Mixed-mode : How to solve the Missing-Not-At-Random (MNAR) issue of online surveys ?

Keywords: *endogenous selection, Heckman's model, mixed-mode.*

1 Introduction

Online surveys are developing rapidly for producing national statistics. However, this survey mode is often associated with a low response rate, about 30-35% in France. We can also expect certain behaviours such as people deciding to participate to the survey because they care about the topic of the survey. Of course, this can also happen with telephone or face-to-face surveys. But it is very likely that this type of behaviour is more often observed with self-administered surveys. This behaviour is a typical endogenous selection.
Mixed-mode is an opportunity to address this issue directly, provided that the protocol is adapted to construct instrumental variables of participation and that endogenous selection correction models are implemented. We proposed to use Heckman's selection model to correct this endogenous selection. This method is illustrated with the French EpiCov survey, collected online and by phone, during the Covid-19 crisis.

2 Methods

2.1 Notations and statistical framework

Let us consider y our target variable and $\mathscr{P}$ the population of interest of size N. We want to estimate the average of this parameter over the whole population $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$.
We select a random sample $\mathbf{s}$ in this population, with $s_i = 1$ when the individual i is sampled and $s_i = 0$ otherwise. The probability to be sampled π_i is known. The sample design $\mathbf{s}$ may depend on a set of variables $\mathbf{Z}$, available over the whole population $\mathscr{P}$. Practically, in surveys, some non-response occurs. Non-response might be seen as a binary random variable $\mathbf{r}$: $r_i = 1$ when the individual i participates to the survey and $r_i = 0$ otherwise. Then it is possible to build an adapted Horvitz-Thompson estimator of the form :

$$\hat{\mu}^{HT} = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{\pi_i \hat{\rho}_i} s_i r_i \tag{1}$$

with $\hat{\rho}_i$ a prediction of r_i from $\mathbf{z}_i$'s.
As $\hat{\rho}_i$ is defined conditionally to Z, a necessary and sufficient condition for $\hat{\mu}^{HT}$ to be unbiased is that $r_i \perp\!\!\!\perp y|\mathbf{Z}$. This hypothesis corresponds to the Missing-At-Random (MAR) case of non-response. But in general case, r_i may depend on y_i and this hypothesis does not hold anymore. In that Missing-Not-At-Random (MNAR) case, the selection is endogenous and the Horvitz-Thompson estimator is biased.

2.2 The use of Heckman selection models

In Heckman's model [Heckman, 1979], y_i and r_i are simultaneously modelled, according to the following :

$$\begin{cases} \text{(i)} & y_i &=& c^1 + \mathbf{z}_i\chi + \epsilon_i^1 \\ \text{(ii)} & r_i^* &=& c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0 \\ \text{(iii)} & r_i &=& \mathbf{1}(r_i^* \geqslant 0) \end{cases} \tag{2}$$

r_i^* is a latent variable we do not observe. We observe r_i and y_i when $r_i = 1$. The set of variables $(\mathbf{z}_i, \mathbf{w}_i)$ is observed for all i's. Identification conditions are the following :

$$\begin{cases} \mathbb{E}\left(\begin{pmatrix}\epsilon_i^0\\ \epsilon_i^1\end{pmatrix}\middle|\, \mathbf{z}_i, \mathbf{w}_i\right) = 0 \\ \begin{pmatrix}\epsilon_i^0\\ \epsilon_i^1\end{pmatrix} \hookrightarrow \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right) \\ \Sigma = \begin{pmatrix}1 & \varrho\sigma\\ \varrho\sigma & \sigma^2\end{pmatrix} \end{cases} \tag{3}$$

The correlation parameter of the two error terms ϱ bears the proof of a possible endogeneity of the selection when it is different from zero.
In this model, participation equation is based on a latent variable, which involves $\mathbf{z}_i$'s, as well as set of instrumental variables $\mathbf{w}_i$'s. These last are instrumental variables in the sense that they explain the participation but not the output variable. Such variables are fundamental in the convergence of the model.
To solve an endogenous selection issue in surveys, it is usual to impute y_i's from non respondents [Galimard et al., 2018]. Here we propose to use a correction for non-response [Castell and Sillard, 2020]. In fact, the model also gives a predictor of $\mathbb{E}(r_i|\mathbf{Z}, \mathbf{w})$, which can be used as the expression of $\hat{\rho}_i$ in the Horvitz-Thompson estimator (1). Indeed, one can show that:

$$\mathbb{P}(r_i^* \geqslant 0|y_i, \mathbf{z}_i, \mathbf{w}_i) = \Phi\left(\frac{c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \frac{\varrho}{\sigma}(y_i - c^1 - \mathbf{z}_i\chi)}{\sqrt{1-\varrho^2}}\right) \tag{4}$$

where Φ is the distribution of the normal law.
Formula (4) applies for continuous y_i. The formulae can be adapted for binary outcome variables, that is : $\mathbb{P}(r_i^* \geqslant 0|y_i = 1, \mathbf{z}_i, \mathbf{w}_i)$ when the outcome is equal to 1 and $\mathbb{P}(r_i^* \geqslant 0|y_i = 0, \mathbf{z}_i, \mathbf{w}_i)$ when the outcome is equal to 0.
Note that we end with one set of weights for each outcome variable $\mathbf{y}$.

3 Results

In the spring of 2020, France, like many other countries, faced a major epidemic due to the SARS-COV-2 virus. In this context, the National Institute for Medical Research and the Department of Statistics of the Ministry of Health decided to set up in emergency a new population survey, with the help of the French NSI. The questionnaire covers a wide spectrum of subjects about health and life conditions during the confinement. Here we focus on 8 questions about existence of symptoms of SARS-CoV-2 virus : fever, headache, unusual fatigue, muscle aches or pains, cough, breathing difficulty or unusual shortness of breath, disorders of taste or smell, chest pain or tightness.

Table 1: Horvitz-Thompson estimates depending on the non-response correction model

	model on observables	Heckman's model
	fever (%)	
total	7.6 (0.09)	5.2 (0.24)
mixed-mode	6.6 (0.20)	4.9 (0.56)
online	8.1 (0.11)	5.3 (0.29)
	at least one symptom (%)	
total	26.0 (0.16)	16.1 (0.61)
mixed-mode	23.3 (0.33)	15.0 (1.07)
online	28.4 (0.17)	16.3 (0.56)

Note: Standard deviations in parentheses.These are estimated by bootstrapping in the sample population (respondents and not).

371 000 individuals have been selected in the French sample frame. Because of constraints of cost and availability of interviewers during the confinement, Internet is chosen as the main mode survey. Four fifth of the sample were surveyed online and one fifth online and by phone. These sub-samples are randomly assigned, independently of y. The participation rate for the online sub-sample is 34.4% whereas the participation rate for the mixed-mode sub-sample is 45.9%. Therefore this protocol can be used as an instrumental variable $\mathbf{w}_i$ in the Heckman's model previously presented. Auxiliary variables $\mathbf{z}_i$ used in the Heckman's model are at individual level - as sex, age, type of income, place of birth -, at household level - as standard of living, type of household, type of contact details, owner or tenant - and at geographical level - as size of the urban unit, density, hospitalisation and death rate, type of neighbourhood.

Two models were estimated for two outcome variables that are : existence of fever and the binary variable associated to the existence of at least one symptom. These two outcomes variables behave in a quite similar way as other symptoms (not presented here). Heckman's model show a clear endogenous selection (the hypothesis $H_0 : \varrho = 0$ is always rejected) and positively associated to the level of symptoms ($\hat{\varrho} > 0$) : the more people suffer from symptoms, the more they are willing to participate.

Table 1 gives the Horvitz-Thompson estimates computed for the whole sample for two types of correction for non-response : the usual non-response correction model on observables and the Heckman's model. The estimates given by these two models are significantly different, and the difference is huge : 10 points (ratio 1.6) for the variable "at least one symptom" and 2.4 points (ratio 1.5) for "fever". In both cases, the difference is significant with respect to confidence intervals. If the non-response correction model is correct, we should estimate the same mean with the mixed-mode sub-sample and with the online sub-sample. It is not the case with the usual model on observables. On the contrary, the difference is no longer significant with the Heckman's model. We observe that the Heckman's estimator comes at the cost of a significant loss of precision, by a factor a bit larger than 3.

4 Conclusions

Mixed-mode can be used to properly handle the general problem that arises in all surveys, particularly in online surveys, when people choose themselves to participate depending on the topic of the survey. The usual non response correction model based on observables may not correct for the bias when the resulting selection is endogenous. Heckman's models can be an effective tool to cope with endogenous selection, provided that the protocol is designed to that. The sample design must incorporate the selection of independent sub-samples with different participation rates. Constructed in this way, the protocol generates natural instruments of participation that can be used to identify possible endogenous selection behaviours.
We focus here on the selection issue, rather than the measurement issue. In fact, the analysis of the EpiCov variables of symptoms is conducted under the hypothesis of no measurement error associated to a given mode. This has been shown reasonable in the context. But it remains a strong hypothesis that must be discussed.

References

J.J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

J.-E. Galimard, S. Chevret, E. Curis, and M. Resche-Rigon. Heckman imputation models for binary or continuous mnar outcomes and mar predictors. *BMC medical research methodology*, 18(90), 2018.

L. Castell and P. Sillard. The Treatment of Endogenous Selection Bias in Household Surveys by Heckman's model. Technical Report Mxxx (to be published), Institut National de la Statistique et des Études Économiques, 2020.