The nonprobsvy package

Maciej Beręsewicz
Department of Statistics, Poznań University of Economics and Business
Centre for the Methodology of Population Studies, Statistical Office in Poznań
Łukasz Chrostowski
Faculty of Mathematics and Informatics, Adam Mickiewicz University

Contents

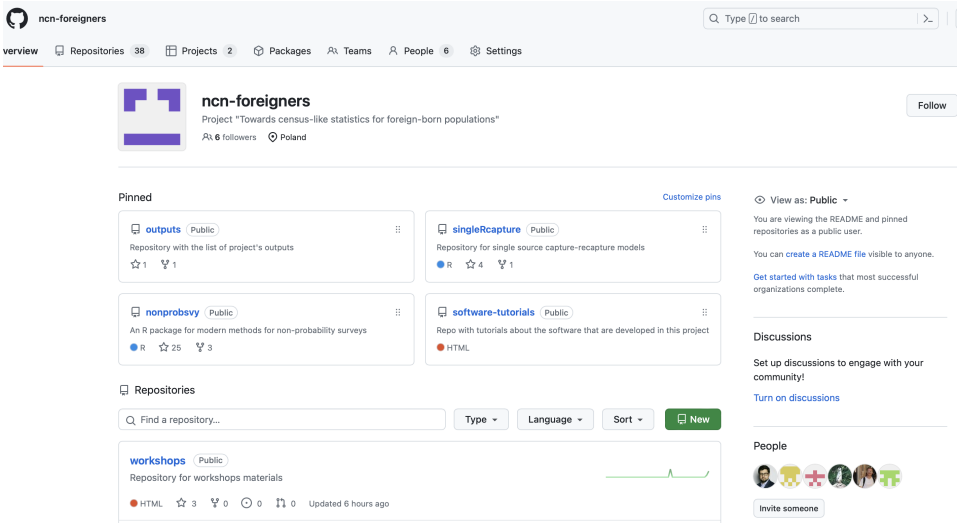
Contents

About funding

- Works on this package was funded by the National Science Centre grant entitled: *Towards census-like statistics for foreign-born populations – quality, data integration and estimation* (no. 2020/39/B/HS4/00941).

- The main objective of the project is to develop methods for estimating the size and characteristics of the foreign population in Poland based on available data sources.

About funding – github



Why another package for non-probability samples?

- It should be noted that there are some packages that can be used for non-probability samples, such as `NonProbEst`, `WeightIt` or `GJRM`.
- These packages are limited in terms of the approaches that can be employed and the variance estimation that can be carried out.
- None of these packages are integrated with the `survey` package.
- There is a lack of implementation of the current solutions presented in the literature.
- The motivation for this project is as follows: The development of a tool that enables the consistent application of different estimation techniques.
- The package is on CRAN and under development so we encourage testing and commenting and tracking on github.

What is unique about the nonprobsvy package?

- It provides an easy to use nonprob function that mimics existing R functions (e.g. uses formulas).
- It has a full integration with the survey package.
- It implements both analytical and bootstrap variance estimators recently proposed in the literature.
- It extends state-of-the-art methods in various ways.

Selected literature

JASA

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 2011–2021: Theory and Methods
https://doi.org/10.1080/01621459.2019.1677241

Doubly Robust Inference With Nonprobability Survey Samples

Yilin Chen, Pengfei Li, and Changbao Wu

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

ABSTRACT

We establish a general framework for statistical inferences with nonprobability survey samples when relevant auxiliary information is available from a probability survey sample. We develop a rigorous procedure for estimating the propensity scores for units in the nonprobability sample, and construct doubly robust estimators for the finite population mean. Variance estimation is discussed under the proposed framework. Results from simulation studies show the robustness and the efficiency of our proposed estimators as compared to existing methods. The proposed method is used to analyze a nonprobability survey sample collected by the Pew Research Center with auxiliary information from the Behavioral Risk Factor Surveillance System and the Current Population Survey. Our results illustrate a general approach to inference with nonprobability samples and highlight the importance and usefulness of auxiliary information from probability survey samples. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2018
Accepted September 2019

KEYWORDS
Design-based inference; Inclusion probability; Missing at random; Propensity score; Regression modeling; Variance estimation

JRSSB

J. R. Statist. Soc. B (2020)

Doubly robust inference when combining probability and non-probability samples with high dimensional data

Shu Yang,

North Carolina State University, Raleigh, USA

Jae Kwang Kim,

Iowa State University, Ames, USA

and Rui Song

North Carolina State University, Raleigh, USA

JRSSA

Received: 8 January 2020 | Accepted: 20 March 2021

DOI: 10.1111/rssa.12696

ORIGINAL ARTICLE

Combining non-probability and probability survey samples through mass imputation

Jae Kwang Kim[1] | **Seho Park**[2] | **Yilin Chen**[3] | **Changbao Wu**[3]

[1]Department of Statistics, Iowa State University, Ames, IA 50011, USA

[2]Department of Biostatistics, Indiana University School of Medicine,

Abstract

Analysis of non-probability survey samples requires auxiliary information at the population level. Such information

Survey Methodology

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 283-311
Statistics Canada, Catalogue No. 12-001-X

Statistical inference with non-probability survey samples

Changbao Wu[1]

Abstract

We provide a critical review and some extended discussions on theoretical and practical issues with analysis of non-probability survey samples. We attempt to present rigorous inferential frameworks and valid statistical procedures under commonly used assumptions, and address issues on the justification and verification of assumptions in practical applications. Some current methodological developments are showcased, and problems which require further investigation are mentioned. While the focus of the paper is on non-probability samples, the essential role of probability survey samples with rich and relevant information on auxiliary variables is highlighted.

Key Words: Auxiliary information; Bootstrap variance estimator; Calibration method; Doubly robust estimator; Estimating equations; Inverse probability weighting; Model-based prediction; Poststratification; Pseudo likelihood; Propensity score; Quota survey; Sensitivity analysis; Variance estimation.

Working papers

arXiv:2403.13750v1 [stat.ME] 20 Mar 2024

Data integration of non-probability and probability samples with predictive mean matching

Piotr Chlebicki[1], Łukasz Chrostowski [2], Maciej Beręsewicz [3]

Abstract

In this paper we study predictive mean matching mass imputation estimators to integrate data from probability and non-probability samples. We consider two approaches: matching predicted to observed ($\hat{y} - y$ matching) or predicted to predicted ($\hat{y} - \hat{y}$ matching) values. We prove the consistency of two semi-parametric mass imputation estimators based on these approaches and derive their variance and estimators of variance. Our approach can be employed with non-parametric regression techniques, such as kernel regression, and the analytical expression for variance can also be applied in nearest neighbour matching for non-probability samples. We conduct extensive simulation studies in order to compare the properties of this estimator with existing approaches, discuss the selection of k-nearest neighbours, and study the effects of model mis-specification. The paper finishes with empirical study in integration of job vacancy survey and vacancies submitted to public employment offices (admin and online data). Open source software is available for the proposed approaches.

Working papers

arXiv:2403.09726v1 [stat.ME] 12 Mar 2024

Inference for non-probability samples using the calibration approach for quantiles

Maciej Beręsewicz[1], Marcin Szymkowiak[2]

Abstract

Non-probability survey samples are examples of data sources that have become increasingly popular in recent years, also in official statistics. However, statistical inference based on non-probability samples is much more difficult because they are biased and are not representative of the target population [75]. In this paper we consider a method of joint calibration for totals [55] and quantiles [59] and use the proposed approach to extend existing inference methods for non-probability samples, such as inverse probability weighting, mass imputation and doubly robust estimators. By including quantile information in the estimation process non-linear relationships between the target and auxiliary variables can be approximated the way it is done in step-wise (constant) regression. Our simulation study has demonstrated that the estimators in question are more robust against model mis-specification and,

Literature (selected)

- Chen, Yilin, Pengfei Li, and Changbao Wu. 2020. "Doubly Robust Inference With Nonprobability Survey Samples." Journal of the American Statistical Association 115 (532): 2011–21. https://doi.org/10.1080/01621459.2019.1677241.
- Kim, Jae Kwang, Seho Park, Yilin Chen, and Changbao Wu. 2021. "Combining Non-Probability and Probability Survey Samples Through Mass Imputation." Journal of the Royal Statistical Society Series A: Statistics in Society 184 (3): 941–63. https://doi.org/10.1111/rssa.12696
- Wu, Changbao. 2023. "Statistical Inference with Non-Probability Survey Samples." Survey Methodology 48 (2): 283–311. https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00002-eng.htm.
- Yang, Shu, Jae Kwang Kim, and Youngdeok Hwang. 2021. "Integration of Data from Probability Surveys and Big Found Data for Finite Population Inference Using Mass Imputation." Survey Methodology 47 (1): 29–58. https://www150.statcan.gc.ca/n1/p
001-x/2021001/article/00004-eng.htm.
- Yang, Shu, Jae Kwang Kim, and Rui Song. 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): 445–65. https://doi.org/10.1111/rssb.12354.

Contents

Basic setup

The package allows two approaches, assuming unit-level data are available from the non-probability sample:

- only population=level data are available (through some vector of totals, means, and population size),
- survey data is available from the reference probability sample (a `survey::svydesign` object can be specified).

Basic setup

Tabela 1: Two sample setting

Sample	ID	Sample weight $d = \pi^{-1}$	Covariates $\boldsymbol{x}$	Study variable y
Non-probability sample (S_A)	1	?	✓	✓
	⋮	?	⋮	⋮
	n_A	?	✓	✓
Probability sample (S_B)	1	✓	✓	?
	⋮	⋮	⋮	?
	n_B	✓	✓	?

Mass imputation estimator

Mass imputation based on regression imputation

```
nonprob(
  outcome = y ~ x1 + x2 + ... + xk,
  data = nonprob,
  pop_totals = c(`(Intercept)`= N,
                 x1 = tau_x1,
                 x2 = tau_x2,
                 ...,
                 xk = tau_xk),
  method_outcome = "glm",
  family_outcome = "gaussian"
)
```

Mass imputation based on regression imputation

```
nonprob(
  outcome = y ~ x1 + x2 + ... + xk,
  data = nonprob,
  svydesign = prob,
  method_outcome = "glm",
  family_outcome = "gaussian"
)
```

Rysunek 1: Mass imputation when population (left) or unit (right) data is available

Inverse probability weighting estimator

Inverse probability weighting

```
nonprob(
  selection =  ~ x1 + x2 + ... + xk,
  target = ~ y,
  data = nonprob,
  pop_totals = c(`(Intercept)` = N,
                 x1 = tau_x1,
                 x2 = tau_x2,
                 ...,
                 xk = tau_xk),
  method_selection = "logit"
)
```

Inverse probability weighting

```
nonprob(
  selection =  ~ x1 + x2 + ... + xk,
  target = ~ y,
  data = nonprob,
  svydesign = prob,
  method_selection = "logit"
)
```

Rysunek 2: IPW when population (left) or unit (right) data is available

Doubly robust estimator

Doubly robust estimator

```
nonprob(
  selection = ~ x1 + x2 + ... + xk,
  outcome = y ~ x1 + x2 + …, + xk,
  pop_totals = c(`(Intercept)` = N,
                 x1 = tau_x1,
                 x2 = tau_x2,
                 ...,
                 xk = tau_xk),
  svydesign = prob,
  method_outcome = "glm",
  family_outcome = "gaussian"
)
```

Doubly robust estimator

```
nonprob(
  selection = ~ x1 + x2 + ... + xk,
  outcome = y ~ x1 + x2 + ... + xk,
  data = nonprob,
  svydesign = prob,
  method_outcome = "glm",
  family_outcome = "gaussian"
)
```

Rysunek 3: DR when population (left) or unit (right) data is available

Example output

```
## mass imputation
result_mi <- nonprob(
  outcome = y1 ~ x1 + x2,
  data = nonprob_df,
  svydesign = sample_prob
)

## IPW
result_ipw <- nonprob(
  selection = ~ x2,
  target = ~y1,
  data = nonprob_df,
  svydesign = sample_prob)
```

Example output

```
#> -------------------------
#> Estimated population mean: 2.95 with overall std.err of: 0.04203
#> And std.err for nonprobability and probability samples being respectively:
#> 0.001227 and 0.04201
#>
#> 95% Confidence inverval for popualtion mean:
#>    lower_bound upper_bound
#> y1    2.867433    3.032186
#>
#>
#> Based on: Mass Imputation method
#> For a population of estimate size: 1e+06
#> Obtained on a nonprobability sample of size: 693011
#> With an auxiliary probability sample of size: 1000
#> -------------------------
#>
#> Regression coefficients:
#> ----------------------
#> For glm regression on outcome variable:
#>             Estimate Std. Error z value P(>|z|)
#> (Intercept) 0.996282   0.002139   465.8  <2e-16 ***
#> x1          1.001931   0.001200   835.3  <2e-16 ***
#> x2          0.999125   0.001098   910.2  <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example output

```
#> -------------------------
#> Estimated population mean: 2.925 with overall std.err of: 0.05
#> And std.err for nonprobability and probability samples being respectively:
#> 0.001586 and 0.04997
#>
#> 95% Confidence inverval for popualtion mean:
#>    lower_bound upper_bound
#> y1     2.82679    3.022776
#> .....
#> .....
#> .....
#> Weights:
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   1.000   1.071   1.313   1.479   1.798   2.647
#> -------------------------
#>
#> Covariate balance:
#> (Intercept)          x2
#>  25062.8473   -517.5862
#> -------------------------
```

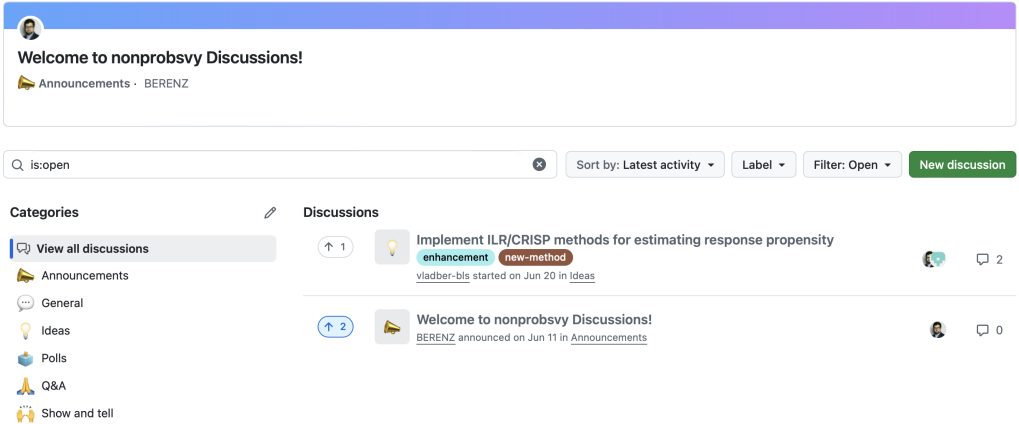
Contents

What is not (yet) implemented?

- Overlapping samples (in the development phase).
- Using replicated weights from the probability sample (in the development phase).
- Model calibration approach (e.g. using empirical likelihood, LASSO).
- GAM or other non-parametric methods for mass imputation /doubly robust estimators.
- Inference for quantiles (to be developed).
- Situations where target variable (Y) is observed in both sources (non-probability and probability sample, i.e. mixed-mode).
- Pseudo-population bootstrap for variance estimation.

If you have other ideas please let us know! ...and leave us a "star" on Github :)

Discussion


Welcome to nonprobsvy Discussions!
Announcements · BERENZ
is:open
Sort by: Latest activity
Label
Filter: Open
New discussion
Categories
View all discussions
Announcements
General
Ideas
Polls
Q&A
Show and tell
Discussions
1
Implement ILR/CRISP methods for estimating response propensity
enhancement
new-method
vladber-bls started on Jun 20 in Ideas
2
2
Welcome to nonprobsvy Discussions!
BERENZ announced on Jun 11 in Announcements
0

Thank you!