Workshop on the `nonprobsvy` package

Łukasz Chrostowski
Faculty of Mathematics and Computer Science, Adam Mickiewicz University

27.08.2024

Contents

Contents

About the workshop

- Part 1: theory (13:00 - 14:15)
- Part 2: practice and discussion session (14:45 - 16:00)

About funding

- 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

Single-source capture-recapture models

Overview

Capture-recapture type experiments are used to estimate the total population size in situations when observing only a part of such population is feasible. In recent years these types of experiments have seen more interest.

Single source models are distinct from other capture-recapture models because we cannot estimate the population size based on how many units were observed in two or three sources which is the standard approach.

Instead in single source models we utilize count data regression models on positive distributions (i.e. on counts greater than 0) where the dependent variable is the number of times a particular unit was observed in source data.

This package aims to implement already existing and introduce new methods of estimating population size from single source to simplify the research process.

Currently we've implemented most of the frequentist approaches used in literature such as:

- Zero truncated Poisson, geometric and negative binomial regression.

Links

View on CRAN

Browse source code

Report a bug

License

Full license

MIT + file LICENSE

Citation

Citing singleRcapture

Developers

Piotr Chlebicki
Author, maintainer

Maciej Beręsewicz
Author, contributor

Dev status

About funding

Calibration for quantiles (and totals)

jointCalib 0.1.2   Reference   Articles ▾   Changelog

Overview

Details

A small package for joint calibration of totals and quantiles (see Beręsewicz and Szymkowiak (2023) working paper for details). The package combines the following approaches:

- Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376–382.
- Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. Survey Methodology, 32(1), 37.
- Wu, C. (2005) Algorithms and R codes for the pseudo empirical likelihood method in survey sampling, Survey Methodology, 31(2), 239.
- Zhang, S., Han, P., and Wu, C. (2023) Calibration Techniques Encompassing Survey Sampling, Missing Data Analysis and Causal Inference, International Statistical Review 91, 165–192.

which allows to calibrate weights to known (or estimated) totals and quantiles jointly. As an backend for calibration sampling (sampling::calib), laeken (laeken::calibWeights) survey

Links

View on CRAN

Browse source code

Report a bug

License

GPL-3

Citation

Citing jointCalib

Developers

Maciej Beręsewicz
Author, maintainer [iD]

Dev status

R-CMD-check passing

CRAN 0.1.0

About funding

Blocking using ANN and graphs

blocking 0.1.0   Reference   Articles ▾   Changelog   Sea

Overview

Description 🔗

An R package that aims to block records for data deduplication and record linkage (a.k.a. entity resolution) based on approximate nearest neighbours algorithms (ANN) and graphs (via the `igraph` package).

Currently supports the following R packages that binds to specific ANN algorithms:

- rnndescent (default, very powerful, supports sparse matrices),
- RcppHNSW (powerful but does not support sparse matrices),
- RcppAnnoy,
- mlpack (see `mlpack::lsh` and `mlpack::knn`).

The package also supports integration with the reclin2 package via `blocking::pair_ann` function.

Funding

Links

Browse source code

Report a bug

License

GPL-3

Citation

Citing blocking

Developers

Maciej Beręsewicz
Author, maintainer ⓘD

Dev status

R-CMD-check passing
test-coverage passing
codecov 94%

Comparison

Tabela 1: Probability and Non-probability Samples

Factor	Probability Sample	Non-probability
Selection	Sampling scheme	Self-selection
Coverage	Usually good	Some groups are excluded
Bias	Usually smaller	Large or very large
Variance	Usually larger	Small or very small
Cost	Large or very large	Usually not large

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.

JASA

Taylor & Francis
Taylor & Francis Group

Check for updates

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.

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

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY | Series A Statistics in Society | A

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

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.

Working papers

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

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.

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!

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 2: 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	✓	✓	?

Estimation methods - consider the following cases
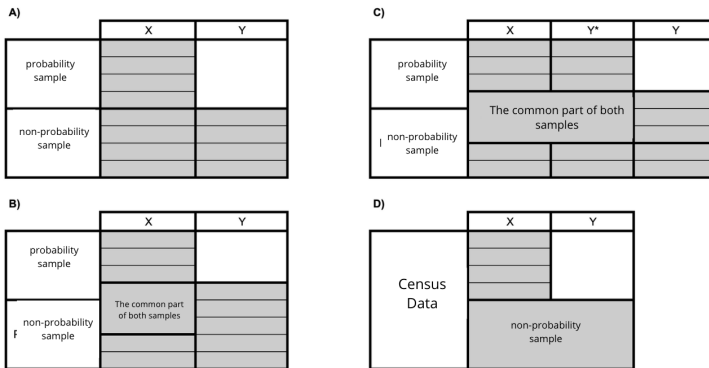
Rysunek 1: Four exemplary cases of data sources, where the goal is to estimate a selected characteristic of the variable Y. The variables $\boldsymbol{X}$ are common, and the variable Y^* is known as a proxy variable.

Elliott and Valliant (2017) distinguish between two approaches

- **Quasi-randomization** – where we construct *pseudo-weights* using a probability sample or known (or estimated) global values.

- **Model-based** – where we assume a certain model.

Basic setup

- Let $U = \{1, ..., N\}$ denote the target population consisting of N labelled units.
- Each unit i has an associated vector of auxiliary variables $\boldsymbol{x}_i$ (a realisation of the random vector $\boldsymbol{X}$ in the super-population) and the study variable y_i (a realisation of the random variable Y in the super-population).
- Let $\{(y_i, \boldsymbol{x}_i), i \in S_A\}$ be a dataset of a non-probability sample of size n_A and let $\{(\boldsymbol{x}_i, \pi_i), i \in S_B\}$ be a dataset of a probability sample of size n_B, where only information about variables $\boldsymbol{X}$ and inclusion probabilities π are available.
- Let $R_i = \mathrm{I}(i \in S_A)$ be an indicator of inclusion into non-probability sample. Each unit in the sample S_B has been assigned a design-based weight given by $d_i = 1/\pi_i$.
- Let $\mathbb{P}(R_i = 1 \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$ be the probability of inclusion into non-probability sample.
- There is no overlap between the two samples.

Implemented approaches

The package implements the following approaches:

- Inverse probability weighting (with possible calibration constraints).
- Mass imputation estimators (nearest neighbours, prediction via GLM and predictive mean matching).
- Doubly robust estimators.

Main nonprob function

```
nonprob(
  data,
  selection,
  outcome,
  target,
  svydesign,
  method_selection = c("logit", "cloglog", "probit"),
  method_outcome = c("glm", "nn", "pmm"),
  family_outcome = c("gaussian", "binomial", "poisson"),
  control_selection = controlSel(),
  control_outcome = controlOut(),
  control_inference = controlInf(),
  verbose = FALSE,
  se = TRUE,
  ...
)
```

Control parameters in nonprob

The nonprob function accepts 3 control functions, holding the values of the custom arguments:

- control_selection - allows to specify control parameters for propensity score model such as epsilon (tolerance for fitting algorithms), optimizer (maxLik or optim), optim_method (e.g. Newton-Raphson) and maxit (maximum number of iteration for fitting algorithms).
- control_outcome - allows to specify control parameters for mass impuation model such as penalty (penalty function, e.g. lasso) or predictive_match (type of predictive mean matching).
- control_inference - allows to specify control parameters for the inference process such as vars_selection, var_method, bias_correction (for bias minimization approach).

Simulation setup

We generate the data using the following recipe:

- $N = 10000$ – population size,
- $p = 50$ – number of X_p variables where $p - 1$ were generated from $N(0, 1)$ distribution,
- A – probability sample about $n_A \approx 500$,
- B – nonprobability sample of size about $n_B \approx 2000$,
- **Selection:**
  - $A \propto 0.25 + |X_{1i}| + 0.03|Y_i|$,
  - B selected according to the model

$$\mathrm{logit}(\pi_{B,i}) = \alpha^T X_i$$

  where $\alpha = (-2.1, 1, 1, 1, 0, \ldots, 0)^T$,

- **Y** generated according to the following model:

$$Y_i = 1 + \beta^T X_i + \epsilon,$$

where $\beta = (1, 0, 0, 1, 1, 1, 1, 0, \ldots, 0)^T$.

IPW – Assumptions

Formally, the assumptions of the *propensity score* method are as follows:

- The selection/inclusion variable R_i and the characteristic y_i under study are conditionally independent given the characteristics $\boldsymbol{x}_i$. In other words,
  $\pi_i = P(R_i = 1|y, \boldsymbol{x}_i) = P(R_i = 1|\boldsymbol{x}_i)$ – **In other words:** the selection is non-informative (in the literature: missing at random, ignorable).

- All units in the population have a non-zero probability of inclusion in the non-probability sample $(\pi_i > 0)$ – **In other words:** there are no coverage errors.

- The variables R_i and R_j are independent when considering $\boldsymbol{x}_i$ for $i \neq j$ – **In other words:** observations are independent (e.g., no duplicates or some unobserved variables $\boldsymbol{z}_i$).

Inverse probability weighting (MLE)

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

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

Then, gradient (for logistic regression)

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

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 $h(\boldsymbol{x})$ be the smooth function and

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

Under $h(\boldsymbol{x}_i) = \boldsymbol{x}_i^{\mathrm{T}}$ and logistic model for propensity score, Equation (6.1) looks like disorted version of the score equation from MLE method.

Inverse probability weighting

With the estimated propensity scores we can consider two approaches for population mean estimation, depending on whether population size is known or not.

$$
\hat{\mu}_{IPW1} = \frac{1}{N} \sum_{i \in S_A} \frac{y_i}{\hat{\pi}_i^A}
$$
$$
\hat{\mu}_{IPW2} = \frac{1}{\hat{N}^A} \sum_{i \in S_A} \frac{y_i}{\hat{\pi}_i^A},
$$
(3)

where $\hat{N}^A = \sum_{i \in S_A} \hat{d}_i^A$.

Model-Based Approach

- In the model-based approach, we assume that we are interested in $E(Y|X)$.
- We assume the model $E(Y|X, R = 1) = E(Y|X) = \mu(y_i|\mathbf{x}_i)$.
- We build a model on the non-probability sample (e.g., linear regression, logistic regression) and apply it to the entire population (or possibly a sample).

Mass imputation

The following approaches are implemented:

- Semi-parametric prediction approach (i.e. imputation are some model $m(.)$ predictions),
- Nearest neighbours imputation (i.e. imputation is based on looking for nearest neighbours based on $\boldsymbol{x}$),
- Predictive mean matching imputation (i.e. imputation is based on matching based on $\hat{y}$; two variants, non-parametric approach via loess).

The following estimator is used

$$\hat{\mu}_{MI} = \frac{1}{N} \sum_{i \in B} w_i \hat{y}_i. \tag{4}$$

Nearest Neighbour estimator

In the first approach, proposed by Rivers (2007), we perform mass imputation through so-called sample matching, which consists of the following steps:

1. For the non-probability sample $\mathcal{S}_{\mathrm{A}}$ and the probability sample $\mathcal{S}_{\mathrm{B}}$, we define a set of common characteristics $\mathbf{X}$.

2. Next, for each unit $i \in \mathcal{S}_{\mathrm{B}}$ we look for the nearest unit from the set $k \in \mathcal{S}_{\mathrm{A}}$ such that

$$d(\mathbf{x}_k, \mathbf{x}_i) = \|\mathbf{x}_k - \mathbf{x}_i\| \tag{5}$$

is minimized. We can use the Euclidean distance for this purpose. We assign the characteristic value y_k from the set $\mathcal{S}_{\mathrm{A}}$ to unit i.

3. After finding appropriate neighbors, we determine the estimator.

Predictive Mean Matching estimator

Considering that the assignment of the values y_i from the non-probability sample for units k from the probability sample should be based solely on one variable, the following approach is used:

1. We build a model $m(\boldsymbol{x}_i; \boldsymbol{\beta})$ on the non-probability sample $\mathcal{S}_A$, obtaining $\hat{y}_i = m(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}})$.
2. We apply the model $m(\boldsymbol{x}; \hat{\boldsymbol{\beta}})$ on the probability sample $\mathcal{S}_B$, obtaining $\hat{y}_k = m(\boldsymbol{x}_k; \hat{\boldsymbol{\beta}})$.
3. For each unit $k \in \mathcal{S}_B$, we find the nearest unit based on $d(\hat{y}_k, \hat{y}_i)$ and assign the value y_i.
4. Then, we determine the estimator.

Prediction Approach

- In the work by Kim, Park, Chen, and Wu (2021) **Combining Non-probability and Probability Survey Samples Through Mass Imputation**, (Journal of the Royal Statistical Society: Series A), a slightly different approach is proposed but based on solid theoretical foundations.
- Instead of searching for the nearest neighbor, only steps 1 and 2 from the previous slide are used.

Doubly robust estimators

- When the population size is known

$$\hat{\mu}_{\mathrm{DR1}} = \frac{1}{N} \sum_{i \in S_A} \frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{N} \sum_{i \in S_B} d_i^B \hat{m}_i \tag{6}$$

- When the population size is unknown

$$\hat{\mu}_{\mathrm{DR2}} = \frac{1}{\hat{N}^A} \sum_{i \in S_A} \frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B} \sum_{i \in S_B} d_i^B \hat{m}_i, \tag{7}$$

where $\hat{N}^A = \sum_{i \in S_A} \left(\hat{\pi}_i^A \right)^{-1}$ and $\hat{N}^B = \sum_{i \in S_B} d_i^B$.

Bias minimization approach

We define the bias as

$$
\begin{aligned}
\text{bias}\left(\hat{\mu}_{DR}\right) =&\mid \hat{\mu}_{DR} - \mu \mid \\
=& \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{R_i^A}{\pi_i^A \left(\mathbf{x}_i^{\mathrm{T}} \theta\right)} - 1 \right\} \left\{ y_i - m\left(\mathbf{x}_i^{\mathrm{T}} \beta\right) \right\} \\
&+ \frac{1}{N} \sum_{i=1}^{N} \left(R_i^B d_i^B - 1\right) m\left(\mathbf{x}_i^{\mathrm{T}} \beta\right)
\end{aligned}
$$

what leads to

$$
\mathrm{U}\left(\theta, \beta\right) = \left(\begin{array}{c} \sum_{i=1}^{N} R_i^A \left\{ \frac{1}{\pi(\mathbf{x}_i, \theta)} - 1 \right\} \left\{ y_i - m\left(\mathbf{x}_i, \beta\right) \right\} \mathbf{x}_i \\ \sum_{i=1}^{N} \frac{R_i^A}{\pi(\mathbf{x}_i, \theta)} \frac{\partial m(\mathbf{x}_i, \beta)}{\partial \beta} - \sum_{i \in \mathcal{S}_{\mathrm{B}}} d_i^{\mathrm{B}} \frac{\partial m(\mathbf{x}_i, \beta)}{\partial \beta} \end{array} \right)
$$

Variable selection

Let $U(\theta, \beta)$ be the join estimating function for (θ, β). When p is large, we consider the penalized estimating functions for (θ, β) as

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

where q_{λ_θ} and q_{λ_β} 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.

We have implemented variable selection for all proposed estimators using LASSO, SCAD or MCP penalty.

Variance estimators

- analytical approach
  1. asymptotic estimation
  2. linearised form
- bootstrap approach
  1. for the non-probability sample $\mathcal{S}_{\mathrm{A}}$, draw a simple sample $\mathcal{S}_{\mathrm{A}}^*$ with replacement, with a size of n_A,
  2. for the probability sample $\mathcal{S}_{\mathrm{B}}$, draw a sample $\mathcal{S}_{\mathrm{B}}$ with replacement, with probability $1/d_i^B$, and a size of n_B

The nonprobsvy package

Summarising the nonprobsvy package implements the following approaches when one have access only to population totals/means and probability sample (we support survey::svydesign objects)

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

Package can be installed from CRAN.

Contents

Case study

- **Data**: We focus on the integration of the Job Vacancy survey with Online/Admin data.
- **Population**: entities with at least one vacancy.
- **Target variable**: whether the entity has at least one vacancy in a one shift.
- **Target quantity**: share of companies that has at least one vacancy on a single shift.
- **Assumptions**:
    - There is an overlap between sources which we ignore.
    - We assume that no measurement errors are present in both sources.
    - Admin population was aligned with the JVS definitions so over-coverage is present at the unit-level.
    - Units in admin data were linked using a business identifier so we assume no linkage errors.

Thank you! Now, we can proceed to the practical part