,982,966). Used with estimated model coefficients to make out-of-sample predictions.

General point estimation procedure

Let the outcome be $y_j = 1$ if person j is food-insecure; $y_j = 0$ otherwise.

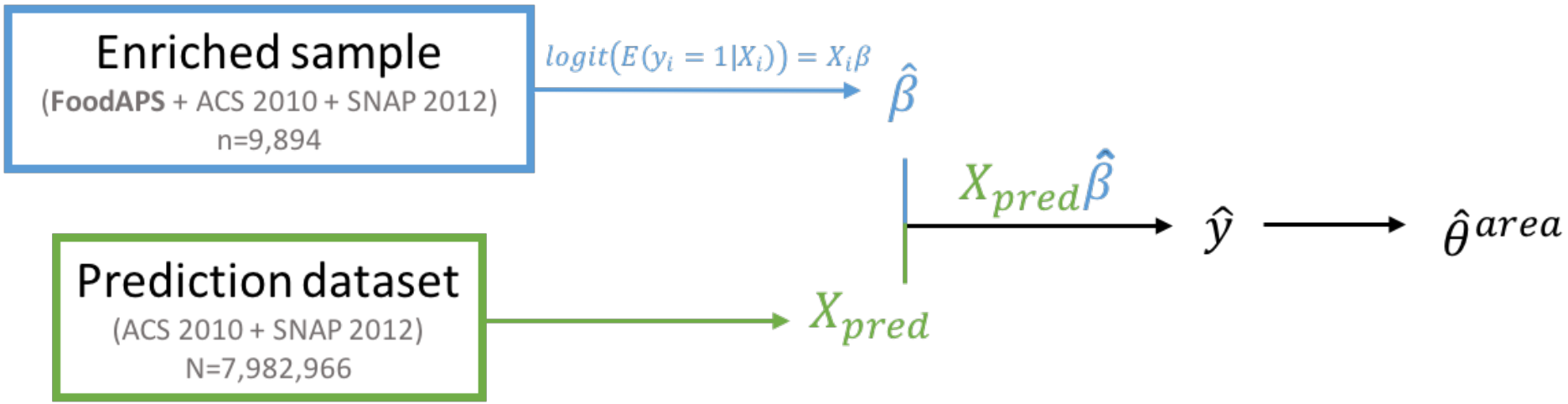
1. Fit the outcome model using the **enriched sample** to obtain coefficient estimates $\hat{\beta}$.

$$logit(P(y_j = 1|X_j)) = X_j\beta$$

2. Use predictors $X_{g,pred}$ from **prediction dataset** to make person-level predictions ($\hat{y}^g = X_{g,pred}\hat{\beta},\ g = 1\ldots, N$)
3. Aggregate relevant $\hat{y}^g$ into small area estimates $\hat{\theta}^{area}$, where *area* is the tract, county, or state of interest.

Note that in step (1), we model the data with weighted logistic regression for the *weighted analysis* and unweighted logistic regression for the *synthetic population generation analysis*.

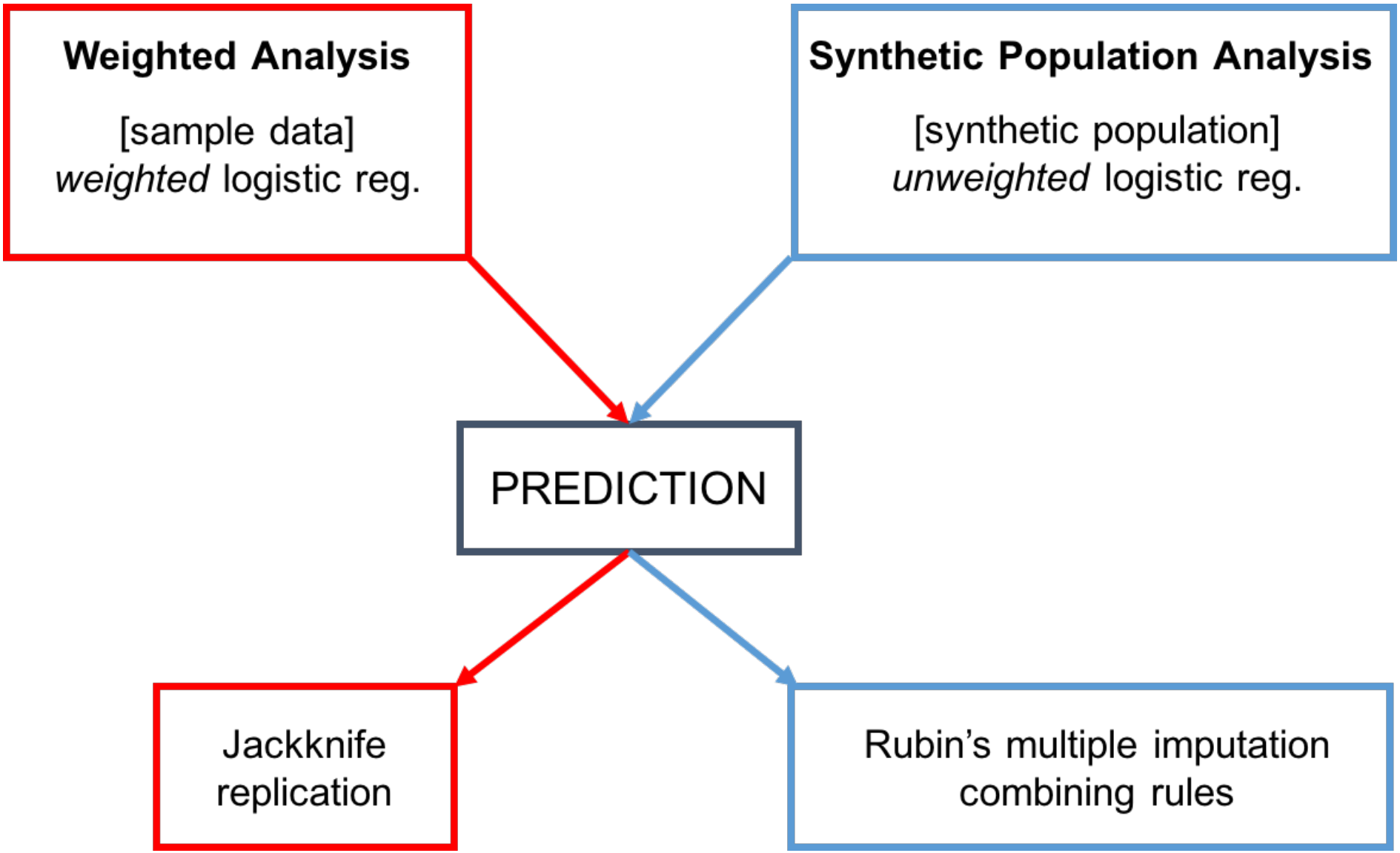
General point estimation workflow



*area refers to a tract, county, state, or US

Variables in X and X_{pred}: **age** (person-level), **race** (person-level)
poverty (tract-level), **education** (tract-level)
poverty (county-level), **NCHS urban/rural status** (county-level)
division (state-level), **poverty** (state-level), **Poverty SNAP** (state-level)

Methods: Overview



Weighted analysis: weighted logit with jackknife SE

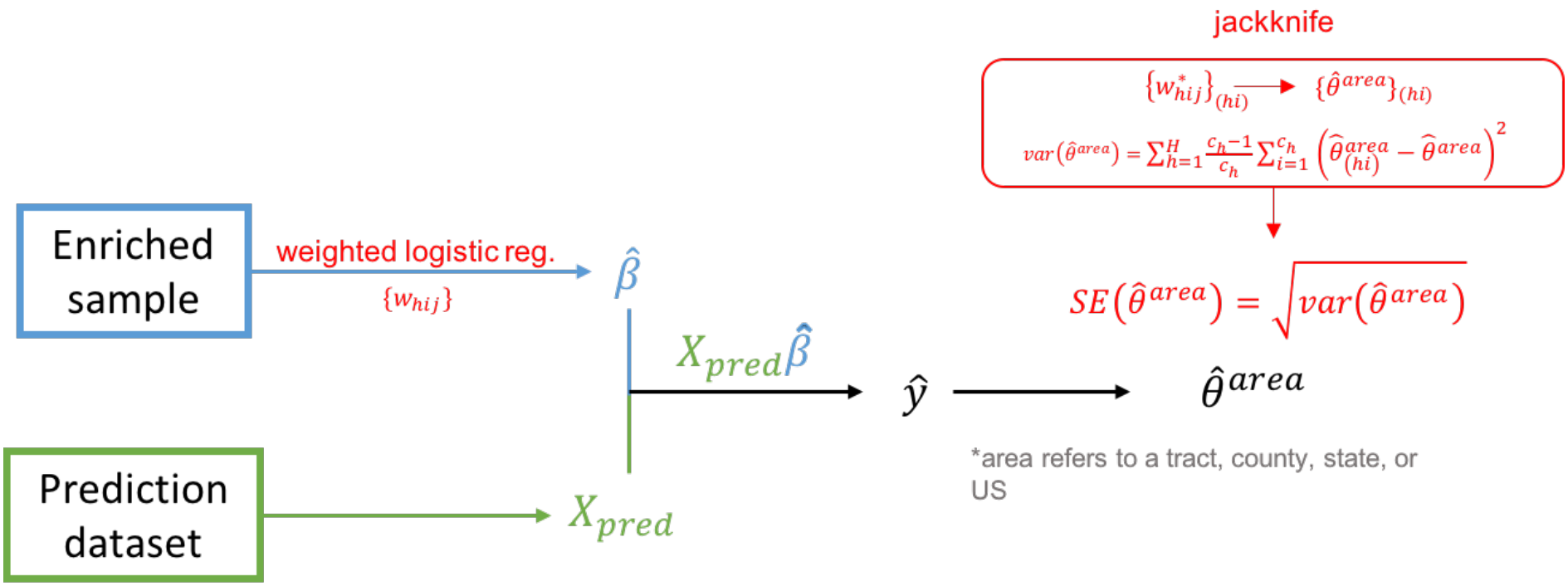
- This approach separates *point* and *variance* estimation steps.
  - Use weighted logistic regression to obtain point estimates
  - Use jackknife replication to obtain variance estimates.

Weighted analysis

Let h index the pseudo-strata; i index the pseudo-PSUs; c_h refer to the number of pseudo-PSUs in stratum h; and $\{w_{hij}\}$ reference the base weights from the enriched sample.

1. Point estimation:
   - Employ the point estimation procedure, using the base weights $\{w_{hij}\}$ in the *weighted* logistic regression.
2. Replicate weight construction:
   - A set of replicate weights $\{w^*_{hij}\}_{(hi)}$ is constructed by dropping observations in the ith PSU and renormalizing weights in stratum h such that $\sum_{i \in h} w^*_{hij} = \sum_{i \in h} w_{hij}$
3. Variance estimation via jackknife replication:
   1. Repeat the point estimation procedure for each set of replicate weights using $\{w^*_{hij}\}_{(hi)}$ in the weighted regression
   2. Obtain $\{\hat{\theta}^{area}_{(hi)}\}$
   3. $var(\hat{\theta}^{area}) = \sum_{h=1}^{H} \frac{c_h}{c_h - 1} \sum_{i=1}^{c_h} (\hat{\theta}^{area}_{(hi)} - \hat{\theta}^{area})^2$

Workflow for weighted analysis



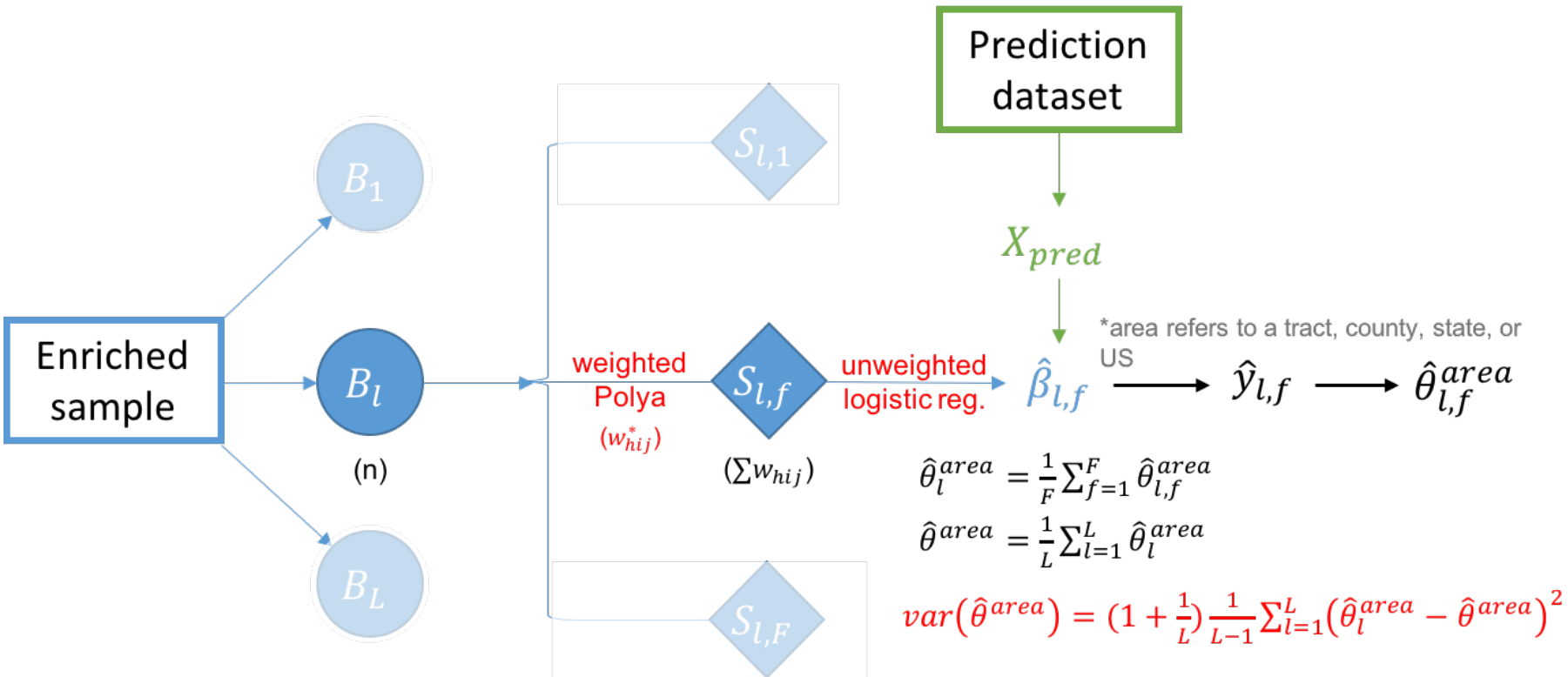
Synthetic Population Generation Analysis

- This approach integrates *point* and *variance* estimation steps.
- The Weighted Finite Population Bayesian Bootstrap (WFPBB) "undoes" the complex sampling design by drawing unobserved units from the weighted Pólya distribution
- These "synthetic populations" can then be analyzed as a SRS without sampling uncertainty.

Synthetic Population Generation Analysis

1. WFPBB replicate weights for bootstrap sample B_l:
   1. For each stratum h, take a SRSWR of $c_h - 1$ PSUs. The number of resamples is m^*_{ih} for the ith PSU.
   2. Construct the replicate weights $\{w^*_{hij}\}$ such that $w^*_{hij} = c_h/(c_h - 1)m^*_{ih}w_{hij}$ and $\sum w^*_{hij} = \sum w_{hij}$.
2. WFPBB synthetic population generation for B_l:
   - Use $\{w^*_{hij}\}$ in the weighted Pólya to draw enough units s.t. after combining with B_l, synthetic population $S_{l,f}$ is of size $\sum_{j=1}^{n} w_{hij}$.
   - Employ the point estimation procedure using the *unweighted* logistic regression on each $S_{l,f}$, $\rightarrow \hat{\theta}^{area}_{l,f}$
   - The estimate for B_l is $\hat{\theta}^{area}_l = \frac{1}{F}\sum_f^F \hat{\theta}^{area}_{l,f}$
3. Point estimation:
   - $\hat{\theta}^{area} = \frac{1}{L}\sum_l^L \hat{\theta}^{area}_l$
4. Variance estimation:
   - $V(\hat{\theta}^{area}) = (1 + 1/L)\frac{1}{L-1}\sum_l^L(\hat{\theta}^{area}_l - \hat{\theta}^{area})^2$

Workflow for WFPBB



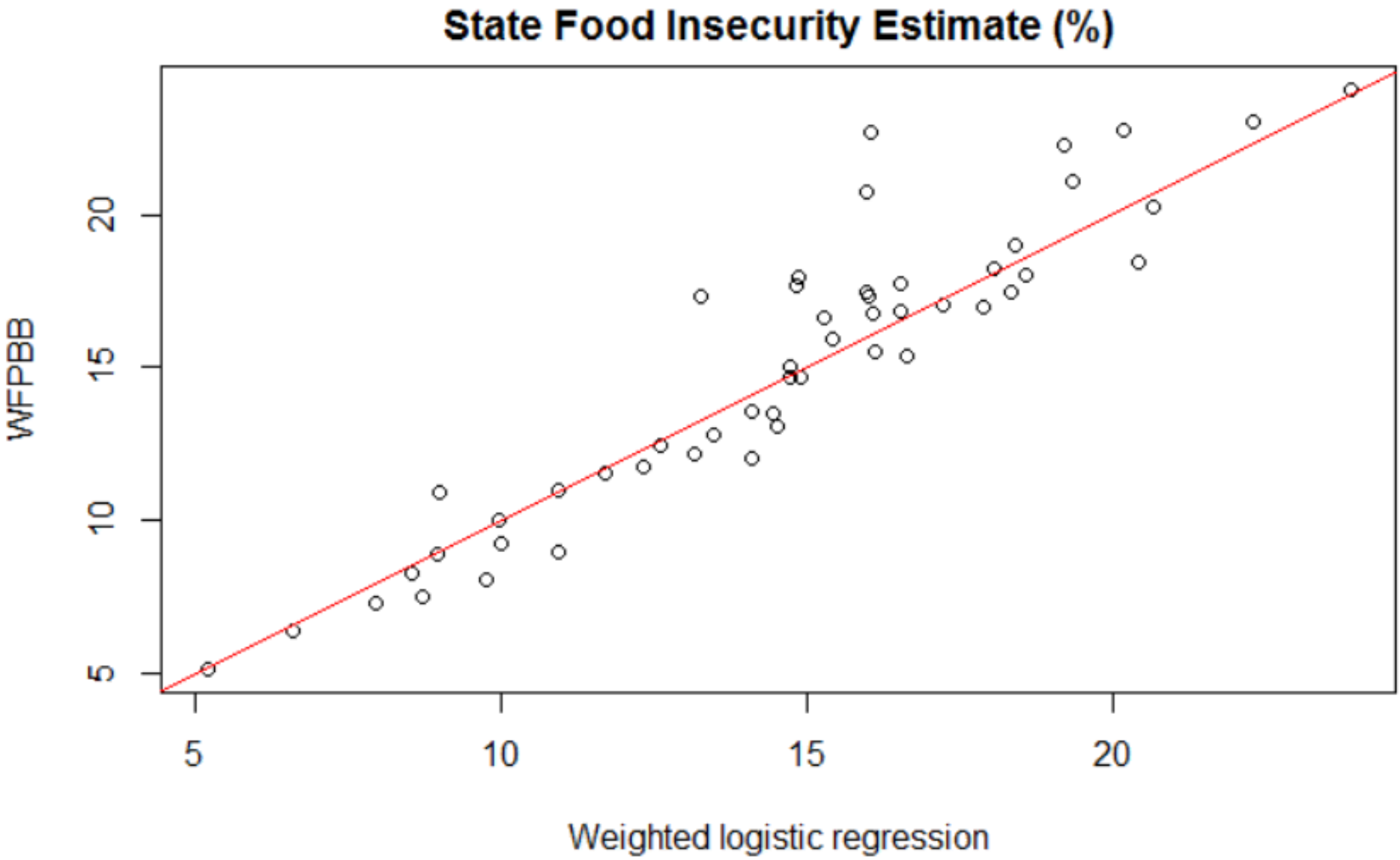
Results

- We report point and total variance estimates for food insecurity prevalence at the U.S. and state levels from the weighted and synthetic population generation analyses.
- Results are compared to external estimates from the **December 2012 CPS Food Security Supplement**.
- 95% confidence intervals are constructed based on the total SE ($\sqrt{\text{sampling SE}^2 + \text{model SE}^2}$)

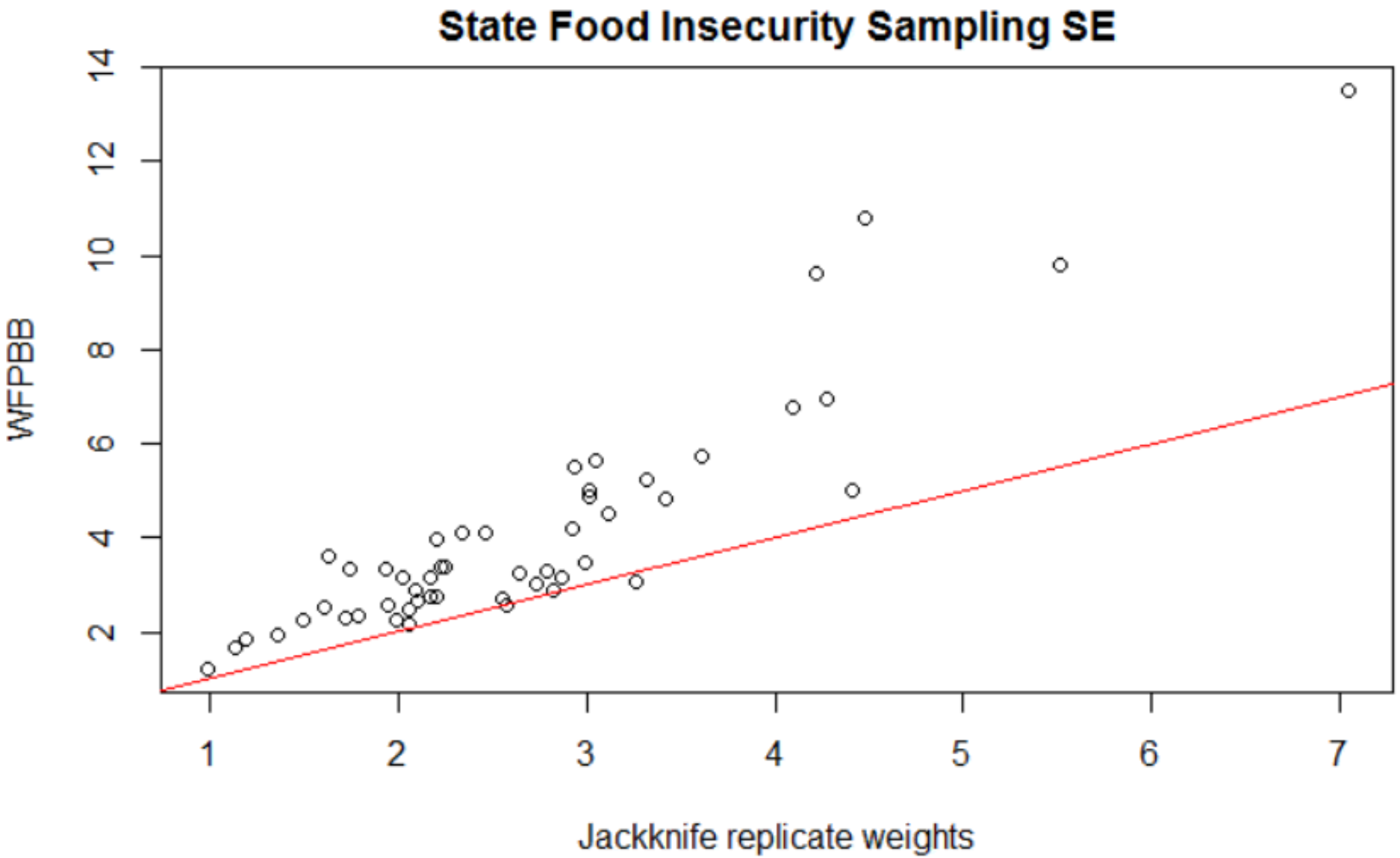
Results: US overall food insecurity prevalence

	Estimate (%)	Total variance	95% CI
Direct estimate	28.2	0.2	(27.3, 29.1)
Weighted estimate	15.7	1.0	(13.6, 17.7)
Weighted analysis	16.4	0.4	(15.0, 17.7)
Synth. pop. analysis	16.5	1.3	(14.2, 18.9)
CPS estimate	14.5	–	–

Results: State-level point estimates



Results: State-level sampling SE



Discussion

- WFPBB estimates higher sampling SE than jackknife because we discard bootstrap samples that do not have all levels of the categorical variables used in the model. SE estimates are much closer to that of jackknife when the model does not include the problematic variables (division, urban/rural status).
- Confidence intervals of the synthetic population analysis cover more of the CPS state-level estimates in out-of-sample areas.
- A substantial portion of tract-level variation is left unaccounted for, and there are no additional predictive tract-level variables in ACS that we can include in the model.
- **Next steps:**
  - Re-structure model so there are fewer sparse variables
  - Use embedded MRP (multilevel regression and poststratification) methods which will allow the model to include FoodAPS variables not found in the prediction data.