State data agency
Statistics Lithuania

Applying Survey Sampling Theory to Web-Scraped Data: An Analysis of OBEC Data Using the IPW Estimator

Vilma Nekrašaitė-Liegė

Outline

Outline

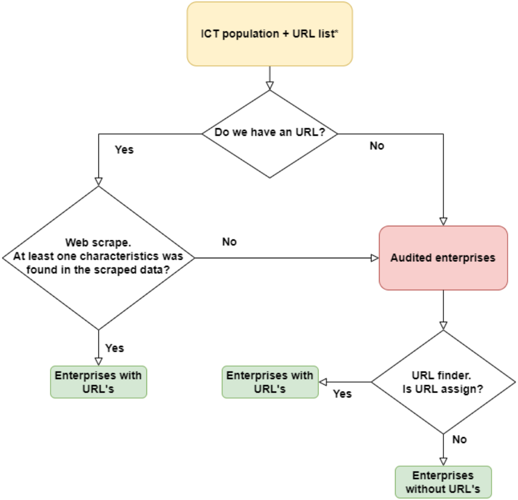
- Introduction
  - Background
  - Research goal and problems
- The use of survey sampling theory
  - Data availability scenarios
  - IPW estimator
- Practical Implementation

Background (ICT survey)

The aim of the European ICT usage surveys is to collect and disseminate harmonised and comparable information on the use of Information and Communication Technologies in enterprises and e-commerce at European level.

- Frequency of data collection: annual.
- Population: enterprises with 10 or more persons employed.
- Statistical unit: enterprise.
- Breakdown: by size class, by NACE Rev 2 categories.

URL search algorithm (1)



* URL list in 2021 was provided by private company, for the other year – the previous year list is used.

URL search algorithm (2)

Web scrape

1. Selenium module is used in Python.
2. All data are saved in the sqLite database.
3. More than 10 different search phrases (Enterprise ID, name, contact information) are used to check if this is the right page.

URL finder

1. Sending search terms to a search engine.
2. Scraping the result URLs.
3. Extracting the scraped data.
4. Creating a machine learning or rule-based model to link websites to enterprises:
   - Logistic regression, random forest models are used.
   - Indicators: enterprise ID, name, municipality, street, zip code, e-mail, telephone, Enterprise's name in URL.
   - No more then one URL is assign to enterprise.

Research goal and problems

Available Data:

- Target population, URL list.

Research Goals:

- Scrape all URLs, find company characteristics such as link to social media presence, e-commerce realization, and estimate proportions for the entire population.

Challenges:

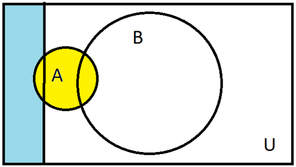
- Not all URL can be scraped.
- The values of the study variables are not directly observable, they are estimated .

Solution:

- Use sample theory methodology.

Outline

Data availability scenarios



Here
$\mathcal{U}$ – a finite population;
A – a probability (reference) sample;
B – a non-probability sample.

Datasets	Scenario I y	Scenario I $\mathbf{x}$	Scenario II y	Scenario II $\mathbf{x}$	Scenario III y	Scenario III $\mathbf{x}$	Scenario IV y	Scenario IV $\mathbf{x}$
$\mathcal{U}$						✓		✓
A		✓	✓	✓		✓	✓	✓
B	✓	✓	✓	✓	✓	✓	✓	✓

Target parameter

Let y be a study bivariate variable with the fixed values $y_1, \ldots, y_N$ in the population $\mathcal{U} = \{1, \ldots, N\}$. The values y_k are observed in the non-probability sample $B \subset \mathcal{U}$.

We aim to estimate the population proportion

$$p_y = \mu_y = \frac{1}{N} \sum_{k=1}^{N} y_k.$$

Let $x^{(0)}, ..., x^{(m)}$ be $m+1$ auxiliary variables completely known in $\mathcal{U}$. For the element $k \in \mathcal{U}$, these variables attain the vector value $\mathbf{x}_k = (x_{k0}, \ldots, x_{km})'$ with $x_{k0} = 1$.

A propensity score

The indicators

$$R_k = \begin{cases} 1 & \text{if } k \in B, \\ 0 & \text{otherwise} \end{cases}$$

describe inclusion of the unit $k \in \mathcal{U}$ to the non-probability sample.

The probability $\pi_k^* = P(R_k = 1 \mid \mathbf{x}_k, y_k)$ called a propensity score is used to describe the inclusion of $k \in \mathcal{U}$ into the non-probability sample B.

A set of assumptions is usually imposed to simplify the modeling of the propensity scores.

A set of assumptions

(A1) $P(R_k = 1 \mid \mathbf{x}_k, y_k) = P(R_k = 1 \mid \mathbf{x}_k)$, $k \in \mathcal{U}$;

(A2) $\pi_k^* > 0$, $k \in \mathcal{U}$;

(A3) $P(R_k = 1, R_l = 1 \mid \mathbf{x}_k, \mathbf{x}_l) = P(R_k = 1 \mid \mathbf{x}_k)P(R_l = 1 \mid \mathbf{x}_l)$;

To estimate the propensity scores, a parametric logistic regression model is applied:

$$\pi_k^* = \pi(\mathbf{x}_k, \beta) = \frac{\exp\{\mathbf{x}_k' \beta\}}{1 + \exp\{\mathbf{x}_k' \beta\}}, \quad k \in \mathcal{U}.$$

Inverse probability weighting (IPW) estimator

Then the ML estimators of the propensity scores are

$$\widehat{\pi}_k^* = \pi(\mathbf{x}_k, \widehat{\beta}), \quad k \in \mathcal{U}.$$

Taking the weights $\widehat{w}_k^* = 1/\widehat{\pi}_k^*$, the estimator

$$\hat{p}_B = \frac{1}{\widehat{N}} \sum_{k \in B} \widehat{w}_k^* y_k, \quad \text{where} \quad \widehat{N} = \sum_{k \in B} \widehat{w}_k^*, \tag{1}$$

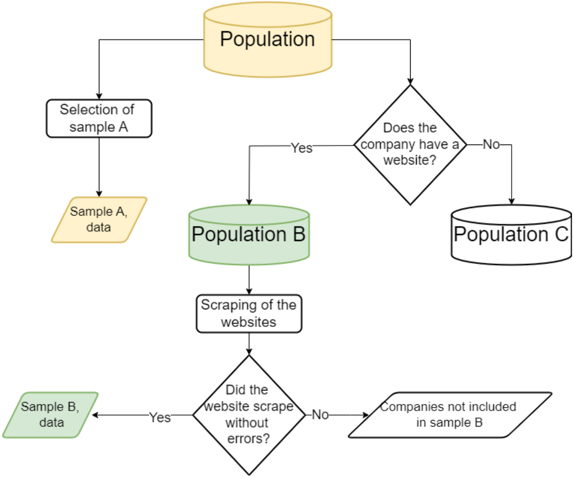
is called the IPW estimator.

Outline

Data

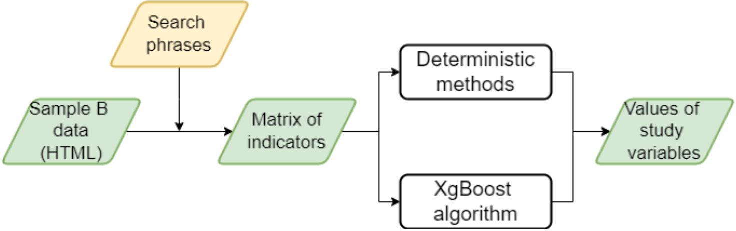
- $\mathcal{U}$ – The Information and Communication Technology (ICT) study population ($N = 13\,403$).
- $y_{(j)}$ – binary study variables indicating whether the company has links to social networks on its website, engages in e-commerce.
- Study paramater is a population mean $\mu_{y_{(j)}} = t_{y_{(j)}}/N$.
- A – probability sample ($n_A = 3\,077$).
- B – non-probability sample ($n_B = 7\,903$).
- $x^{(0)}, \ldots, x^{(m)}$ – auxiliary variables, such as number of employees, income, NACE and region indicators.

Algorithm



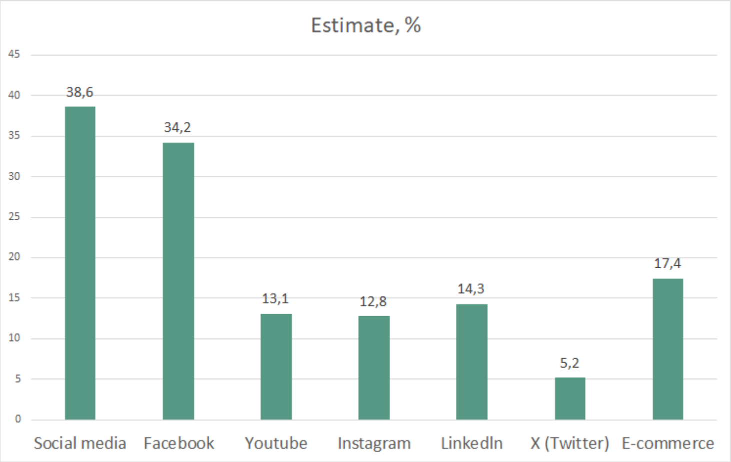
Population
Selection of sample A
Does the company have a website?
Yes
No
Sample A, data
Population B
Population C
Scraping of the websites
Did the website scrape without errors?
Sample B, data
Yes
No
Companies not included in sample B

Study variables



- Deterministic methods are used to determine the values of study variables that indicate whether a company website has a link to certain social network;
- A study variable's values indicating whether the company is engaged in e-commerce are obtained by applying the XgBoost algorithm.

Results



Estimate, %
45
40
35
30
25
20
15
10
5
0
38,6
34,2
13,1
12,8
14,3
5,2
17,4
Social media
Facebook
Youtube
Instagram
LinkedIn
X (Twitter)
E-commerce

R code (1)

```
#-----------------------------------------------------------------------------------------------
# --------------------------Part were you have to make changes-----------------------------------
#-----------------------------------------------------------------------------------------------
rm(list=ls(all=TRUE))
# set your working directory
setwd("C:/Users/VilmaNe/Documents/Statistika/Imoniu_statistika/OBEC_group/SocialMedia")

#provide a dataset of population
population <- read_excel("population.xlsx")
#provide information about variables in population:
#1. name of identificator:
id="ID"
#2. how many categorical variables are in the population?
categorical_n=2
categorical_variables=c("Region", "Nace")
#3. how many numeric variables are in the population?
numeric_n=1
numeric_variables=c("x")
#4. Name of indicator about the www
www=c("www")


#provide a dataset of the sample (scraped)
sample <- read_excel("sample.xlsx")
#provide information about variables in the sample:
#1. name of identificator:
ids="ID"
#2. how many indicators are in the sample?
indicators_n=3
indicators_variables=c("ind1", "ind2", "ind3")

#-----------------------------------------------------------------------------------------------
# --------------------------Part were you just run the code--------------------------------------
#-----------------------------------------------------------------------------------------------
```

R code (2)

```
#-----------------------------Estimation----------------------------------------

# Creating a dataset, where all results will be saved
REZ = data.frame(matrix(ncol = 4, nrow = 0))
colnames(REZ) <- c('Variable', 'total', 'percentage_www', 'percentage_pop')

First=indicators_variables[1]
Last=indicators_variables[length(indicators_variables)]

i=1
for(i in which(indicators_variables == First):which(indicators_variables == Last)){
  y=dplyr::pull(sample_w, indicators_variables[i])
  table(y)
  tt_B=sum(y*sample_w$w)/sum(sample_w$w)*NN_www
  mu_www=tt_B/NN_www*100
  mu_pop=tt_B/NN*100
  REZ[nrow(REZ)+1,]=c(indicators_variables[i],
                      round(tt_B, 2),
                      round(mu_www, 2),
                      round(mu_pop,2))
}

# -------------- All results are exported in xlsx file to the same working directory
write_xlsx(REZ,"Results_LT_case.xlsx")
```

Literature

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology* 46:1–28.

Burakauskaitė, I., Čiginas, A. (2023). An approach to integrating a non-probability sample in the population census. *Mathematics* 11:1782–1795.

Čiginas, A., Krapavickaitė, D., Nekrašaitė-Liegė. (2024). *Evaluating the impact of a non-probability sample-based estimator in a linear combination with an estimator from a probability sample.* Submitted.

Chen, Y., Li, P., Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association* 115:2011–2021.

Golini N., Ridhi, P. (2024). Integrating probability and big non-probability samples data to produce Official Statistics. *Statistical Methods and Applications* 33:555-580.

Kim, J.K., Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review* 89:382–401.

Kim, J.K., Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review* 87:177–191.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics* 12:685–726.

Rueda, M.d.M, Pasadas-del-Amo, S., Rodríguez, B.C., Castro-Martín, L., Ferri-García, R. (2023). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal* 65:1–19.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology* 48:283–311.

State data agency
Statistics Lithuania

Applying Survey Sampling Theory to Web-Scraped Data: An Analysis of OBEC Data Using the IPW Estimator

Vilma Nekrašaitė-Liegė