

Maximum likelihood estimation of response propensity to a nonprobability survey

Vladislav Beresovsky ¹

¹ U.S. Bureau of Labor Statistics, OSMR

2023 FCSM Research & Policy Conference

October 24-26, 2023

This is a joint work with Julie Gershunskaya (OEUS) and Terrance Savitsky (OSMR)

The views expressed are those of the authors and do not reflect the official policy or position of U.S. Bureau of Labor Statistics

Outline

Survey statistics: from probability samples to data integration

Modeling response propensity to a nonprobability survey

Theoretical properties of the estimators

Simulation study to compare estimates by different methods

Discussion and next steps

References

- ▶ Problems with traditional probability surveys
 - ▶ Labor intensive and time consuming data collection
 - ▶ Increasing costs
 - ▶ Decreasing response rates
 - ▶ Restricted geography of data collection
- ▶ Proliferation of relatively inexpensive and expedient nonprobability data sources:
 - ▶ Web surveys (opt-in or online panels)
 - ▶ Extracts from administrative records

Outline

Survey statistics: from probability samples to data integration

Modeling response propensity to a nonprobability survey

Theoretical properties of the estimators

Simulation study to compare estimates by different methods

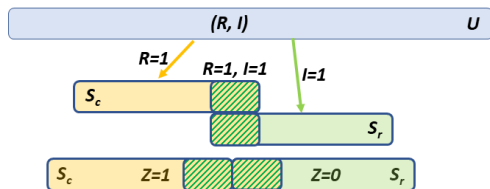
Discussion and next steps

References

IPW estimation from nonprobability samples

- ▶ Naïve unweighted estimates of population parameters from nonprobability sample may be biased
- ▶ To correct for a bias, we focus on Inverse Propensity Weighting (IPW) methods, based on estimated latent response propensity $\pi_c(\mathbf{x})$ to a nonprobability survey
- ▶ Methods for estimation of $\pi_c(\mathbf{x})$ from combined nonprobability (convenience) sample S_c and probability (reference) sample S_r :
 - ▶ pseudo-likelihood based approaches (Chen, P. Li, and Wu (2020), Wang, Y. Li, and Valliant (2021))
 - ▶ likelihood based approach (Savitsky et al., 2022)
- ▶ We present and compare theoretical properties of these methods and illustrate the differences using numerical study and simulations

Combined sample $S = S_c \oplus S_r$



$\pi_r(\mathbf{x}_i) = P(I_i = 1 | i \in U, \mathbf{x}_i)$ are *known* reference sample selection probabilities

$\pi_c(\mathbf{x}_i) = P(R_i = 1 | i \in U, \mathbf{x}_i)$ are *unknown* response propensity to convenience survey

Stack samples: $S = S_c \oplus S_r$ (overlapping units included twice)

Z is *observed* indicator on S : $Z_i = 1 | i \in S_c$ and $Z_i = 0 | i \in S_r$

$\pi_z(\mathbf{x}_i) = P(Z_i = 1 | i \in S, \mathbf{x}_i)$ can be estimated

Core Relationship for Independent Sampling Probabilities (CRISP)

Response propensity $\pi_c(\mathbf{x}_i)$ to convenience survey can be found from CRISP:

$$\pi_z(\mathbf{x}_i) = \frac{\pi_c(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i) + \pi_r(\mathbf{x}_i)}$$

Elliott (2009) and Elliott and Valliant (2017) present this formula assuming negligible fraction of overlapping units.

The proof by Savitsky et al. (2022) makes no assumption about the overlapping units.

Implicit Logistic Regression (ILR)



Bernoulli log-likelihood for *observed* Z on 'stacked' $S = S_c \oplus S_r$:

$$l^{\text{ILR}}(Z, \pi_z(\phi)) = \sum_{S_c \oplus S_r} Z_i \log \pi_{zi} + (1 - Z_i) \log (1 - \pi_{zi})$$

Use CRISP to parameterize $\pi_{zi}(\phi)$ as composite function:

$$\pi_{zi}(\phi) = \frac{\pi_{ci}(\phi)}{\pi_{ci}(\phi) + \pi_{ri}}, \text{ where } \text{logit}[\pi_{ci}(\phi)] = \mathbf{x}_i \phi$$

Parameters ϕ can be estimated by solving score equations $\frac{\partial l^{\text{ILR}}}{\partial \phi} = 0$ with Newton-Raphson iterations.

Pseudo-ILR (PILR)



Bernoulli log-likelihood on 'stacked' $S_c \oplus U$ (Wang, Y. Li, and Valliant, 2021):

$$l(Z, \pi_z(\phi)) = \sum_{S_c \oplus U} Z_i \log \pi_{zi} + (1 - Z_i) \log(1 - \pi_{zi})$$

Pseudo log-likelihood on $S = S_c \oplus S_r$ with $w_{ri} = \pi_{ri}^{-1}$:

$$l^{\text{PILR}}(Z, \pi_z(\phi)) = \sum_{S_c \oplus S_r} Z_i \log \pi_{zi} + w_{ri} (1 - Z_i) \log(1 - \pi_{zi}).$$

Use CRISP to parameterize $\pi_{zi}(\phi)$ as composite function:

$$\pi_{zi}(\pi_{ci}(\phi)) = \frac{\pi_{ci}(\phi)}{\pi_{ci}(\phi) + 1}, \text{ where } \text{logit}[\pi_{ci}(\phi)] = \mathbf{x}_i \phi.$$

If $S_r \equiv U$ then $l^{\text{PILR}}(\phi) \equiv l^{\text{ILR}}(\phi)$.

Likelihood Estimating Equations (CLW)

S_c $R=1$

$R=0$ U

Bernoulli log-likelihood of *unobserved* R on U (Chen, P. Li, and Wu, 2020):

$$\begin{aligned}l(R, \pi_c(\phi)) &= \sum_U R_i \log \pi_{ci} + (1 - R_i) \log (1 - \pi_{ci}) \\ &= \sum_{S_c} \log \left\{ \frac{\pi_{ci}}{1 - \pi_{ci}} \right\} + \sum_U \log (1 - \pi_{ci})\end{aligned}$$

Pseudo log-likelihood with survey weights $w_{ri} = \pi_{ri}^{-1}$

$$l^{\text{CLW}}(\pi_c(\phi)) = \sum_{S_c} \log \left\{ \frac{\pi_{ci}}{1 - \pi_{ci}} \right\} + \sum_{S_r} w_{ri} \log (1 - \pi_{ci}),$$

Parameterize $\text{logit}(\pi_{ci}(\phi)) = \mathbf{x}'_i \phi$ and solve score equations for ϕ .

A key difference between the methods

- ▶ CLW *directly* models the response indicator to nonprobability survey $R_i \sim \text{Bernoulli}(\pi_c(\mathbf{x}_i))$. If R_i is observed, then CLW is optimal. However, R_i is generally *unobserved*, unless $S_r = U$.
- ▶ ILR and PILR use CRISP to *implicitly* model R_i by modeling $Z_i \sim \text{Bernoulli}(\pi_z(\pi_c(\mathbf{x}_i)))$, the *observed* indicator of S_c on 'stacked' sample $S = S_c \oplus S_r$

Outline

Survey statistics: from probability samples to data integration

Modeling response propensity to a nonprobability survey

Theoretical properties of the estimators

Simulation study to compare estimates by different methods

Discussion and next steps

References

Hajek estimator of population mean

Hajek IPW estimator of finite population mean

$\mu = N^{-1} \sum_{i \in U} Y_i$ from the convenience sample with estimated response propensity:

$$\hat{\mu} = \frac{1}{\hat{N}} \sum_{i \in S_c} \frac{y_i}{\hat{\pi}_{ci}},$$

where $\hat{N} = \sum_{i \in S_c} (\hat{\pi}_{ci})^{-1}$.

How $\text{Var}(\hat{\mu})$ is inflated due to estimation of $\hat{\pi}_{ci}$?

Estimating equations for $\eta = (\mu, \phi)$

μ - population mean, ϕ - propensity model parameters.
Consider the system of estimating equations:

$$\Phi(\eta) = \begin{pmatrix} S(\mu) \\ \mathbf{S}(\phi) \end{pmatrix} = 0,$$

where

$$S(\mu) = N^{-1} \sum_{i \in S_c} \pi_{ci}^{-1} (y_i - \mu) = 0$$

and $\mathbf{S}(\phi) = \mathbf{S}_c(\phi) + \mathbf{S}_r(\phi)$ are score equations for ILR, PILR or CLW.

Variances of $\eta = (\mu, \phi)$

Variances of parameter estimates are

$$\text{Var}(\hat{\eta}) = \mathbf{H}^{-1} \text{Var}\{\Phi(\eta)\} \mathbf{H}^{-1},$$

where

$$\text{Var}\{\Phi(\eta)\} = \begin{pmatrix} \text{Var}(S(\mu)) & \text{Cov}(S(\mu), \mathbf{S}_c(\phi)) \\ \text{Cov}(S(\mu), \mathbf{S}_c(\phi)) & \text{Var}(\mathbf{S}(\phi)) \end{pmatrix},$$

$$\mathbf{H} = -E \left[\frac{\partial \Phi(\eta)}{\partial \eta} \right] \text{ is Hessian.}$$

Variances of $\eta = (\mu, \phi)$ (cont'd)

$$\text{Var}(\hat{\mu}) = \text{Var}(S(\mu)) - 2\mathbf{b}\text{Cov}(S(\mu), \mathbf{S}_c(\phi)) + \mathbf{b}\text{Var}(S(\phi))\mathbf{b}^T$$

$$\text{Var}(\hat{\phi}) = \mathbf{H}_{\phi\phi}^{-1}\text{Var}(S(\phi))\mathbf{H}_{\phi\phi}^{-1}$$

where

$$\text{Var}(S(\phi)) = \text{Var}(\mathbf{S}_c(\phi)) + \text{Var}(\mathbf{S}_r(\phi))$$

$$\mathbf{b} = \mathbf{H}_{\phi\phi}^{-1}\mathbf{H}_{\phi\mu}, \mathbf{H}_{\alpha\beta} = -E\left[\frac{\partial\Phi(\beta)}{\partial\alpha}\right]$$

All methods are asymptotically equivalent:

$$\text{Var}(\hat{\eta}) \sim O(n_r^{-1} + n_c^{-1})$$

However, $\text{Var}(\mathbf{S}_r(\phi))$ is a major contributor to the differences between ILR, PILR and CLW methods.

Design variance and “overlap” between S_c and S_r

$V_d(\cdot)$ is design variance with respect to sampling indicator I_{ri} :

$$\text{ILR: } \text{Var}(S_r(\phi)) = V_d \left(\sum_U I_{ri} \frac{g_i}{1+g_i} (1 - \pi_{ci}) \mathbf{x}_i \right)$$

$$\text{PILR: } \text{Var}(S_r(\phi)) = V_d \left(\sum_U I_{ri} g_i \frac{1 - \pi_{ci}}{1 + \pi_{ci}} \mathbf{x}_i \right)$$

$$\text{CLW: } \text{Var}(S_r(\phi)) = V_d \left(\sum_U I_{ri} g_i \mathbf{x}_i \right)$$

$g_i = \pi_c(\mathbf{x}_i) / \pi_r(\mathbf{x}_i)$ - depends on overlap between S_c and S_r in covariate-defined domains \mathbf{x}_i .

- ▶ “High” overlap: $\pi_c(\mathbf{x}_i)$ and $\pi_r(\mathbf{x}_i)$ are similar for all \mathbf{x}_i , and $g_i \sim n_c/n_r$ fluctuate around constant.
- ▶ “Low” overlap: $\pi_c(\mathbf{x}_i)$ and $\pi_r(\mathbf{x}_i)$ are dissimilar, and $g_i \in (0, \infty)$.

For ILR, $\frac{g_i}{1+g_i} \in (0, 1)$.

For PILR, variability due to large g_i is reduced by $\frac{1-\pi_{ci}}{1+\pi_{ci}}$.

High and Low overlaps between S_c and S_r in covariate domains

- ▶ Covariates: $X \sim N(0, 1)$
- ▶ S_c response propensity: $\text{logit}(\pi_{ci}) = \phi_0 + \phi_c * x_i$, $\phi_c = 1.0$
- ▶ S_r PPS size measure: $\text{logit}(M_{ri}) = 1.0 + \phi_r * x_i$

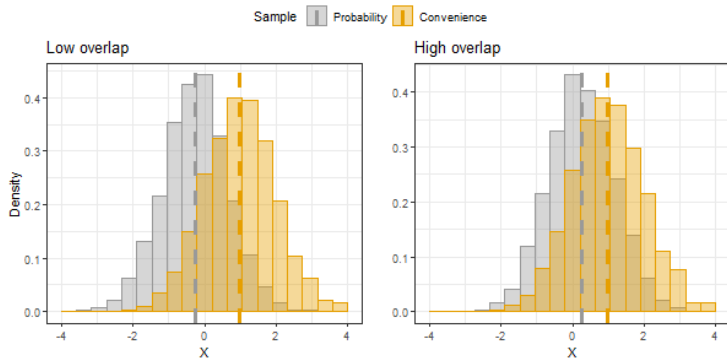


Figure: Low overlap: $\phi_r = -\phi_c$; High overlap: $\phi_r = \phi_c$

Ratio of $SE(\hat{\phi}_1^{\text{ILR}})$ to $SE(\hat{\phi}_1^{\text{CLW}})$

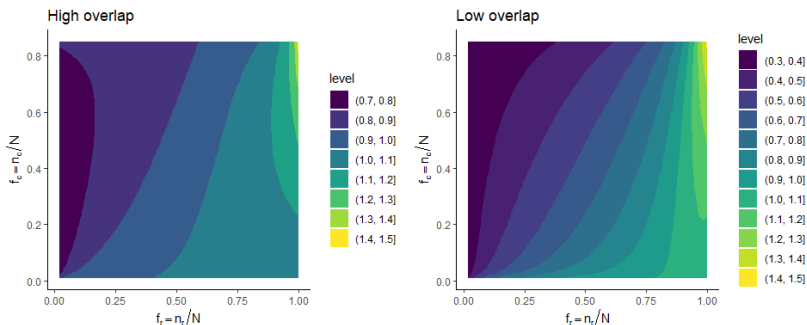


Figure: Contour plots of the ratio of SE depending on sampling fractions (f_r, f_c)

Ratio of $SE(\hat{\mu}^{\text{LR}})$ to $SE(\hat{\mu}^{\text{CLW}})$

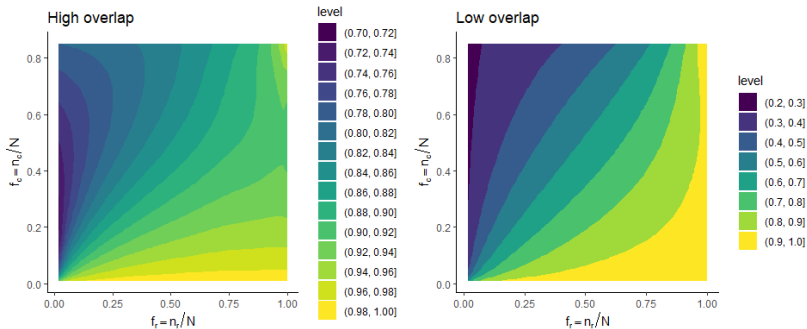


Figure: Contour plots of the ratio of SE depending on sampling fractions (f_r, f_c)

Outline

Survey statistics: from probability samples to data integration

Modeling response propensity to a nonprobability survey

Theoretical properties of the estimators

Simulation study to compare estimates by different methods

Discussion and next steps

References

Simulations scenarios

The purpose of simulations is to illustrate effects of sample fractions, sizes and overlap (“high” and “low”) between S_r and S_c on estimates $\hat{\eta} = (\hat{\mu}, \hat{\phi})$.

Sample fraction	PPS sample S_r $\text{logit}(M_{ri}) = 1.0 + \phi_{r1} * x_i$	Poisson sample S_c $\text{logit}(\pi_{ci}) = \phi_{c0} + 1.0 * x_i$
$f_r = .1, f_c \approx .1$	$n_r = 100$	$n_c \approx 100$
	$n_r = 600$	$n_c \approx 600$
$f_r = .01, f_c \approx .01$	$n_r = 100$	$n_c \approx 100$
	$n_r = 600$	$n_c \approx 600$
$f_r = .1, f_c \approx .01$ $f_r = .01, f_c \approx .1$	$n_r = 1,000$	$n_c \approx 100$
	$n_r = 100$	$n_c \approx 1,000$

Covariates: $x_i \sim N(0, 1)$

$f_c \approx .1, \phi_{c0} = -2.5$ and $f_c \approx .01, \phi_{c0} = -5.0$

High overlap: $\phi_{r1} = 1.0$, Low overlap: $\phi_{r1} = -1.0$

Outcome variable on S_c : $y_i \sim N(1 + x_i, 1.5^2)$

S_c and S_r are sampled 500 times from a simulated population.

Propensity parameter $\hat{\phi}_1$ (true value is $\phi_1 = 1.0$) for sample fractions $(f_c, f_r) = (0.01, 0.1)$ and $(0.1, 0.01)$, population size $N = 10,000$

	High overlap					Low overlap				
	$\hat{\phi}_1$	SD	\widehat{SE}	SE	95%CI	$\hat{\phi}_1$	SD	\widehat{SE}	SE	95%CI
$n_c = 100, n_r = 1,000$										
ILR	1.00	0.12	0.11	0.11	0.95	1.01	0.14	0.13	0.13	0.95
PILR	1.00	0.12	0.12	0.11	0.95	1.03	0.16	0.15	0.18	0.94
CLW	1.01	0.12	0.12	0.12	0.95	1.05	0.19	0.16	0.23	0.93
$n_c = 1,000, n_r = 100$										
ILR	1.00	0.14	0.14	0.15	0.94	1.01	0.15	0.15	0.15	0.95
PILR	1.01	0.16	0.16	0.16	0.95	1.05	0.26	0.25	0.29	0.94
CLW	1.05	0.23	0.22	0.22	0.95	1.29	0.62	0.51	0.68	0.92

SD - SD of $\hat{\phi}_1$ over simulations

\widehat{SE} - estimated SE averaged over simulations.

SE - calculated SE for population.

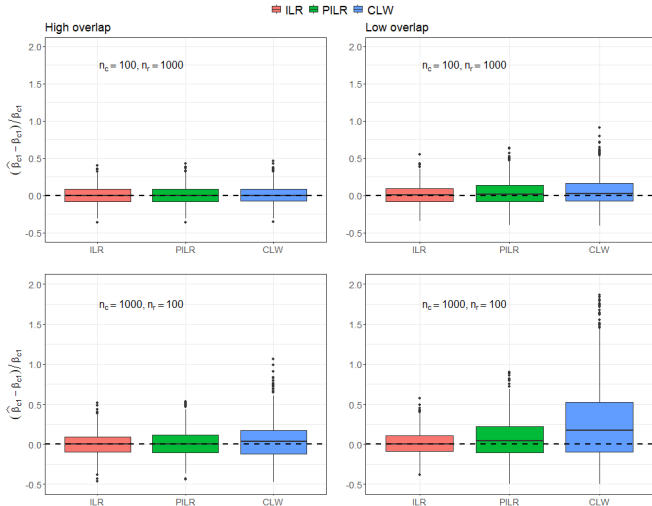


Figure: Relative residuals of the estimated propensity parameters $\hat{\phi}_1$ for sample fractions (f_c, f_r) equal to $(0.01, 0.1)$ (upper row) and to $(0.1, 0.01)$ (lower row) over the simulations

Estimate of population mean $\hat{\mu}$ (true value is $\mu = 1.0$) for sample fractions $(f_c, f_r) = (0.01, 0.1)$ and $(0.1, 0.01)$, population size $N = 10,000$

	High overlap					Low overlap				
	$\hat{\mu}$	SD	\widehat{SE}	SE	95%CI	$\hat{\mu}$	SD	\widehat{SE}	SE	95%CI
$n_c = 100, n_r = 1,000$										
ILR	1.07	0.30	0.29	0.31	0.87	1.03	0.32	0.28	0.31	0.88
PILR	1.07	0.30	0.30	0.32	0.88	1.00	0.35	0.34	0.36	0.90
CLW	1.07	0.30	0.31	0.32	0.88	0.98	0.37	0.37	0.39	0.91
$n_c = 1,000, n_r = 100$										
ILR	1.02	0.14	0.14	0.14	0.94	1.01	0.13	0.13	0.13	0.95
PILR	1.00	0.17	0.18	0.16	0.96	0.96	0.24	0.27	0.26	0.96
CLW	0.97	0.23	0.29	0.21	0.95	0.75	0.54	Inf	0.60	0.95

SD - SD of $\hat{\mu}$ over simulations

\widehat{SE} - estimated SE averaged over simulations.

SE - calculated SE for population.

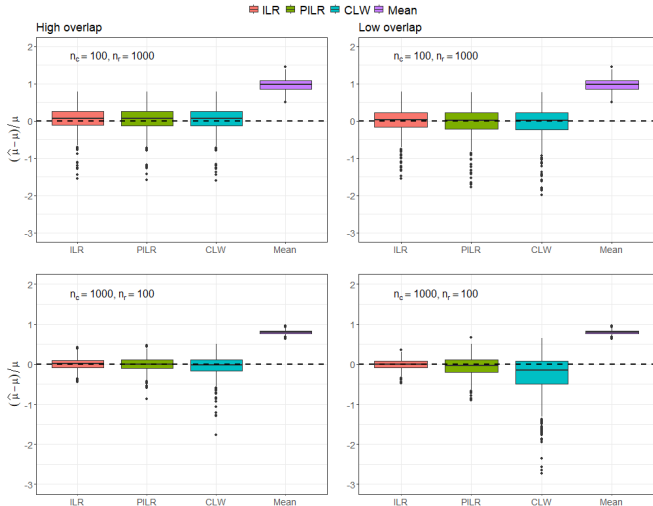


Figure: Relative residuals of the estimated population mean $\hat{\mu}$ for sampling fractions (f_c, f_r) equal to $(0.01, 0.1)$ (upper row) and to $(0.1, 0.01)$ (lower row) over the simulations

Outline

Survey statistics: from probability samples to data integration

Modeling response propensity to a nonprobability survey

Theoretical properties of the estimators

Simulation study to compare estimates by different methods

Discussion and next steps

References

Comparison of the three methods for estimation of π_c

- ▶ All three methods are asymptotically consistent and similarly efficient under favorable conditions.
- ▶ ILR and PILR use “stacked” sets and employ CRISP formula $\pi_z = \pi_c / (\pi_c + \pi_r)$.
CLW models *unobserved* indicator of S_c on the population.
- ▶ ILR uses *likelihood*; PILR and CLW use *pseudo-likelihood*.
- ▶ ILR requires sampling probabilities π_r for convenience sample units. If they cannot be inferred from the probability sample design, PILR should be used.
- ▶ ILR is more efficient especially in case of low overlap in covariate domains, which is important for multivariate models.

What's next ...

- ▶ Explore applying ILR to integrate BLS probability surveys with administrative data. This may substantially improve estimates in small domains (SAE) and for rare populations.
- ▶ The presented methods use CRISP to define composite dependence of likelihood on model parameters. The CRISP formula is general. Bayesian estimation has been implemented by Savitsky et al. (2022). The methodology can be extended to popular ML algorithms: CART, LASSO, Random Forest, Neural Networks, etc.

Outline

Survey statistics: from probability samples to data integration

Modeling response propensity to a nonprobability survey





Theoretical properties of the estimators

Simulation study to compare estimates by different methods


Discussion and next steps

References

References I

-  Chen, Y., P. Li, and C. Wu (2020). “Doubly Robust Inference With Nonprobability Survey Samples”. In: *Journal of the American Statistical Association* 115.532, pp. 2011–2021. eprint: <https://doi.org/10.1080/01621459.2019.1677241>.
-  Elliott, M. R. (2009). “Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights”. In: *Survey Practice* 2 (6), pp. 813–845.
-  Elliott, M. R. and R. Valliant (2017). “Inference for Nonprobability Samples”. In: *Statistical Science* 32.2, pp. 249–264.
-  Savitsky, T. et al. (2022). “Methods for Combining Probability and Nonprobability Samples Under Unknown Overlaps”. In: eprint: <https://arxiv.org/abs/2208.14541>.

References II

-  Wang, L., Y. Li, and R. Valliant (2021). “Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts”. In: *Stat Med.* 40.4, pp. 5237–5250.

Thank you for attending FCSM!

Vladislav Beresovsky

Beresovsky.Vladislav@bls.gov