COVID-19 REPORTS

Recommendations for eliminating selection bias in household surveys during the coronavirus disease (COVID-19) pandemic

May 2020

Summary

In the context of the global coronavirus disease (COVID-19) pandemic, the Economic Commission for Latin America and the Caribbean (ECLAC) is preparing a series of short publications with relevant policy recommendations. A number of recommendations are made in this note to address bias problems that may arise in household surveys carried out during this exceptional period, as a complement to the suggestions made in a previous note on the sample designs for this type of survey.

Introduction

In an attempt to slow the spread of COVID-19, countries have imposed movement restrictions on people, which have prevented the face-to-face collection of information for household surveys. In order to address this issue and to continue producing relevant and timely official statistics, some national statistical offices (NSOs) have resorted to conducting surveys by telephone or via the Internet. The document "Recommendations for the publication of official statistics from household surveys in the context of the coronavirus disease (COVID-19) pandemic" contains some possible lines of action for drawing the sample of households to participate in the surveys, in particular using a selected panel from a recent period for which the telephone contact information is available (ECLAC, 2020). This note complements the recommendations made in that document, by proposing two approaches to minimize the bias generated by non-response that will be encountered when carrying out surveys by telephone. In addition, a third approach is proposed for those instances in which it is difficult to obtain auxiliary information.

A. Detecting bias

Changing the household survey data collection modality from face-to-face interviews to a telephone- or web-based modality may have unintended consequences and, in particular, may generate biases (of selection, coverage and non-response) among survey respondents. In a scenario where a sample of households from a previous period is being used (hereinafter the "original sample") and where every effort is being made to contact those selected households, the process is inevitably exposed to the following difficulties:

- Not all the households in the original sample provided their telephone contact information.
- Some households provided their contact information, but at the time of the interview they do not live at the selected address.
- Some households provided their contact information, but they have since changed their contact telephone number.
- Not all households that provided their contact information are willing to answer the survey questionnaire.

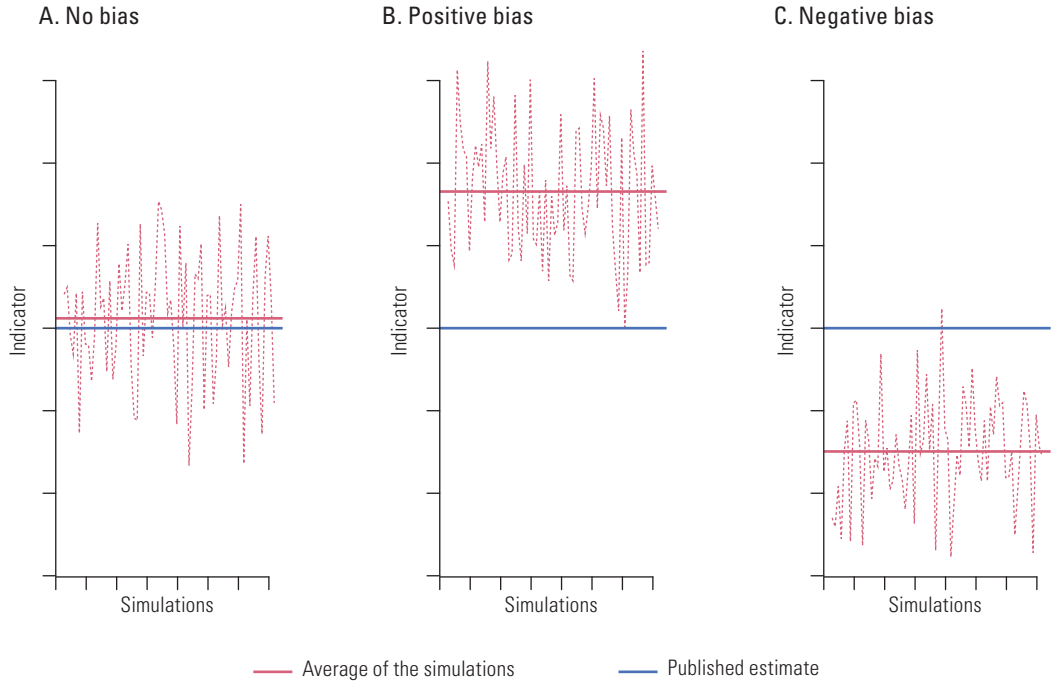
For example, if it is assumed that 85% of the sample did provide contact details and that the probability of a contacted household responding to the entire survey is 80%, then responses would be obtained from just 68% of the original sample. In addition to these considerations, the possible attrition effect on the panel —loss of participants the longer the panel is used for— should also be taken into account, since there will be households that will stop responding to the survey because they are contacted repeatedly.

In this scenario, it is highly likely that respondent households do not have similar characteristics to non-respondent or non-covered households, and that the reasons for household non-response to the survey are associated with the phenomenon being measured (for example, there are more unemployed people in non-respondent households, or non-response rates are higher among households living in poverty). This suggests that the information obtained from the respondent households will be biased, so it cannot be used without making some sort of correction.

Therefore, once the information has been collected in a certain period, the first step should be to estimate the magnitude of the bias. In line with what was proposed by ECLAC (2020), one possibility is to use simulated scenarios, on the basis of the final distribution of households that actually completed the survey within the primary sampling units (PSUs). Using simulations, it is possible to try to predict how the estimators would have behaved in the month of the sample selection if only current partial information had been available. The difference between the published (unbiased) estimates and the simulated (possibly biased) estimates will give an idea of the magnitude of the bias.

Three possible scenarios for identifying bias are presented in figure 1. The graph on the left illustrates a case in which there is no bias, while the graphs in the centre and on the right depict scenarios corresponding to a significant bias. The blue horizontal line corresponds to the published estimate for the month in which the original sample was first selected, while the red horizontal line represents the average of the simulations using the effective sample. Each of the simulations' results are represented by the dotted lines. In the last two scenarios (figures 1.B and 1.C), most of the simulations do not include the published estimate and, therefore, it can be assumed that there is a bias.

Figure 1
Three possible scenarios when identifying selection bias



Source: Economic Commission for Latin America and the Caribbean (ECLAC).

In general, the published confidence interval can be used to determine whether and to what extent a bias exists. It is recommended to review all scenarios that are, on average, further than a half standard deviation from the point estimate of the original sample. In the most optimistic scenario, that with no bias, an NSO would be well placed to replicate the usual processes of inference. However, if a bias is suspected, and depending on the auxiliary information available, one of the alternatives described below may be applied.

B. Propensity score adjustment

Telephone surveys have been conducted in many countries of the region, based on a sample from previous periods, as an alternative given the restrictions on face-to-face information collection. ECLAC (2020) suggested that the best strategy for selecting the panel was to follow up on a complete sample from previous months (for example, February 2020), since choosing a subsample from an amalgamation of samples from previous months necessitates rather complex inclusion probability calculations. Therefore, if a probability sample is used that is in line with the aforementioned recommendations, the expansion factors can be adjusted in a differential manner to correct for the selection bias.

The approach proposed by Rosenbaum and Rubin (1983) is useful for elucidating the structure of non-response and therefore for correcting coverage bias and non-response bias (Lensvelt-Mulders, Lugtig and Hubregtse, 2009). For the effective management of non-response, the dichotomous variables I_k and D_k are considered, which indicate whether the household belongs to the original sample and whether it has responded to the telephone survey, respectively. Assuming that the distribution of effective responses can be estimated, the propensity score of a household in the sample is given by:

$$\phi_k = Pr(D_k = 1 | I_k = 1)$$

This score is different for each household and can be estimated using the panel data. The original sample, for which all the information in the questionnaire was obtained in a previous period, is an excellent starting point for efforts to eliminate bias, since a set of covariates x will be available to determine the best model to estimate the pattern of non-response in the sample of effective respondents. For example, useful covariates for estimating the propensity score may include sex, age, educational level, area and geographical region of residence, occupation status and per capita household income in the original survey.

Assuming that the propensity score depends on some linear combination of the covariates available in the original sample, it is possible to fit a model where the dependent variable is D_k and the covariant vector is denoted by x. Kim and Riddles (2012) show that it is possible to use a model based on the propensity-score adjustment of the telephone sample using the following expression:

$$\text{logit}(\hat{\phi}_k) = x_k \hat{\beta}$$

where $\hat{\beta}$ is the estimated coefficient vector of the logistic regression. Particular attention should be paid to the choice of predictors in the logistic regression model, which should work well if the available auxiliary information variables are relevant and explanatory of the telephone response; otherwise, this methodology will not help to reduce bias (and could possibly exacerbate it) and will result in larger standard errors.

Given that the original weights from the telephone survey are denoted as d_k, and having estimated $\hat{\phi}_k$ for respondents and non-respondents of the telephone sample, then the adjusted expansion factor would take the following form:

$$w_k = \frac{d_k}{\hat{\phi}_k}$$

Using the expansion factor w_k in the calculation of the desired estimators would minimize the selection bias generated by the change in data collection modality. The factors associated with coverage bias may not be the same as those associated with non-response bias, so it would probably be helpful to model these problems separately and then use the two propensity scores as independent adjustment factors.

C. Two-stage calibration method

Särndal and Lundström (2006) assert that when sample surveys are affected by non-response, it is desirable to have a weighting system that reproduces the auxiliary information available and that is efficient when estimating any characteristic of interest in a multipurpose study. Calibration estimators (Deville and Särndal 1992) satisfy these conditions and can be easily adjusted to mitigate the bias generated by the change in collection modality.

In principle, two sources of auxiliary information are available. On the one hand, there is the information that is usually used to calibrate the expansion factors in a regular survey (denoted as x_{1k}). On the other hand, the variables that were measured in the original sample (denoted as x_{2k}) are available. This means that, after calculating the weights for the telephone survey (s_t), it is possible to calibrate them at the level of the auxiliary information available in the original sample (s_m), at the national level (u), or by strata of interest.

The first stage consists therefore of finding a set of calibrated weights subject to the following restriction (Särndal, 2007):

$$\sum_{s_m} w_{1k}\, x_{1k} = \sum_{U} x_{1k}$$

In the second stage, the intermediate weights w_{1k} must be used to calculate the final calibration weights w_k of the telephone sample, subject to the following restriction:

$$\sum_{s_t} w_k\, x_{2k} = \sum_{s_m} w_{1k}\, x_k = \left(\begin{array}{c} \sum_{U} x_{1k} \\ \sum_{s_t} w_{1k}\, x_{2k} \end{array} \right)$$

To ensure consistency between the official figures already published and those that the telephone survey may produce, the use of calibration estimators is desirable. Adopting this approach ensures a robust inferential structure in the presence of available information, since both the sampling error (increasing accuracy) and the non-response error (eliminating bias) are reduced.

For example, a two-stage calibration procedure could use the following reference variables:

(i) In the first stage, the calibration of the original sample weights could be based on the totals for age, subnational region, area and sex, available from robust population projections (or census counts, if the last census was conducted recently).

(ii) In the second stage, the calibration of the weights of the telephone sample could be based on the variables indicated above, and also on the totals of per capita income, occupation status, branch of activity and level of education, obtained from the publication containing the results of the original survey.

The general purpose of the calibration process is to find a moderate number of constraints that allow approximately unbiased estimates to be calculated with a lower variance than that generated with the original expansion factors. In general, calibration processes can be classified into one of the following three categories:

(i) Calibration with continuous variables, which is where the calibration is performed with the totals of continuous variables such as income and expenditure, among others.

(ii) Poststratification with categorical variables, which is where the calibration is carried out with the population sizes (based on demographic projections or administrative records) of subgroups of interest.

(iii) Raking using categorical variables, which is defined as a calibration based on the marginal sizes of contingency tables of subgroups of interest. Unlike the previous categories, raking does not take into account the sizes of the crosses, but only the marginal sizes; therefore, this method leads to fewer restrictions.

D. Poststratification based on multilevel models

In the event that an NSO has not used a panel defined by a probability sample from a previous period, none of the options outlined above will be able to be implemented, since the auxiliary information needed to decipher the response mechanism of the telephone survey will not be available. In these cases, and as a last resort, it is possible to perform some empirical exercises based on predictive models to get an idea of the magnitude of the bias and correct for it.

Multilevel regression with poststratification (MRP) is a useful technique for predicting a parameter of interest within small domains by modelling the mean of the conditional variable of interest on poststratification counts. This method was initially proposed by Gelman and Little (1997) and expanded by Park, Gelman, and Bafumi (2004). This technique is widely used to correct for survey selection bias and its ultimate goal is to estimate a parameter of interest (totals, means or proportions, among others) for all strata (domains, categories or subgroups) in a finite population.

Similar to the two-stage calibration model, auxiliary information on the totals of people according to the characteristics of interest are needed to apply this methodology; for example, census information aggregated at the regional level on the total number of people for all possible combinations of the variables sex, age and education level.

The MRP model is composed of two parts: the first involves adjusting a multilevel regression model based on the household survey; and the second is the poststratification, using census counts. The methodology steps are described below.

(i) Both the characteristic of interest y and the auxiliary covariates x (demographic and geographical) are observed in the survey. It can be assumed that covariates define a set of J cells or poststrata $(j=1,...,J)$. For example, it might be that the poststrata are made up of crosses between 5 age categories, 4 education categories, 2 area categories (rural and urban), 2 sex categories and 20 regions (provinces, departments or states). Thus, $j=5 \times 4 \times 2 \times 2 \times 20=1,600$ poststrata.

(ii) A multilevel regression model $y\sim(X|Región)$ is fitted to obtain an average prediction $\hat{y}$ of the characteristic of interest conditional on its demographic and geographical variables. For this step, this value must be predicted for each poststratum; that is $\hat{y}_j$, must be obtained for $j=1,...,J$.. Please note that this regression model is multilevel, since it depends on the 20 regions defined above.

(iii) For each cell j, information on the population N_j is available, extracted from the demographic projections. Therefore, the national average of the variable of interest can be estimated as a weighted average of the estimates $\hat{y}_j$:

$$\hat{\bar{y}} = \frac{\sum_j N_j \hat{\bar{y}}_j}{\sum_j N_j}$$

In the particular case that the variable of interest is the national unemployment rate, it is defined as a dichotomous variable y_{ij}, which describes occupation status (employed or unemployed) of the nth person in the workforce, who belongs to the poststratum j. The model seeks to relate the components of y_{ij} with the auxiliary information x_{ij}, which can be achieved by using a multilevel logistic regression model on the probability of being unemployed $p_{ij}=Pr(y_{ij}=1)$, defined as:

$$\text{logit}(p_{ij}) = \beta_0^j + \mathbf{x}_i\boldsymbol{\beta}$$

In the previous expression, the coefficients $\boldsymbol{\beta}$ denote the fixed effects of the variables (age, education, area and sex) on the model's probabilities. Meanwhile, the first summand represents the random intercept of the model that depends on the region.

$$\beta_0^j = \alpha_0 + \alpha_1^j$$

where the coefficients $\alpha_0 + \alpha_1^l$ represent the random intercept induced by the region l (l=1,...,L). After estimating the coefficients of the multilevel regression model, the probability that any person in the workforce is unemployed can be estimated, conditioned on his or her personal information in terms of the region and other auxiliary information variables. Considering that n_j denotes the sample size in the poststratum j, the prediction of the average number of unemployed individuals in this crossing will correspond to the average of the probabilities $\hat{p}_{ij}$ that were predicted by the model in the same poststratum j. In other words:

$$\hat{\bar{y}}_j = \frac{\sum_i \hat{p}_{ij}}{n_j}$$

The most important point of this technique is the adjustment of the differences between the sample and the population. To carry out the poststratification process, the census counts for each poststratum are used, that is, how many people in the workforce are in each of the 1,600 combinations of all the possible crosses of the auxiliary variables. These quantities will be denoted by $N_1,...,N_j,...,N_{1600}$. Lastly, the estimate from the intention of the national unemployment rate is given by:

$$\hat{\bar{y}} = \frac{\sum_{j=1}^{1600} N_j \, \hat{\bar{y}}_j}{\sum_{j=1}^{1600} N_j}$$

E. Conclusions

This document presents a simulation approach that can be applied by NSOs to correct possible selection biases in household surveys carried out remotely, amid the movement restrictions that prevail in the countries of the region.

If a probability sample from a previous period (panel) has been used in the survey, it is recommended to use one of the first two methodologies described, propensity score adjustment or the two-stage calibration method, to eliminate bias. If, however, a panel was not used, it is recommended to opt for the MRP method, as a last resort, to provide the country with data for evidence-based public policies.

Bibliography

Deville, J. C. and C. E. Särndal (1992), "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, vol. 87, No. 418.

ECLAC (Economic Commission for Latin America and the Caribbean) (2020), "Recommendations for the publication of official statistics from household surveys in the context of the coronavirus disease (COVID-19) pandemic", April [online] https://repositorio.cepal.org/bitstream/handle/11362/45382/1/S2000273_en.pdf.

Gelman, A. and T. Little (1997), "Poststratification into many categories using hierarchical logistic regression", *Survey Methodology*, vol. 27, No. 2.

ILO (International Labour Organization) (2013), "Resolution concerning statistics of work, employment and labour underutilization", nineteenth International Conference of Labour Statisticians, October 2013 [online] http://www.ilo.org/wcmsp5/groups/ public/---dgreports/---stat/documents/normativeinstrument/wcms_230304.pdf.

Kim, J. K. and M. K. Riddles (2012), "Some theory for propensity-score-adjustment estimators in survey sampling", *Survey Methodology*, vol. 38, No. 2.

Lensvelt-Mulders, G., P. Lugtig and M. Hubregtse (2009), "Separating selection bias and non-coverage in Internet panels using propensity matching", *Survey Practice*, 2, No. 6.

Park, D., A. Gelman and J. Bafumi (2004), "Bayesian multilevel estimation with poststratification: State-level estimates from national polls", *Political Analysis,* vol. 12, No. 4.

Rosenbaum, P. R. and D. B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, vol. 70, No. 1.

Särndal, C. E. (2007), "The calibration approach in survey theory and practice", *Survey Methodology*, vol. 33, No. 2.

Särndal, C. E. and S. Lundström (2006), "Estimation in surveys with nonresponse", *Wiley Series in Survey Methodology,* Wiley.

This document is part of a series of reports prepared by the Economic Commission for Latin America and the Caribbean (ECLAC) on the evolution and effects of the COVID-19 pandemic in Latin America and the Caribbean. It was prepared by the Statistics Division, directed by Rolando Ocampo, under the general coordination of Alicia Bárcena, Executive Secretary of ECLAC.

ECLAC

Economic Commission for Latin America and the Caribbean (ECLAC)
Comisión Económica para América Latina y el Caribe (CEPAL)
www.eclac.org