

nonprobsvy – An R package for statistical inference with non-probability samples

Łukasz Chrostowski⁽¹⁾, Maciej Beręsewicz^(2,3)

(1) Adam Mickiewicz University in Poznań, Poland

(2) Poznań University of Economics and Business, Poland

(3) Statistical Office in Poznań, Poland

RSS 2023 International Conference, 07.09.2023

Outline

- 1 Introduction
- 2 Theory
- 3 `nonprobsvy` package and examples
- 4 Conclusions

Outline

- 1 Introduction
- 2 Theory
- 3 nonprobsvy package and examples
- 4 Conclusions

Why create another package?

- **GJRM** (Marra et al. 2017) – generalized Heckman's sample selection models,
- **NonProbEst** (Rueda, Ferri-García & Castro 2020) – model-assisted, model-based, model-calibrated and propensity score adjustment estimators among others,
- **DoubleML**, **CBPS**, etc. (Bach et al. 2022, Imai & Ratkovic 2014) – covariate balancing propensity score weighting,
- **WeightIt** – various methods to reweight treatment and control groups,

However, none of these packages implements *state-of-the-art* methods recently proposed in the literature, in particular doubly robust estimators, analytical variance estimation or bias minimisation. i.e. Yang Kim, & Song (2020 JRSSB), Kim, Park, Chen & Wu, C. (2021 JRSSA) or Chen, Li, & Wu (2020 JASA).

Outline

- 1 Introduction
- 2 Theory
- 3 nonprobsvy package and examples
- 4 Conclusions

Basic idea

Sample		Sampling weight π^{-1}	Covariate \mathbf{x}	Study variable \mathbf{y}
Non-probability sample (A)	1	?	✓	✓
	\vdots	\vdots	\vdots	\vdots
	n_A	?	✓	✓
Probability sample (B)	$n_A + 1$	✓	✓	?
	\vdots	\vdots	\vdots	\vdots
	$n_A + n_B$	✓	✓	?

Notation

Notation	Meaning
N	population size
\mathcal{U}	finite population with N units
y	response variable
\mathbf{x}	auxiliary variables
p	number of auxiliary variables
π_i^A	unknown probability inclusion into non-probability sample
π_i^B	probability of inclusion into probability sample
$d_i^B = 1/\pi_i^B, w_i^B$	design and calibrated weight
$\left\{ (\mathbf{x}_i, y_i, \delta_i^A), i \in \mathcal{S}_A \right\}$	dataset from non-probability sample A
$\left\{ (\mathbf{x}_i, d_i^A), i \in \mathcal{S}_B \right\}$	dataset from probability sample B
n_A	size of \mathcal{S}_A dataset
n_B	size of \mathcal{S}_B dataset
μ_y	mean of population for response variable
R_i^A	indicator function for non-probability sample
R_i^B	indicator function for probability sample

Inverse probability weighting (MLE)

Let $P(R_i = 1 \mid \mathbf{x}_i) = \pi(\mathbf{x}_i, \boldsymbol{\theta}_0)$. The maximum likelihood estimator is computed as $\hat{\pi}_i^A = \pi(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_0)$, where $\hat{\boldsymbol{\theta}}_0$ is the maximizer of the following pseudo-log-likelihood function:

$$l^*(\boldsymbol{\theta}) = \sum_{i \in S_A} \log \left\{ \frac{\pi(\mathbf{x}_i, \boldsymbol{\theta})}{1 - \pi(\mathbf{x}_i, \boldsymbol{\theta})} \right\} + \sum_{i \in S_B} d_i^B \log\{1 - \pi(\mathbf{x}_i, \boldsymbol{\theta})\} \quad (1)$$

Then, gradient (for logistic regression) is given by

$$U(\boldsymbol{\theta}) = \sum_{i \in S_A} \mathbf{x}_i - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\theta}) \mathbf{x}_i$$

Inverse probability weighting (GEE)

The pseudo score equations $U(\theta) = \mathbf{0}$ derived from Maximum Likelihood Estimation methods may be replaced by a system of general estimating equations. Let $h(\mathbf{x})$ be the smooth function and

$$U(\theta) = \sum_{i \in S_A} h(\mathbf{x}_i, \theta) - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \theta) h(\mathbf{x}_i, \theta). \quad (2)$$

Under $h(\mathbf{x}_i) = \mathbf{x}_i$ and logistic model for propensity score, equation (2) looks like distorted version of the score equation from MLE method. Population mean estimator has following form:

$$\hat{\mu}_{IPW} = \frac{1}{\hat{N}^A} \sum_{i \in S_A} \frac{y_i}{\hat{\pi}_i^A} \quad (3)$$

Mass imputation

This method is based on parametric model on sample S_A in the form of

$$\mathbb{E}(y_i | \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta}) \quad (4)$$

for some unknown $\boldsymbol{\beta}$ and known function $m(\cdot)$.

Based on this approach we can obtain the population mean estimator:

$$\hat{\mu}_{\text{MI}} = \frac{1}{\hat{N}^{\text{B}}} \sum_{i \in \mathcal{S}_B} d_i^{\text{B}} m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \quad (5)$$

Alternatively one can use nearest neighbours algorithm or predictive mean matching imputation estimator.

Doubly robust estimators

This method involves using units from both probability and non-probability samples. In particular, estimator of the population mean is as follows

$$\hat{\mu}_{\text{DR}} = \frac{1}{\hat{N}^A} \sum_{i \in \mathcal{S}_A} d_i^A \left\{ y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \right\} + \frac{1}{\hat{N}^B} \sum_{i \in \mathcal{S}_B} d_i^B m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}),$$

where $d_i^A = \pi(\mathbf{x}_i, \boldsymbol{\theta})^{-1}$, $\hat{N}^A = \sum_{i \in \mathcal{S}_A} d_i^A$ and $\hat{N}^B = \sum_{i \in \mathcal{S}_B} d_i^B$.

Variable selection

Let

$$U(\boldsymbol{\theta}, \boldsymbol{\beta}) = \left(\frac{\sum_{i \in S_A} \mathbf{h}(\mathbf{x}_i, \boldsymbol{\theta}) - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\theta}) \mathbf{h}(\mathbf{x}_i, \boldsymbol{\theta})}{\sum_{i \in S_A} \{y_i - m(\mathbf{x}_i, \boldsymbol{\beta})\} \mathbf{x}_i} \right) \quad (6)$$

be the joint estimating function for $(\boldsymbol{\theta}, \boldsymbol{\beta})$. When p is large, we consider the penalized estimating functions for $(\boldsymbol{\theta}, \boldsymbol{\beta})$ as

$$U^p(\boldsymbol{\theta}, \boldsymbol{\beta}) = U(\boldsymbol{\theta}, \boldsymbol{\beta}) - \begin{pmatrix} q_{\lambda_{\boldsymbol{\theta}}}(|\boldsymbol{\theta}|) \operatorname{sgn}(\boldsymbol{\theta}) \\ q_{\lambda_{\boldsymbol{\beta}}}(|\boldsymbol{\beta}|) \operatorname{sgn}(\boldsymbol{\beta}) \end{pmatrix},$$

where $q_{\lambda_{\boldsymbol{\theta}}}$ and $q_{\lambda_{\boldsymbol{\beta}}}$ are some smooth functions. We let $q_{\lambda}(x) = \frac{\partial p_{\lambda}}{\partial x}$, where p_{λ} is some penalization function.

Selection of relevant tuning parameters are based on minimizing covariate balancing loss function.

Outline

- 1 Introduction
- 2 Theory
- 3 nonprobsvy package and examples**
- 4 Conclusions

The nonprobsvy package

The `nonprobsvy` package allows the following approaches when one have access only to population totals/means or probability sample (we support `survey::svydesign` objects)

- IPW: MLE (with different optimizers), GEE with two $h()$ functions,
- MI: model-based (GLM) or NN imputation,
- DR: with different methods of IPW, MI estimators, bias minimization technique,
- variable selection: SCAD, LASSO, MCP,
- GLM: gaussian, binomial (logit, probit, cloglog), Poisson,
- (initial implementation of) samples overlap.

Package can be installed from github github.com/ncn-foreigners/nonprobsvy.

The nonprobsvy package – ncn-foreigners/software-tutorials

Basic use cases of the **nonprobsvy** package

AUTHOR

Maciej Beręsewicz, Łukasz Chrostowski

1 Introduction

This tutorial shows basic usage of the [nonprobsvy](#) package developed in this project based on example from the paper

Yang, S., Kim, J. K., & Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 82(2), 445.

All technical details regarding implementation can be found [here](#)

1.1 Basic information regarding the package

Table of contents

[1 Introduction](#)

1.1 Basic information regarding the package

1.2 Install and load the required packages

1.3 Basic setup for examples

2 Inverse probability weighting estimator



3 Mass imputation estimator

4 Doubly robust estimator


5 Summary

6 Literature

The nonprobsvy package – ncn-foreigners/nonprobsvy-book

Modern inference methods
for non-probability
samples with R  



- 1 Welcome!
 - 2 Introduction and Overview
 - 3 Inverse probability weighting
 - 4 Mass imputation
 - 5 Doubly robust methods
 - 6 [Techniques of variables selection for high-dimensional data](#)
 - 7 Summary
 - 8 References
- Appendices 
- A Appendices
- B Nomenclature

6 Techniques of variables selection for high-dimensional data

6.1 Motivation

In the presence of high-dimensional data, variable selection is important, because it can reduce variability in estimation resulting from using irrelevant variables for model building. There is a considerable body of literature on variable selection, but little about techniques for data integration that can successfully recognize the strengths and the limitations of each source of data. The selection of variables is the basis of a two-step approach to estimation, where in first one we select important variables and in the second one re-estimate model. For the first step it is proposed penalized logistic regression model for propensity score estimation (Yang et al, 2020), but we expand this approach on complementary log-log and probit models. For a mass imputation based on a parametric model it is considered penalized OLS (Ordinary least squared) method. It is worth mentioning that Yang, Kim and Rui (2020), in their article on this topic, used the SCAD (Smoothly Clipped Absolute Deviation) penalization, but one can use LASSO (Least Absolute Shrinkage and Selection Operator) and MCP (Minimax Concave Penalty) techniques as well, what will be considered in the next subsection.

6.2 Existed techniques

Let $U(\boldsymbol{\theta}, \boldsymbol{\beta})$ be the joint estimating function for $(\boldsymbol{\theta}, \boldsymbol{\beta})$. When p is large, we consider the penalized estimating functions for $(\boldsymbol{\theta}, \boldsymbol{\beta})$ as

$$U^p(\boldsymbol{\theta}, \boldsymbol{\beta}) = U(\boldsymbol{\theta}, \boldsymbol{\beta}) - \begin{pmatrix} q_{\lambda_\theta}(|\boldsymbol{\theta}|) \operatorname{sgn}(\boldsymbol{\theta}) \\ q_{\lambda_\beta}(|\boldsymbol{\beta}|) \operatorname{sgn}(\boldsymbol{\beta}) \end{pmatrix},$$

Table of contents

- [6.1 Motivation](#)
- [6.2 Existed techniques](#)
- [6.3 Solution](#)

 Report an issue

The nonprobsvy package - example usage (IPW)

```
est_ipw <- nonprob(  
  selection = ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8,  
  target = ~ Y_21,  
  data = sample_B2,  
  svydesign = sample_A_svy_cal,  
  family_outcome = "binomial",  
  method_selection = "logit",  
  control_selection = controlSel(penalty = "SCAD",  
                                est_method_sel = "mle"),  
  control_inference = controlInf(vars_selection = TRUE))
```

```
-----  
Estimated population mean: 0.607 with overall std.err of: 0.03408  
And std.err for nonprobability and probability samples being respectively:  
0.01164 and 0.03203
```

Based on: Inverse probability weighted method

95% Confidence interval for population mean:

	lower_bound	upper_bound
Y_21	0.5401836	0.6737786

For a population of estimate size: 10350.43
Obtained on a nonprobability sample of size: 1927
With an auxiliary probability sample of size: 493

The nonprobsvy package - example usage (MI)

```
est_mi <- nonprob(  
  outcome = Y_21 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8,  
  data = sample_B2,  
  svydesign = sample_A_svy_cal,  
  family_outcome = "binomial",  
  method_outcome = "glm",  
  control_outcome = controlOut(penalty = "SCAD"),  
  control_inference = controlInf(vars_selection = TRUE))
```

Estimated population mean: 0.6389 with overall std.err of: 0.01092
And std.err for nonprobability and probability samples being respectively:
0.009853 and 0.004709

Based on: Mass Imputation method

95% Confidence interval for population mean:

	lower_bound	upper_bound
Y_21	0.6174754	0.6602813

For a population of estimate size: 10000
Obtained on a nonprobability sample of size: 1927
With an auxiliary probability sample of size: 493

The nonprobsvy package - example usage (DR)

```
est_dr <- nonprob(  
  outcome = Y_21 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8,  
  selection = ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8,  
  data = sample_B2,  
  svydesign = sample_A_svy_cal,  
  family_outcome = "binomial",  
  method_outcome = "glm",  
  method_selection = "logit",  
  control_outcome = controlOut(penalty = "SCAD"),  
  control_selection = controlSel(penalty = "SCAD",  
                                est_method_sel = "mle"),  
  control_inference = controlInf(vars_selection = TRUE))
```

```
-----  
Estimated population mean: 0.6388 with overall std.err of: 0.008773  
And std.err for nonprobability and probability samples being respectively:  
0.007413 and 0.004691
```

Based on: Doubly-Robust method

```
95% Confidence interval for population mean:  
      lower_bound upper_bound  
Y_21  0.6216019   0.6559903
```

For a population of estimate size: 10388.52
Obtained on a nonprobability sample of size: 1927
With an auxiliary probability sample of size: 493

Outline

- 1 Introduction
- 2 Theory
- 3 nonprobsvy package and examples
- 4 Conclusions

Conclusions

- Package is under development so comments are welcome!
- Further plans: predictive mean matching, empirical likelihood, multiply robust estimators, not missing at random, GAM and other models, performance improvements.



Referecens I

- Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2022). DoubleML: an object-oriented implementation of double machine learning in Python. *The Journal of Machine Learning Research*, 23(1), 2469-2474.
- Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.
- Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1), 243-263.
- Kim, J. K., Park, S., Chen, Y., & Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3), 941-963.
- Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., & McGovern, M. E. (2017). A simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, 112(518), 484-496.
- Rueda, M., Ferri-García, R., & Castro, L. (2020). The R package NonProbEst for estimation in non-probability surveys. *R J*, 12, 406-418.
- Yang, S., Kim, J. K., & Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 82(2), 445.

Referecens II

- Yang, S., Kim, J. K., & Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation.