5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16415

Estimation by kernel weighting of parameters related to employment in the confinement period

Beatriz Cobo[1], Luis Castro[1], Jorge Rueda[2]
[1]Department of Quantitative Methods for Economics and Business , University of Granada, Spain, [2]Department of Statistics and Operations Research, University of Granada, Spain.

Abstract

During the period of COVID-19's confinement, new working methods that were not normally used began to be relevant. This is the case of teleworking or the use of new techniques for conducting surveys. The gold standard for carrying out surveys is probability sampling based on face-to-face interviews, but due to this situation of social isolation, non-probabilistic methods, such as online or web surveys, began to be used. However, in order to make reliable estimates from non-probability samples we must use special techniques to reduce the bias that appears in them.

In this paper we will study a technique for bias reduction in non-probabilistic surveys that stands out for its promising results, known as Kernel Weighting. It requires a probabilistic sample as auxiliary information, and its performance can be improved using Machine Learning techniques, such as regularised logistic regression. We will use a non-probabilistic survey focused on studying the employment situation of the Spanish population during COVID-19, and as probabilistic survey the CIS Barometer of May 2020. We will compare the new estimates with those obtained in the original survey, observing important differences.

Keywords: *Regularized logistic regression; kernel weighting; employment; confinement period; COVID-19.*

1. Introduction

The COVID-19 brought with it a situation of confinement never experienced before that affected the habits and behaviors of Spaniards both economically, at work and socially. Focusing on the first Pérez et al. (2022) conducted a non-probabilistic survey using snowball or chain sampling to recruit respondents during the confinement period in Spain and obtained a data set with sociodemographic variables, variables related to residence during confinement with employment status with household chores, with health, and with politics, that it to say, how they lived and felt as well as their perceptions while in lockdown. Cowan (2020) used IPUMS-CPS survey data to examine how workers transitioned between labor market states and which workers have been most affected by the pandemic. Coibion et al. (2020) used the Nielsen Homescan survey to show the large effects of the pandemic on job losses. Cajner et al. (2020) measured the evolution of the US labor market during the first four months of the pandemic using weekly administrative payroll data from the largest US payroll processing company. Fairlie et al. (2020) used CPS data to focus on how the job market crisis due to the pandemic has affected racial and ethnic minorities compared to whites. Schieman et al. (2021) found that the conflict between work and life decreased among people without children at home; in the case of having children, these changes were limited to the age of the youngest child in the home and the degree of work-home integration.

In most of these studies, to carry out an investigation in which we want to know what happens in the population, it is essential to carry out a survey. The ideal in these cases is to use a probabilistic sampling to ensure that the sample is random and we will be able to make inferences later. However, years ago researchers began using non-probability surveys. Non-probability sampling is a method of selecting units from a population using a subjective (ie, non-random) method. Since non-probability sampling does not require a complete survey framework, it is a quick, easy, and inexpensive way to collect data.

Following the onset of the COVID-19 pandemic, conventional survey data collection efforts (e.g., pencil-and-paper surveys as part of population-representative household surveys) came to a halt due to lockdowns, mobility limitations and social distancing requirements. Because of this, what has occurred is an increase in telephone surveys among others to meet the rapidly evolving needs for timely and policy-relevant microdata to understand socioeconomic impacts and responses to the pandemic. Gourlay et al. (2021) provided an overview of options for the design and implementation of telephone surveys to collect representative data from households and individuals. In addition, they identified the requirements for telephone surveys to be used in national statistical offices. De Boni (2020) briefly presented some of the advantages that may have driven web surveys. She talks about the growing popularity of these in the COVID-19 context, since in addition to the possibility of collecting data remotely, it allows social distancing. She also tells us about

the side effects and ethical issues that need to be considered when planning such surveys and interpreting their results. Hlatshwako et al. (2021) tell us about online health research surveys, the challenges in the implementation and interpretation of data from this type of survey, and the considerations that must be made to make the most of the research. Ali et al. (2020) show us how the COVID-19 pandemic has forced researchers to explore innovative ways to collect public health data in an efficient and timely manner. Social media platforms were explored as a recruiting tool for researchers in other settings; however, its feasibility for collecting representative survey data during infectious disease epidemics remained unexplored. Roig et al. (2022) conducted a study of the gender gap related to tasks within the home during the COVID-19 health crisis, introducing the variable size of the municipality in the analyses.

Our work focuses on the combination of data obtained through probabilistic and non-probabilistic surveys with the aim of obtaining more reliable estimates through regularized logistic regression and kernel weighting techniques. As a non-probabilistic survey, we will base ourselves on the survey carried out by Pérez et al. (2022) and we will study the block of questions referring to employment and study as a probabilistic survey the CIS Barometer of May 2020.

2. Methodology

Let s_v be the non-probabilistic sample, obtained from the population of interest U, from a survey of volunteers which has a size n_v, the variable of interest y, and the vector of auxiliary variables $x = (x_1, \dots, x_p)$. In order to eliminate the volunteer bias in non-probability surveys, we must have available auxiliary information that is accurate and closely related to the topic under study. Depending on the type of auxiliary information available, we distinguish different types of bias reduction techniques (Rueda et al., 2020), although in our study we will focus on one, the kernel weighting method, due to its recent appearance and excellent results. This technique is included in the group of those that need a probabilistic reference sample in order to be carried out, from which we only need to know its auxiliary variables.

We define s_r as the probability sample, obtained from the population of interest U, which we will use as a reference and which has a size n_r, and $x = (x_1, \dots, x_p)$ as the vector of auxiliary variables that we have measured both in the reference probability sample s_r and in the non-probability (or volunteer) sample s_v. In the case of s_r, the probability that each individual has of participating in the survey is known and greater than zero, conditions that must be met for the sample to be probabilistic. This probability is known as the first-order inclusion probability (π_r). From these probabilities we obtain what are known as design weights (w_r), which will be key when forming our estimators, and which are obtained by

calculating the inverse of the π_r. In the case of non-probabilistic samples, the inclusion probabilities are not usually known since we lack a probabilistic theory to support them, making it impossible to accurately determine their value. We will therefore assume their value or, as in this case, seek to estimate these probabilities, also called propensities, for the non-probabilistic case (π_v). We define the probability of belonging to the non-probabilistic sample, $\forall i \in U$, as:

$$\pi_{vi} = P[1_i = 1|x_i] \;\; where \;\; 1_i = \begin{cases} 1 & for \;\; i \in s_v \\ 0 & otherwise \end{cases}$$

We seek to estimate the expected value of the probability of inclusion in the non-probability sample, through some model M, $\forall i \in s_v \cup s_r$:

$$\hat{\pi}_{vi} = E_M\left[\hat{1}_i = 1|x_i\right] \;\; where \;\; \hat{1}_i = \begin{cases} 1 & for \;\; i \in s_v \\ 0 & for \;\; i \in s_r \end{cases}$$

To estimate these probabilities, logistic regression models are usually considered, due to the binary nature of the variable to be estimated, although machine learning techniques can be used for the same purpose (Ferri-García and Rueda, 2020). In our work we will use what is known as regularized logistic regression, which adds penalty parameters when estimating the coefficients of the logistic regression. Other methods for obtaining these probabilities can be found at Castro-Martín et al. (2020). In the case of logistic regression, the estimated probabilities are obtained as follows:

$$\hat{\pi}_{vi} = \frac{1}{1 + exp(-\beta x_i)}, \;\; i \in s_v \cup s_r$$

being β the vector of regression coefficients. To calculate these coefficients we must maximize the log-likelihood function:

$$l(\beta) = \sum_{i \in s_v \cup s_r} \hat{1}_i \ln(\hat{\pi}_{vi}) + \left(1 - \hat{1}_i\right)\ln(1 - \hat{\pi}_{vi}) = \sum_{i \in s_v \cup s_r} \left[\hat{1}_i \ln\left(\frac{\hat{\pi}_{vi}}{1 - \hat{\pi}_{vi}}\right) + \ln(1 - \hat{\pi}_{vi})\right]$$
$$= \sum_{i \in s_v \cup s_r} \left[\hat{1}_i \beta x_i - \ln\left(1 + e^{\beta x_i}\right)\right]$$

Regularization strategies introduce penalties in order to avoid overfitting, reduce variance and minimize the influence of less relevant predictors in the model. We apply ridge regularization, which introduces an L2 penalty to the log-likelihood function. The estimated coefficients β will also be obtained by maximizing the log-likelihood function with this penalty:

$$l_R(\beta) = \sum_{i \in s_v \cup s_r} \left[\hat{1}_i \beta x_i - \ln\left(1 + e^{\beta x_i}\right)\right] - \lambda \sum_{j=1}^{p} \beta_j^2$$

The penalty depends on the parameter $\lambda \geq 0$ which measures the regularization intensity of the fit. This parameter depends on the real parameters of the regression, and since they are unknown, we will have to choose it arbitrarily or through the hyperparameter adjustment. In our work the selection of this parameter is made by means of cross-validation. This consists of dividing our data into two complementary sets, performing the analysis on one subset (training set) and validating the analysis on the other (test set).

Once we have explained how we estimate the π_v, we move on to the explanation of the technique in charge of reducing the bias, which in our work will be the kernel weighting method.

2.1. Kernel Weighting Method (KW)

Developed by Wang et al. (2020), it is based on the creation of new weights, called pseudo-weights, from the design weights of the individuals in the probability sample weighted according to the similarity they have to the non-probability sample individuals. This similarity is measured with the distance between individuals, through the difference of the estimated probabilities of belonging to the probability sample (analogous to the non-probabilistic case) and to the non-probability sample.

$$d_{ij} = \hat{\pi}_{vi} - \hat{\pi}_{rj}, \;\; i \in s_v, \;\; j \in s_r$$

These distances will have a value between -1 and 1, so we will try to smooth them. For this purpose, we make use of kernel functions centred at zero, which are continuous, symmetric, and positive functions, and can be used as density functions of statistical distributions. The closer the distance is to zero, the more similar the individuals will be with respect to their auxiliary variables, since propensities are estimated as a function of these variables. The more similar the individuals are, the more the KW will assign a higher percentage of the design weight from the individual in the probability sample to the individual in the non-probability sample. These percentages are called kernel weights, and are obtained:

$$k_{ij} = \frac{K\{d_{ij}/h\}}{\sum_{i \in s_v} K\{d_{ij}/h\}}$$

where K{.} is a kernel function centred at zero, and h is the corresponding bandwidth (Epanechnikov, 1969). The kernel weight values will be between zero and one, and the sum of all of the volunteer sample values is one. To calculate the pseudoweights KW we will sum the reference sample design weights w_r of each $j \in s_r$, weighted by the kernel weights of the i-th individual, from the non-probability sample, with each j-th individual from the reference sample. The expression of these pseudoweights is as follows:

$$w_i^{KW} = \sum_{j \in s_r} w_{rj}\, k_{ij}$$

Finally obtaining the estimator of the total:

$$\hat{Y}_{KW} = \sum_{i \in s_v} w_i^{KW}\, y_i$$

3. Application

We have applied the proposed methodology in order to correct the bias of the survey conducted during the COVID-19 lockdown in Spain carried out by Pérez et al. (2022). The survey was distributed following a snowball method via email and social networks. Therefore, a significant lack of representativity is to be expected. However, it was conducted between 28th April and 14th May, 2020, and it included some variables in common with the CIS Barometer of May 2020. The latter can be considered a reference probabilistic survey since it was carried out by an official Spanish institution following strict methodologies.

In particular, the following covariates will be used for applying the kernel weighting method: state, province, urban density, sex, age, education level, employment status, last electoral vote, intended electoral vote and confidence in the government during the pandemic.

Once we have obtained representative weights, these can be used in order to produce estimations for the economical variables included in the survey of interest. In Table 1, the differences between the estimations obtained with the naive mean and the estimations obtained after correcting the bias can be observed. Depending on the variable, the change can be quite significant.

Table 1. Naive and corrected estimations for the variables of interest (those directed to workers and students)

Variable	Naive	KW	Variable	Naive	KW
WORK.PROD	64.8	63.7	STU.PROD	86	72.8
WORK.EXP.1	17	21	STU.EXP.1	0.9	0.1
WORK.EXP.2	23.7	23.3	STU.EXP.2	7.9	8.6
WORK.EXP.3	26.2	30.2	STU.EXP.3	23.1	25.4
WORK.EXP.4	24.8	19.4	STU.EXP.4	28	26.7
WORK.EXP.5	44.9	47.2	STU.EXP.5	68	61.1
WORK.PERCEP.IN.1	27.3	32.4	STU.EXP.6	24.3	27.7
WORK.PERCEP.IN.2	5.7	8.4	STU.EXP.7	47.9	45.2
WORK.PERCEP.IN.3	10.6	11.8	STU.EXP.8	26.4	19.2
WORK.PERCEP.IN.4	3.3	1.9	STU.EXP.9	10.7	17.4
WORK.PERCEP.IN.5	59.3	55.8	STU.PERCEP.1	11.4	16.6
WORK.PERCEP.OUT.1	31.7	34	STU.PERCEP.2	49.8	46.6
WORK.PERCEP.OUT.2	8.7	8.9	STU.PERCEP.3	25	34.7
WORK.PERCEP.OUT.3	9	8.3	STU.PERCEP.4	39	45.1
WORK.PERCEP.OUT.4	55.5	53.4	STU.PERCEP.5	12.1	5.7

Note: For more information on the variables of interest, see the article on Pérez et al. (2022)

The results show the percentages of individuals which agree with each of the statements surveyed. For some general questions, such as the percentage of people who feel their work performance has been affected due to the lockdown, the results stay similar. This may be explained either by a lack of bias for those questions or because more relevant covariates are needed in order to reduce some underlying bias. In fact, when the students are asked the same question, we observe a 13.2% difference in the estimation. These changes also occur for some specific questions. For example, 5.1% more individuals than those initially observed thought that their work would be affected by an economical crisis after the pandemic.

4. Conclusions

We have evaluated a reweighting method for correcting bias in non-probabilistic surveys, by using a reference probabilistic survey. We have also confirmed with a practical application the significance of the differences in the estimations obtained. However, these differences will depend on the presence of an actual bias and on the importance of choosing the right covariates.

References

Ali, S.H., Foreman, J., Capasso, A., Jones, A.M., Tozan, Y. & DiClemente, R.J. (2020). Social media as a recruitment platform for a nationwide online survey of COVID-19 knowledge, beliefs, and practices in the United States: methodology and feasibility analysis. *BMC Medical Research Methodology.* 20, 116. https://doi.org/10.1186/s12874-020-01011-0.

Cajner T., Crane L.D., Decker R.A., Grigsby J., Hamins-Puertolas A., Hurst E., Kurz C. & Yildirmaz A. (2020). The US Labor Market during the Beginning of the Pandemic Recession. *National Bureau of Economic Research*, w27159.

Castro-Martín, L., Rueda, M. D. M., & Ferri-García, R. (2020). Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. Mathematics, 8(6), 879.

Coibion, O., Gorodnichenko Y. & Weber M. (2020). Labor Markets During the COVID-19 Crisis: A Preliminary View. *National Bureau of Economic Research*. w27017

Cowan, B.W. (2020). Short-run effects of COVID-19 on U.S. worker transitions. *National Bureau of Economic Research.* w27315. 10.3386/w27315.

De Boni, R.B. Web surveys in the time of COVID-19. (2020). *Cadernos de Saúde Pública: Reports in Public Health.* 36(7):e00155820.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications, 14(1), 153-158.

Fairlie, R.W., Couch K.A., & Xu H. (2020) The Impacts of COVID-19 on Minority Unemployment: First Evidence from April 2020 CPS Microdata. *National Bureau of Economic Research,* w27246.

Ferri-García, R., & Rueda, M. D. M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PloS one, 15(4), e0231500.

Gourlay, S., Kilic, T., Martuscelli, A., Wollburg, P., & Zezza, A. (2021). High-frequency phone surveys on COVID-19: good practices, open questions. Food Policy, 105, 102153.

Hlatshwako, T.G., Shah, S. J., Kosana, P., Adebayo, E., Hendriks, J., Larsson, E.C., Hensel, D.J., Erausquin, J.T., Marks, M., Michielsen, K., Saltis, H., Francis, J.M ., Wouters, E., & Tucker, J.D. (2021). Online health survey research during COVID-19. *The Lancet: Digital Healt.* 3. https://doi.org/10.1016/S2589-7500(21)00002-9.

Peréz, V., Aybar, C. & Pavía, J.M. (2022). Dataset of the COVID-19 lockdown survey conducted by GIPEyOP in Spain. *Data in Brief,* 40, https://doi.org/10.1016/j.dib.2021.107700.

Roig, R., Aybar, C., & Pavía, J. M. (2022). COVID-19, gender housework division and municipality size in Spain. Social Sciences, 11(2), 37.

Rueda, M.D.M., Ferri-García, R., & Castro, L. (2020). The R package Non-ProbEst for estimation in non-probability surveys. The R Journal, 12(1), 406-418.

Schieman, S., Badawy, P.J., Milkie, M.A. & Bierman A. (2021). Work-life conflict during the COVID-19 pandemic. *Socius: Sociological Research for a Dynamic World*, 7, 10.1177/2378023120982856.

Wang, L., Graubard, B. I., Katki, H. A., & Li, A. Y. (2020). Improving external validity of epidemiologic cohort analyses: a kernel weighting approach. Journal of the Royal Statistical Society: Series A (Statistics in Society), 183(3), 1293-1311.