Statistical Data Integration

Partha Lahiri
University of Maryland College Park, USA

Based on joint research with Aditi Sen, Doctoral Student,
University of Maryland College Park, USA

MET2023, Warsaw, Poland, July 3-5, 2023

Method

Let

- N_i: the finite population size for the ith area (e.g., state in a nationwide sample survey);
- m: number of areas of interest (e.g., $m = 51$ if we are interested in all US states and the District of Columbia);
- Y_{ij}: value of the outcome variable for the jth unit of the ith area, $i = 1, \cdots, m$; $j = 1, \cdots, N_i$.

Parameter of interest:

$$\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}, \; i = 1, \cdots, m,$$

For the estimation, we have:

- A small sample $\tilde{s}$ of size $\tilde{n}$ from the finite population
  - $\tilde{s}$ contains information on the study variable Y and a vector of auxiliary variables X related to Y for all units.
  - The area sample sizes $\tilde{n}_i$ of $\tilde{s}$ are small; $\tilde{n}_i$ could be zero for some areas.
  - Units of $\tilde{s}$ cannot be linked to the finite population units.

- A big sample s of size n from the same finite population
  - s does not contain information on Y for any unit, but contains the same vector of auxiliary variables X.
  - The area sample sizes n_i of s are large.
  - Units of s cannot be linked to the finite population.
  - There is no or negligible overlap of units between the big sample s and small sample $\tilde{s}$.

- A vector of auxiliary variables at the area level.

Define

$$\bar{Y}_{iw} = \sum_{j=1}^{n_i} w_{ij} Y_{ij},$$

where

- Y_{ij}: unobserved study variable for the jth unit of ith area in the big sample s, $i = 1, \cdots, m;\ j = 1, \cdots, n_i$.

- w_{ij}: known weight assigned to the jth unit of the ith area; we assume $\sum_{j=1}^{n_i} w_{ij} = 1$.

- n_i: sample size for the ith area; we assume n_i is large for each area.

We assume

$$\bar{Y}_i \approx \bar{Y}_{iw},\ \ i = 1, \cdots, m.$$

Under certain assumptions, such an approximation can be justified appealing to the law of large numbers since n_i's are large for all i.

Prediction of $\bar{Y}_i, \; i = 1, \cdots, m$

- For the prediction problem, we assume a *working* model for the entire finite population.
- We assume noninformative sampling so the population working model will hold for both s and $\tilde{s}$.
- We predict Y_{ij} for all units of s using information on both Y and X from $\tilde{s}$, X contained in s, and other state level available auxiliary variables.
- The working model can be fitted using $\tilde{s}$ because it contains information on both Y and X for all units.
- For all units in s, we predict Y_{ij} by:

$$\hat{Y}_{ij} = E(Y_{ij}|\tilde{s})$$

  because this will minimize the mean squared prediction error (MSPE).

A Working Model

For $i = 1, \cdots, m; j = 1, \cdots, N_i$, assume

$$\begin{aligned}
\text{Level 1:} \quad & Y_{ij}|\theta_{ij} \overset{ind}{\sim} \text{Bernoulli}(\theta_{ij}), \\
\text{Level 2:} \quad & \theta_{ij} = \frac{\exp(x_{ij}'\beta + v_i)}{1 + \exp(x_{ij}'\beta + v_i)}, \\
\text{Level 3:} \quad & v_i \overset{iid}{\sim} N(0, \sigma^2),
\end{aligned}$$

where

- β is a vector of unknown fixed effects;
- v_i is random effect specific to the ith area with unknown variance component σ^2.

The best predictor (BP) of Y_{ij} for any unit in s is given by:

$$\hat{Y}_{ij}^{BP} \equiv \hat{Y}_{ij}^{BP}(\beta, \sigma^2) = E\left[\frac{\exp(x_{ij}'\beta + v_i)}{1 + \exp(x_{ij}'\beta + v_i)} \mid \tilde{s}\right],$$

where the expectation is with respect to the conditional distribution of v_i given $\tilde{s}$.

Empirical Best Prediction (EBP) Approach

- Estimate β and σ^2 by a classical method (e.g., maximum likelihood, residual likelihood, adjusted maximum likelihood).
- Let $(\hat{\beta}, \hat{\sigma}^2)$ be an estimator of (β, σ^2).
- EBP of Y_{ij} for any unit in s:

$$\hat{Y}_{ij}^{EBP} = \hat{Y}_{ij}^{BP}(\hat{\beta}, \hat{\sigma}^2).$$

- EBP of $\bar{Y}_i$:

$$\hat{\bar{Y}}_i^{EBP} \approx \sum_{j=1}^{n_i} w_{ij} \hat{Y}_{ij}^{EBP},$$

Data Analysis

Data sources:

- `PEW`: Pew Research Organization's October 2016 Political Survey.
- `CPS`: 2016 Voting and Registration Supplement to the Current Population Survey (CPS).
- `Actual 2016 Election Result`

Study variable: Voting preference, a binary variable taking on the value 1 if the person prefers to vote for Clinton and 0 otherwise.

Areas of interest example: States and DC

Table: List of unit level auxiliary variables

Predictor	Levels	Values
Age	4	18-29 years, 30-44 years, 45 - 64 years, 65+ years
Gender	2	Male or female.
Race	3	White, Black or Hispanic.
Education	4	Higher Secondary, Some college, College Graduate or Postgraduate
Region	4	Northeast, South, North Central or West

Area level auxiliary variable: state specific percentage of voters who voted for Obama in the 2012 presidential election.

Figure: Multiple survey data and structure

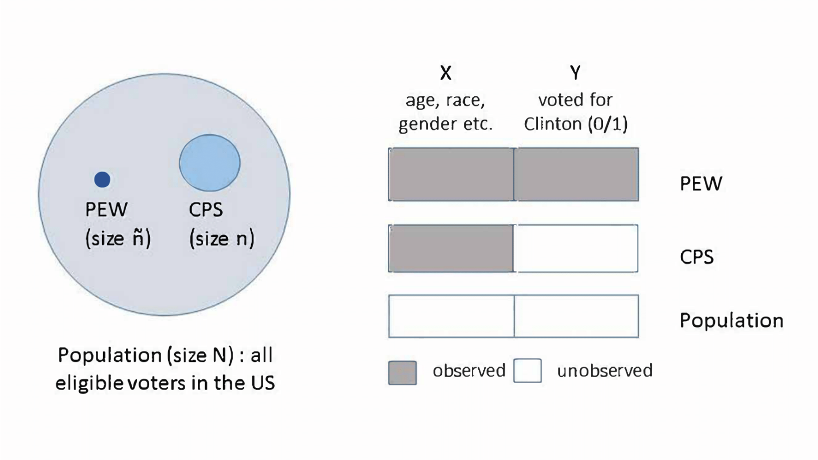


Table: Model 1: All auxiliary variables

	Est. Std.	Error	z value	$Pr(> \|z\|)$	
(Intercept)	-0.81	0.23	-3.53	4.21e-4	***
age: 30-44 years	-0.14	0.20	-0.71	0.48	
age: 45-64 years	-0.34	0.18	-1.95	0.05	.
age: 65+ years	-0.18	0.19	-0.95	0.34	
gender: female	0.64	0.11	5.83	5.61e-09	***
race: black	3.05	0.32	9.68	¡ 2e-16	***
race: hispanic	1.12	0.21	5.39	7.19e-08	***
some college	0.11	0.16	0.66	0.51	
college graduate	0.48	0.16	3.06	2.2e-3	**
postgraduate	1.06	0.17	6.33	2.43e-10	***
South	-0.25	0.19	-1.31	0.19	
North Central	-0.09	0.19	-0.48	0.63	
West	0.14	0.19	0.75	0.45	
voting % Obama	0.97	0.20	4.78	1.82e-06	***

Table: Model 2: Significant auxiliary variables only

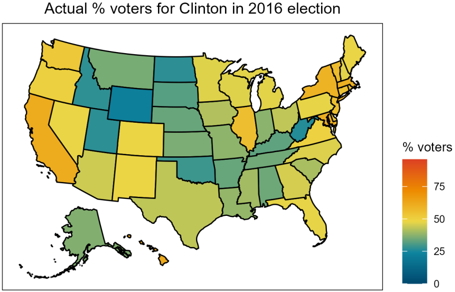
	Est. Std.	Error	z value	$Pr(>\|z\|)$	
(Intercept)	-0.96	0.12	-8.35	¡ 2e-16	***
age: 45-64 years	-0.22	0.11	-1.98	0.048	*
gender: female	0.63	0.11	5.75	8.77e-09	***
race: black	3.01	0.31	9.58	¡ 2e-16	***
race: hispanic	1.13	0.21	5.51	3.62e-08	***
college graduate	0.42	0.13	3.34	8.48e-4	***
postgraduate	0.99	0.14	6.99	2.75e-12	***
voting % Obama	1.10	0.19	5.70	1.13e-08	***

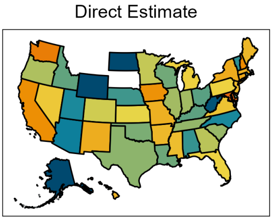
Estimated standard deviation of random effect is 0.21.

Table: Model selection

Model	Model Description	AIC	BIC
Model 1	Mixed effect all covariates	2032.5	2114.1
Model 2	Mixed effect significant covariates only	2026.3	2075.2

From the model selection criteria AIC and BIC, we conclude that **model 2 is better than Model 1** and we choose Model 2 for prediction of voting %.


Actual % voters for Clinton in 2016 election
% voters
75
50
25
0


Direct Estimate

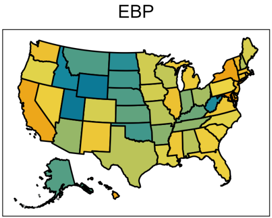

EBP

Figure: Actual and predicted valus for all states

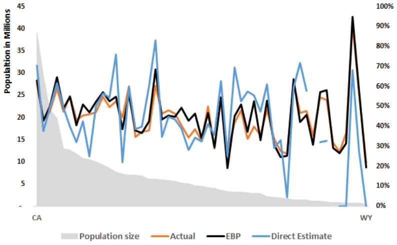


Figure: Direct Estimator SE and Root MSPE from EBP

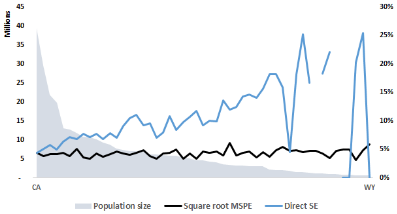


Table: Comparison of direct and EBP for a few selected states

State	Pop Size	Actual (%)	Direct Est SE (%)	EBP Root MSPE (%)
CA	39 mil	61.5	70.5 (4.3)	63.1 (4.4)
FL	21 mil	47.4	49.2 (5.7)	50.0 (4.1)
MD	6 mil	60.3	83.2 (7.0)	68.4 (3.3)
MT	110 k	35.4		30.8 (4.7)
SD	895 k	31.7		29.3 (4.8)
AK	732 k	36.6	0 (0)	31.4 (5.0)
DC	670 k	90.9	68.2 (20.2)	95.0 (3.1)
WY	578 k	21.9	0 (0)	19.5 (5.8)

Table: Summary evaluation measures

Measure	Formula	Direct	EBP
ASD	$\sum_{i=1}^{51}(\hat{Y}_i^{est} - \hat{Y}_i^{act})^2$	2137.3	18.9
RASD	$\sqrt{ASD}$	46.2	4.3
AAD	$\sum_{i=1}^{51}\left\lvert \hat{Y}_i^{est} - \hat{Y}_i^{act} \right\rvert$	44.6	3.5

Concluding Remarks and Extensions

- In our application, EBP method improves on the direct method considerably. There was **no sample** for Montana and South Dakota in PEW survey data. But we can obtain estimates for those using EBP method.
- Use of a bigger survey allows us to include many relevant auxiliary variables in the working model.
- In our application, estimate of the random effects variance is positive. However, for parametric bootstrap samples, we observed 0 estimates for the random effects variance.
- We have extended an adjusted maximum likelihood method to get around the problem associated with the boundary value problem.

Contact Information

Partha Lahiri
Professor and Director, Joint Program in Survey Methodology
& Professor, Department of Mathematics
1218 Lefrak Hall
University of Maryland
College Park, MD 20742
Email: plahiri@umd.edu

Thank You!