The Treatment of Endogenous Selection Bias in Household Surveys by Heckman Model

Documents de travail
N° M2021-02 - March 2021

Laura CASTELL
Patrick SILLARD

Insee
Mesurer pour comprendre

THE NATIONAL INSTITUTE OF STATISTICS AND ECONOMIC STUDIES

Directorate for Methodology and Statistical and International Coordination
Methodology Statistics « Working papers »

M 2021/02

The Treatment of Endogenous Selection Bias in Household Surveys by Heckman Model

Laura CASTELL
Patrick SILLARD

Insee

March 2021

The authors would like to thank Jérôme Accardo, Pascal Ardilly, Pauline Givord, Sylvie Lagarde, Stéfan Lollivier and Amandine Schreiber for their careful review of the document and their valuable advice, as well as the participants in the DMCSI seminars held on 10 September and 19 November 2020.

Directorate for Methodology and Statistical and International Coordination
Statistical Methods Department -Stamp L001 -
88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : -DG75-L001@insee.fr - Website Insee : http://www.insee.fr

The Treatment of Endogenous Selection Bias in Household Surveys by Heckman Model

Laura Castell and Patrick Sillard, 1er septembre 2021

Abstract — **The objective of this working paper is to describe the conditions under which selection bias related to non-response in household surveys can be corrected. Generally, the correction methods implemented assume an ignorable non-response (i.e. missing at random) mechanism. However, when there is an endogenous non-response problem, then the non-response mechanism is no longer ignorable, and the estimators derived from conventional correction methods are biased.**

To correct this bias, we propose a weighting based on a Heckman model. This model consists of simultaneously modelling the participation and the variable of interest that we are trying to estimate. However, the identification of the model is conditional on a certain number of hypotheses, such as the existence of an instrument that explains participation but not the variable of interest. In order to have such an instrument, an adapted protocol with independent sub-samples can be set up. This paper details the conditions under which this type of protocol allows an estimate corrected for endogenous selection.

Keywords : non-response, Heckman model, survey, sampling
Classification JEL : C18, C83, C34, C36

In several household surveys, one can suspect an endogenous selection phenomenon leading to a non-ignorable selection bias. In fact, participation is often linked to the respondents' interest in the subject matter of the survey, particularly in the case of self-administered surveys. If this interest in the topic influences participation and output variables in the survey, conditionally to the observables, a problem of endogenous selection is encountered. In this case, classical correction methods lead to biased estimators. This bias is all the more important when the response rate is low and the omitted variable is correlated with the variables of interest.

To begin with, we recall the analytical framework of the estimation of a variable of interest in the case of a household survey carried out by sampling (part I). We then give the conditions under which the selection linked to ignorable non-response on the one hand (part II) and non-ignorable on the other hand (part III) can lead to a biased estimate. In part IV, we present the method for correcting the endogenous selection proposed from the Heckman model and develop its conditions of application and, more generally, of identification. Finally, part V specifies the relatively rare conditions under which it is possible to separate measurement errors, for example linked to a mode of data collection, and an endogenous selection bias, in the framework of the Heckman model.

I. Notations and general principles

We note y the variable of interest, conceptually collectable on the individuals of a population $\mathscr{P}$. Each individual of the population is identified by its index i. We are interested in the average of this variable in the population :

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1}$$

In a survey, μ cannot be observed, because only some individuals i are actually observed. Let us note s_i the indicator random variable equals 1 if the individual i is sampled, and 0 otherwise.

This variable is therefore a binary variable. The sample design is a vector of random variables

$$\mathbf{s} = \left(s_1, \ldots, s_N\right)'$$

We note $\mathbf{Z}$ a set of variables $\mathbf{z}_i$ known *ex-ante* on the whole population $\mathscr{P}$ (because appearing in the sampling frame). Its distribution $f_{\mathbf{Z}}$ is thus known on $\mathscr{P}$. This variable $\mathbf{Z}$ is used, for example, to stratify the $\mathbf{s}$ sampling design or to balance the first degree in the case of a sampling design with several degrees.

The survey collects the variable of interest y_i (whose vector form on $\mathscr{P}$ is noted $\mathbf{y}$), as well as the characteristic variables of the surveyed individuals $\mathbf{x}_i$ (whose matrix form on $\mathscr{P}$ is noted $\mathbf{X}$). These variables are collected only for the respondents but exist for the whole population $\mathscr{P}$.

The sampling design is set upon the variables $\mathbf{Z}$. Consequently, the sampling design is characterised by a $f_{\mathbf{s}}$ distribution. As the law of $\mathbf{Z}$ is known for the population $\mathscr{P}$, we can deduce $\mathbb{E}(s_i)$, from (27) :

$$\mathbb{E}(s_i) = \mathbb{E}\left[\mathbb{E}(s_i|\mathbf{Z})\right] \tag{2}$$

We note $\mathbb{E}(s_i) = \pi_i$ the probability of inclusion of i in the sample (or probability that i is selected).

The Horvitz-Thompson estimator of μ, based on the design $\mathbf{s}$ is classically :

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{\pi_i} s_i \tag{3}$$

This is observable since y_i is observed as soon as $s_i = 1$.

We will show that $\hat{\mu}$ estimates μ without bias [1]. To justify that the Horvitz-Thompson estimator is convergent, it is advisable to examine $E(\hat{\mu}|\mathbf{y})$ (and not $E(\hat{\mu})$ in the absolute), and to

1. To do this, it is necessary to specify the framework of thought of the theory of surveys. In this framework, the variable $\mathbf{y}$ is not a random variable. It is a set of observable parameters on the people surveyed. Nevertheless, this quantity can be linked, formally, to the random variables characteristic of the survey, in particular of the survey design (i.e. the variable $\mathbf{s}$). To be precise, it can condition these variables. Consequently, to establish the conditions of convergence of the estimators, it is theoretically appropriate to treat $\mathbf{y}$ as a random variable, even if it is an assumption which is not necessary in the framework of the sampling theory. One way out of this debate of principle is to consider that $\mathbf{y}$ is indeed a random variable (e.g. drawn from a "super-population") but that all estimators based on the survey are conditioned by $\mathbf{y}$. This is what we do in this text.

show that this conditional expectation is equal to μ such as defined in (1).

Let's calculate :

$$\begin{aligned}\mathbb{E}(\hat{\mu}|\mathbf{y}) &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\frac{y_i}{\pi_i}s_i|\mathbf{y}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{E}(s_i|y_i)\end{aligned}$$

We have seen that $\mathbf{s}$ only depends on $\mathbf{Z}$, so

(H1) :	$\mathbf{s} \perp\!\!\!\perp \mathbf{y}|\mathbf{Z}$

This is the hypothesis (H1) of this text. It is interesting to stop for a moment on this hypothesis, to specify the underlying mechanism. Little and Rubin (1987) explain that the distribution of sample ($\mathbf{s}$) is set **before** any observation $\mathbf{y}$ is made, then the distribution $f_{\mathbf{s}|\mathbf{y},\mathbf{Z}}$ cannot depend on $\mathbf{y}$ which is unknown to the sampler in the *ex-ante* framework in which he fixes $\mathbf{s}$ and its realisation. Then, $f_{\mathbf{s}|\mathbf{y},\mathbf{Z}} \equiv f_{\mathbf{s}|\mathbf{Z}}$, according to relations (25) and (29).This can also be written: $\mathbf{s} \perp\!\!\!\perp \mathbf{y}|\mathbf{Z}$. Dawid (1979), who introduced this notation, indicates that intuitively, this means that given $\mathbf{Z}$, any information received about $\mathbf{y}$ does not alter uncertainty about $\mathbf{s}$.

It follows, according to (2), the formula of iterated expectations (28) and the consequence, on the conditional expectations, of the relations of orthogonality between variables (31), that :

$$\begin{aligned}\mathbb{E}(s_i|y_i) &= \mathbb{E}\left[\mathbb{E}(s_i|y_i,\mathbf{Z})|y_i\right] \\ &= \mathbb{E}\left[\mathbb{E}(s_i|\mathbf{Z})\right] \\ &= \pi_i\end{aligned}$$

Finally,

$$\mathbb{E}(\hat{\mu}|\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu \tag{4}$$

We note that the condition under which this result is obtained is the knowledge of the law of $\mathbf{Z}$ (i.e. on $\mathscr{P}$) and of that of $\mathbf{s}$ (relation (2)).

II. Missing-At-Random: the standard model

Non-response in surveys is a complementary process to in-sample selection but operates in a similar way. As a complement to selection characterised by the sampling design $\mathbf{s}$, non-response is a binary variable r_i which is 1 if the individual i, selected in the survey, responds to the survey, and 0 otherwise. The response to the survey is therefore characterised by the product of the variables $s_i r_i$. The variables $(\mathbf{y}, \mathbf{X})$ are observed on the sole set $(\mathbf{sry}, \mathbf{srX})$. The uncorrected Horvitz-Thompson estimator

$$\hat{\mu}^0 = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}s_i r_i$$

is biased because $\mathbb{E}(s_i r_i) \neq \pi_i$. Nevertheless, it is possible to derive the assumptions under which it is possible to construct a corrected unbiased Horvitz-thompson estimator.

Quite naturally, we look for an estimator of the following type:

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

where $\hat{\rho}_i$ would model of r_i, with a form under conditions to be precised later on, in such a way that $\mathbb{E}(\hat{\mu}^1|\mathbf{y}) = \mu$. We are going to study these conditions.

As before, the calculation of $\mathbb{E}(\hat{\mu}^1|\mathbf{y})$ will involve a conditioning by the variable $\mathbf{Z}$, i.e. by the variables available on the whole sample $\mathbf{s}$. In fact, to identify a model of $\mathbf{r}$, we cannot be satisfied with working only on the observables, because in this case, all r_i are equal to 1. There would therefore be no way of identifying such a model.

By construction,

$$\mathbb{E}(y_i s_i r_i|\mathbf{y},\mathbf{Z}) = y_i\mathbb{E}(s_i r_i|\mathbf{y},\mathbf{Z})$$

For the same reasons as those used to justify the hypothesis (H1), s_i is determined *ex-ante* by the knowledge of $\mathbf{Z}$. In this context, the knowledge of r_i does not bring anything to that of s_i (and vice versa by virtue of the symmetry of the $\perp\!\!\!\perp$ – see Appendix A). This results in the following hypothesis :

(H2) :	$s_i \perp\!\!\!\perp r_i|(\mathbf{y},\mathbf{Z})$

Under this assumption and applying the relation (30), it is possible to separate the contributions in the previous expression:

$$\mathbb{E}(y_i s_i r_i|\mathbf{y},\mathbf{Z}) = y_i\mathbb{E}(s_i|\mathbf{y},\mathbf{Z})\mathbb{E}(r_i|\mathbf{y},\mathbf{Z})$$

Using, in addition, the hypothesis (H1), it comes:

$$\mathbb{E}(y_i s_i r_i|\mathbf{y},\mathbf{Z}) = y_i\mathbb{E}(s_i|\mathbf{Z})\mathbb{E}(r_i|\mathbf{y},\mathbf{Z})$$

Let us assume that we are able to model r_i by an unbiased and convergent $\hat{\rho}_i$ estimator.

To fix the ideas, $\hat{\rho}_i$, in this scheme, is obtained by regression (linear probability model for example) of r_i on y_i and $\mathbf{z}_i$. This regression is problematic for the following two reasons. First, this regression uses y_i as an explanatory variable. This is reasonable because it is possible that the propensity r_i to answer is explained by y_i, or by a variable correlated with it. This is, at least in the general case, a hypothesis that cannot be excluded. However, the observed sample, which would be used here to identify the r_i model, would be such that for all i, $r_i = 1$, since y_i is only observed when $r_i = 1$. Such a model would not be identifiable.

Once this observation has been made, we can imagine, in a second step, predicting r_i by a truncated model, conditional on $\mathbf{Z}$. But if y_i really appears as an explanatory factor of r_i, then y_i appears as a variable omitted in the model, and whose absence leads to a biased $\hat{\rho}_i$ estimator. In short, the $\hat{\rho}$ model is identifiable and unbiased only when the following double hypothesis is made :

(H3)	
	• $r_i \perp\!\!\!\perp y_i|\mathbf{Z}$
	• $\hat{\rho}_i \equiv \rho(\mathbf{z}_i;\hat{\gamma})$ where
	– ρ is a known function, continuous in $\mathbf{z}_i$ and $\hat{\gamma}$
	– $\hat{\gamma}$ is a convergent estimator of unknown parameters γ^*(i.e. $\operatorname{plim}\hat{\gamma} = \gamma^*$); $\hat{\gamma}$ depends on $\mathbf{y}$ and $\mathbf{Z}$
	– $\rho(\mathbf{z}_i;\gamma^*) = \mathbb{E}(r_i|\mathbf{Z})$

Under the previous assumptions (H1), (H2) and (H3), the corrected Horvitz-Tompson estimator $\hat{\mu}^1$ estimates μ asymptotically unbiased (in $\hat{\gamma}$). Indeed,

$$\begin{aligned}\mathbb{E}(\hat{\mu}^1|\mathbf{y}) &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(\pi_i\hat{\rho}_i)^{-1}y_i s_i r_i|\mathbf{y}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\{\mathbb{E}\left[(\pi_i\hat{\rho}_i)^{-1}y_i s_i r_i|\mathbf{y},\mathbf{Z}\right]\Big|\mathbf{y}\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{E}\left\{\mathbb{E}\left[\hat{\rho}_i^{-1}s_i r_i|\mathbf{y},\mathbf{Z}\right]\Big|\mathbf{y}\right\}\end{aligned}$$

By hypothesis, $\hat{\rho}_i^{-1}$ depends only on $\mathbf{Z}$ and $\mathbf{y}$, through $\hat{\gamma}$. Therefore,

$$\begin{aligned}\mathbb{E}(\hat{\mu}^1|\mathbf{y}) &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{E}\left\{\hat{\rho}_i^{-1}\mathbb{E}(s_i r_i|\mathbf{Z},\mathbf{y})\Big|\mathbf{y}\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{E}\left\{\hat{\rho}_i^{-1}\mathbb{E}(s_i|\mathbf{Z})\mathbb{E}(r_i|\mathbf{Z})\Big|\mathbf{y}\right\}\end{aligned}$$

Then, given (H3), $\operatorname{plim}(\hat{\rho}_i^{-1}) = [\mathbb{E}(r_i|\mathbf{Z})]^{-1}$. It follows that, asymptotically, $\mathbb{E}(\hat{\mu}^1|\mathbf{y}) = \mu$.

The previous development deals with the *Missing completely at random* (MCAR) and *Missing at random* (MAR) cases in the sense of Little and Rubin (1987). The first corresponds to the case where $\hat{\rho}_i \equiv \rho$, with ρ a constant. The probability of responding is independent of $\mathbf{Z}$ and $\mathbf{y}$. Therefore, the Hajek estimator (Tillé 2019), defined by:

$$\hat{\mu}_H^0 = \sum_{i=1}^{N}\frac{y_i}{\pi_i}s_i r_i \Big/ \sum_{i=1}^{N}\frac{1}{\pi_i}s_i r_i \tag{5}$$

estimates μ asymptotically unbiased.
The second corresponds to the case where it is possible to construct an unbiased estimator of r_i from the observables $\mathbf{z}_i$, as assumed in (H3). The probability of responding is independent of $\mathbf{y}$, but not of $\mathbf{Z}$.

Let us now return to the $\hat{\rho}_i$ non-response model. Like all models, it is based on the identification of unknown parameters γ^*. As indicated by the hypothesis (H3), we can write :

$$\begin{cases} r_i = \rho(\mathbf{z}_i;\hat{\gamma}) + \nu_i \\ \text{with}\ \ \mathbb{E}(\nu_i|\mathbf{z}_i) = 0 \end{cases} \tag{6}$$

In practice, however, (H3), like (6) which is analogous, can be complicated to justify. We have so far assumed that it is verified. We will now study some deviations from this assumption.

III. Non-Missing-At-Random

In this paragraph we propose to examine what happens when a variable, although explanatory of the participation in the survey r_i, is omitted in the expression of $\hat{\rho}_i$. Suppose that the variable ξ_i explains r_i but that this dependence is omitted in the modelling :

$$r_i = c + \mathbf{z}_i\beta + \xi_i + u_i \tag{7}$$

with $\mathbb{E}(u_i|\mathbf{Z}) = 0$, while the applied model is:

$$\tilde{\rho}_i = \tilde{c} + \mathbf{z}_i\tilde{\beta} + \tilde{u}_i \tag{8}$$

Throughout this paragraph III, for the sake of simplicity and readability, we adopt the linear dependency framework corresponding to the two previous relations (7-8).

First of all, let us note that it is possible that $\tilde{\rho}_i$ follows the assumption (H3), despite the omission of ξ_i. Indeed, from the two preceding relations, we derive :

$$\begin{cases} \mathbb{E}(r_i|\mathbf{Z}) &= c + \mathbf{z}_i\beta + \mathbb{E}(\xi_i|\mathbf{z}_i) + \mathbb{E}(u_i|\mathbf{z}_i) \\ \mathbb{E}(\tilde{\rho}_i|\mathbf{Z}) &= \tilde{c} + \mathbf{z}_i\tilde{\beta} + \mathbb{E}(\tilde{u}_i|\mathbf{z}_i) \end{cases}$$

Now, by hypothesis, $\mathbb{E}(u_i|\mathbf{z}_i) = 0$. Consequently, as soon as $\mathbb{E}(\xi_i|\mathbf{z}_i) = 0$, it is possible to obtain, for example by linear regression of r_i on $\mathbf{z}_i$, an estimator $\tilde{\rho}_i$ verifying the second condition of (H3), while omitting ξ_i in the model. The first condition of the hypothesis (H3) can also be verified by projecting this condition onto ξ_i, using the relation (7). Thus, any ξ_i variable verifying :

$$\begin{cases} \xi_i \perp\!\!\!\perp y_i|\mathbf{Z} \\ \mathbb{E}(\xi_i|\mathbf{Z}) = 0 \end{cases}$$

does not cause an omission problem, i.e. it can be omitted in the nonresponse model without generating any (asymptotic) bias in the Hajek estimator.

Let us now look at the case of an omitted variable that does not meet either of the two previous conditions. For example, let's say:

$$\left|\begin{array}{l} \xi_i = \kappa + \vartheta y_i + \mathbf{z}_i\theta + \upsilon_i \\ \mathbb{E}(\upsilon_i|y_i,\mathbf{z}_i) = 0 \end{array}\right. \tag{9}$$

One can imagine two types of problem:

— a problem of endogeneity of ξ_i in the modelling of $\tilde{\rho}_i$ which biases the estimation of the model coefficients (8). This corresponds to the case where $\theta \neq 0$ and $\vartheta = 0$ in the expression of ξ_i above.

— an endogenous self-selection problem in which y_i is at the same time a variable of interest and an explanatory variable of the non-response. This corresponds to the case where $\theta = 0$ and $\vartheta \neq 0$.

Of course, the two types of problem are likely to overlap, but their consequences are very different. And for presentation purposes, it is easier to separate them.

Let us first consider the case where ξ_i is endogenous in the non-response model (i.e. $\theta \neq 0$ and $\vartheta = 0$). This is the case if we estimate $\tilde{\beta}$, in the framework of a linear probability model, by linear regression of r_i on $\mathbf{z}_i$. Classically,

$$\operatorname{plim}\begin{pmatrix}\tilde{c}\\ \tilde{\beta}\end{pmatrix} = \left(\mathbb{E}\left[\begin{pmatrix}1\\ \mathbf{z}_i'\end{pmatrix}\begin{pmatrix}1 & \mathbf{z}_i\end{pmatrix}\right]\right)^{-1}.\mathbb{E}\left[\begin{pmatrix}1\\ \mathbf{z}_i'\end{pmatrix}r_i\right] \tag{10}$$

Using (7) and (9), we note that r_i can also be written as:

$$r_i = \begin{pmatrix}1 & \mathbf{z}_i\end{pmatrix}.\left[\begin{pmatrix}c\\ \beta\end{pmatrix} + \begin{pmatrix}\kappa\\ \theta\end{pmatrix}\right] + u_i + \upsilon_i$$

Substituting this last expression for r_i in (10), and noting that by hypotheses, $\mathbb{E}(u_i|\mathbf{z}_i) = 0$ and $\mathbb{E}(\upsilon_i|\mathbf{z}_i) = 0$, we conclude that:

$$\operatorname{plim}\begin{pmatrix}\tilde{c}\\ \tilde{\beta}\end{pmatrix} = \begin{pmatrix}c\\ \beta\end{pmatrix} + \begin{pmatrix}\kappa\\ \theta\end{pmatrix}$$

And then:

$$\begin{aligned}\mathbb{E}(\tilde{\rho}_i|\mathbf{Z}) &= \begin{pmatrix}1 & \mathbf{z}_i\end{pmatrix}\begin{pmatrix}c+\kappa\\ \beta+\theta\end{pmatrix}\\ &= c + \mathbf{z}_i\beta + \underbrace{\kappa + \mathbf{z}_i\theta}_{\mathbb{E}(\xi_i|\mathbf{Z})}\\ &= \mathbb{E}(r_i|\mathbf{Z})\end{aligned}$$

In the end, the endogeneity of ξ_i in the modelling of r_i is not critical since, if it biases the regression coefficients, it does not however bias the $\tilde{\rho}_i$ predictors which result from this estimation, as estimators of r_i. In this case, the non-response mechanism remains ignorable, despite the omission of ξ_i.

Let us now consider the case of an endogenous selection, where ξ_i depends on y_i (i.e. $\theta = 0$ and $\vartheta \neq 0$). The $\tilde{\rho}_i$, obtained by regression of r_i on $\mathbf{z}_i$ estimates without bias r_i, conditional to $\mathbf{Z}$. But we cannot identify the dependence of r_i on y_i by regressing r_i on $(\mathbf{z}_i, y_i)$ since y_i is only observed for the respondents, i.e. the i for which $r_i = 1$. Thus the identifying regression for $\tilde{\rho}_i$ is based on the $\mathbf{z}_i$ alone. If the second part of the hypothesis (H3) is valid, the first part, on the other hand, is no longer verified since r_i depends on y_i via ξ_i. In this case, the non-response mechanism is then non-ignorable.

In the development of $\mathbb{E}(\hat{\mu}^1|\mathbf{y})$, the first part remains true under the assumptions (H1) and (H2) which are, themselves, verified. Therefore, under these assumptions,

$$\begin{aligned}\mathbb{E}(\hat{\mu}^1|\mathbf{y}) &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{E}(s_i|\mathbf{y})\mathbb{E}\left\{\mathbb{E}\left[\tilde{\rho}_i^{-1}r_i|y_i,\mathbf{z}_i\right]\middle|\, y_i\right\}\\ &= \frac{1}{N}\sum_{i=1}^{N} y_i\mathbb{E}\left\{\mathbb{E}\left[\tilde{\rho}_i^{-1}r_i|y_i,\mathbf{z}_i\right]\middle|\, y_i\right\}\end{aligned} \tag{11}$$

Using (7 – 9), we then have :

$$\mathbb{E}\left[\tilde{\rho}_i^{-1}r_i\middle|\, y_i,\mathbf{z}_i\right] = \mathbb{E}\left[\frac{c+\mathbf{z}_i\beta+u_i+\kappa+\vartheta y_i+\upsilon_i}{\tilde{c}+\mathbf{z}_i\tilde{\beta}}\middle|\, y_i,\mathbf{z}_i\right]$$

For the same reasons that lead to the relation (10), in a linear probability model,

$$\operatorname{plim}(\tilde{c}+\mathbf{z}_i\tilde{\beta}) = c+\kappa+\mathbf{z}_i\beta$$

Here we can make the reasonable complementary assumption that $u_i \perp\!\!\!\perp y_i|\mathbf{z}_i$, which is equivalent to consider that all the dependence of r_i on y_i passes through ξ_i. All this is essentially formal since these are assumptions about the structure of r_i and the additive separation of its different components. It follows from this additional assumption and the above that, asymptotically:

$$\mathbb{E}\left[\tilde{\rho}_i^{-1}r_i\middle|\, y_i,\mathbf{z}_i\right] = 1+\vartheta\mathbb{E}\left[\frac{y_i}{c+\kappa+\mathbf{z}_i\beta}\middle|\, y_i,\mathbf{z}_i\right] = 1+\vartheta\frac{y_i}{\tilde{\rho}_i(\mathbf{z_i})}$$

Finally (asymptotically),

$$\mathbb{E}(\hat{\mu}^1|\mathbf{y}) = \mu + \vartheta\frac{1}{N}\sum_{i=1}^{N} y_i^2\mathbb{E}\left[\,1/\tilde{\rho}_i(\mathbf{z}_i)\,|\,y_i\right] \tag{12}$$

The bias linked to the existence of an endogenous selection is therefore of the sign of the dependence of participation on the variable of interest (i.e. of ϑ) : if participation increases with the variable of interest, $\hat{\mu}^1|\mathbf{y}$ is positively biased ; the latter is negatively biased if participation decreases with the variable of interest. And all other things being equal, the bias increases with the variance of the variable of interest on $\mathscr{P}$.

IV. Correction of endogenous non-response

The correction of endogenous non-response is a known problem in the exploitation of missing data (Little and Rubin 1987). It has been the subject of numerous econometric developments aimed at fitting a suitable model of the variable of interest (see for example Boutchenik, Coudin, and Maillard (2019)). In doing so, these developments are also interesting for survey analysis since they allow in particular to estimate $\mathbb{E}(y)$ without bias (see for example Ardilly (2006)). Two main classes of methods can be distinguished in this context. The latent participation variable methods, as in Heckman's models. More recently, these methods have been extended and developed further (Vella 1998, Galimard, Chevret, Curis, and Resche-Rigon 2018, Wing 2019), including towards non-parametric modelling, for example in the field of treatment evaluation with endogenous selection (see for example the article by Lee (2009)). Generalized calibration methods have also been developed, based on different identification conditions (see for example the article by Lesage, Haziza, and D'Haultfœuille (2019)).

The remainder of this text is devoted to the presentation of the use of the classical Heckman model for the treatment of endogenous non-response in surveys.

A. The framework of the Heckman model

The Heckman model was popularised by the econometer of the same name (Heckman 1979). It is also known as the *Tobit II* model (Wooldridge 2010, Cameron and Trivedi 2005). It consists in modelling simultaneously y_i and r_i in the following way:

$$\begin{cases}\text{(i)} & y_i &= c^1 + \mathbf{z}_i\chi + \epsilon_i^1\\ \text{(ii)} & r_i^* &= c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0\\ \text{(iii)} & r_i &= \mathbb{1}(r_i^* \geqslant 0)\end{cases} \tag{13}$$

r_i^* is a latent variable that is not observed. We observe r_i and y_i when $r_i = 1$. $(\mathbf{z}_i, \mathbf{w}_i)$ is observed for all i.

$(\epsilon_i^0, \epsilon_i^1)$ are unknown parameters. In this model, the participation equation (13-(iii)) is based on a latent variable r_i^* (relation (13-(ii)) which involves the explanatory variables of y_i (here on the whole sample, respondents and non-respondents). This latent variable also involves $\mathbf{w}_i$ instruments, i.e. variables that explain participation but are not explanatory of y_i. These are referred to as exclusion conditions. Formally, the identification conditions of the previous model are:

$$\begin{cases}\mathbb{E}\left(\begin{pmatrix}\epsilon_i^0\\ \epsilon_i^1\end{pmatrix}\middle|\,\mathbf{z}_i,\mathbf{w}_i\right) = 0\\ \begin{pmatrix}\epsilon_i^0\\ \epsilon_i^1\end{pmatrix} \hookrightarrow \mathscr{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix},\Sigma\right)\\ \Sigma = \begin{pmatrix}1 & \varrho\sigma\\ \varrho\sigma & \sigma^2\end{pmatrix}\end{cases} \tag{14}$$

Σ is a variance matrix. It is therefore through this matrix that the simultaneous formation of the explained variable y_i and the participation r_i is modelled. In practice and in the model (13), not all the parameters of Σ can be identified. As in a probit model, it is appropriate to adopt a unit variance for ϵ_i^0 since the coefficients of (13-(ii)) are identifiable down to a multiplicative factor. Hence, the form proposed for Σ above is the most general possible in the context of the Heckman model [2].

Let us note that this model concerns the whole population $\mathscr{P}$. Apart from the selection linked to the sampling ($s_i = 1$) which we have shown to have no effect on the expectation of the estimators, we only observe the respondents, i.e. the i such that $r_i = 1$. It may be interesting to study, in this model, how $(\mathbf{ry})$ behaves. Let us thus calculate $\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 1)$. Under the previous assumptions, we have (see Appendix A):

$$\begin{aligned}\mathbb{E}(y_i|\mathbf{z}_i, \epsilon_i^0) &= c^1 + \mathbf{z}_i\chi + \mathbb{E}(\epsilon_i^1|\epsilon_i^0) \\ &= c^1 + \mathbf{z}_i\chi + \varrho\sigma\epsilon_i^0\end{aligned}$$

It follows that:

$$\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 1) = c^1 + \mathbf{z}_i\chi + \varrho\sigma\mathbb{E}(\epsilon_i^0|\epsilon_i^0 \geqslant -c^0 - \mathbf{z}_i\beta - \mathbf{w}_i\psi)$$

ϵ_i^0 is a Gaussian variable, so the expression of $\mathbb{E}(\epsilon_i^0|\epsilon_i^0 \geqslant -a)$ is a known function $\lambda(a)$, corresponding to the inverse of the Mills ratio [3]. Finally,

$$\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 1) = c^1 + \mathbf{z}_i\chi + \varrho\sigma\lambda(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi) \quad (15)$$

We show in the same way that [4] :

$$\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 0) = c^1 + \mathbf{z}_i\chi - \varrho\sigma\lambda(-(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi))$$

We note that:

$$\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 1) - \mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 0) = \varrho\sigma\left[\lambda(\breve{r}_i^*) + \lambda(-\breve{r}_i^*)\right] \quad (16)$$

where $\breve{r}_i^* = c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi$. Thus defined, $\breve{r}_i^*$ is comparable, in the reasoning, to a predictor of r_i^*.

We first observe, from the previous expression, that if $\varrho = 0$, then $\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 1) = \mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 0)$. Thus, the endogeneity of the selection occurs in the Heckman model when the correlation of the residuals $(\epsilon_i^0, \epsilon_i^1)$ is non-zero. And conversely, there is no endogeneity of selection when $\varrho = 0$. In the expression (16), the term between square brackets is positive since the function λ is. It follows that if $\varrho > 0$, then the observed y_i are, all other things being equal, larger on average than the unobserved y_i. Conversely, if $\varrho < 0$, then the observed y_i are, all other things being equal, smaller, on average, than the unobserved y_i.

The resolution of the Heckman model allows us to construct two estimators : one by imputation of y_i for non-respondents, the other by reweighting.

First of all, the knowledge of $\mathbb{E}(y_i|\mathbf{z}_i, r_i = 0)$ allows us to build a new estimator of μ free of the endogenous selection step r_i, by imputing the y_i of the non-respondents, in the following way:

$$\left\{\begin{array}{ll} \hat{\mu}_H^{0H} & = \displaystyle\sum_{i=1}^N \frac{\hat{y}_i^H}{\pi_i}s_i \Big/ \sum_{i=1}^N \frac{1}{\pi_i}s_i \\ \text{where} & \left|\begin{array}{l} \hat{y}_i^H(r_i = 0) = \hat{c}^1 + \mathbf{z}_i\hat{\chi} - \hat{\varrho}\hat{\sigma}\lambda(-(\hat{c}^0 + \mathbf{z}_i\hat{\beta} + \mathbf{w}_i\hat{\psi})) \\ \hat{y}_i^H(r_i = 1) = y_i \end{array}\right. \end{array}\right. \quad (17)$$

The parameters $(\hat{c}^1, \hat{\beta}, \hat{\psi}, \hat{c}^0, \hat{\chi}, \hat{\varrho}, \hat{\sigma})$ are estimated on the basis of the model (13). The methods for estimating them are discussed in the following paragraph.

We can also derive another μ estimator by basing it on the conditional inclusion probability from the Heckman model. Let us start again from the expression (11) :

$$\mathbb{E}(\hat{\mu}^1|\mathbf{y}) = \frac{1}{N}\sum_{i=1}^N y_i\mathbb{E}\left\{\mathbb{E}\left[\tilde{\rho}_i^{-1}r_i|y_i, \mathbf{z}_i\right]\Big|\, y_i\right\}$$

One can here make use of $\mathbf{w}_i$ as an additional conditioning variable. In this way, if it is possible to construct a convergent estimator $\tilde{\rho}_i$ of $\mathbb{E}[r_i|y_i, \mathbf{z}_i, \mathbf{w}_i]$. Then, as before (cf. II), we will be able to construct an asymptotically unbiased estimator of $\mathbb{E}(\hat{\mu}^1|y)$.

In order to construct $\tilde{\rho}_i$, let us calculate $\mathbb{E}(r_i|y_i, \mathbf{z}_i, \mathbf{w}_i)$ from the model (13). $(r_i|y_i, \mathbf{z}_i, \mathbf{w}_i)$ is a binary variable, so $\mathbb{E}(r_i|y_i, \mathbf{z}_i, \mathbf{w}_i) = \Pr(r_i^* \geqslant 0|y_i, \mathbf{z}_i, \mathbf{w}_i)$. Under the previous hypotheses, the law of $(r_i^*|y_i, \mathbf{z}_i, \mathbf{w}_i)$ is known. Indeed [5],

$$\mathscr{L}(\epsilon_i^0|y_i, \mathbf{z}_i, \mathbf{w}_i) = \mathscr{L}(\epsilon_i^0|\epsilon_i^1 = y_i - c^1 - \mathbf{z}_i\chi)$$

And $(\epsilon_i^0, \epsilon_i^1) \hookrightarrow \mathscr{N}\left(0, \Sigma = \begin{pmatrix} 1 & \varrho\sigma \\ \varrho\sigma & \sigma^2 \end{pmatrix}\right)$, then the conditional law $(\epsilon_i^0|\epsilon_i^1)$ is a normal law. This leads to [6] :

$$\begin{aligned}&\mathscr{L}(r_i^*|y_i, \mathbf{z}_i, \mathbf{w}_i) \\ &\quad = \mathscr{L}(\epsilon_i^0 + c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi|\epsilon_i^1 = y_i - c^1 - \mathbf{z}_i\chi) \\ &\quad = \mathscr{N}\left(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \frac{\varrho}{\sigma}(y_i - c^1 - \mathbf{z}_i\chi); (1 - \varrho^2)\right)\end{aligned}$$

It follows that:

$$\begin{aligned}&\Pr(r_i^* \geqslant 0|y_i, \mathbf{z}_i, \mathbf{w_i}) \\ &\quad = \Phi\left(\frac{c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \frac{\varrho}{\sigma}(y_i - c^1 - \mathbf{z}_i\chi)}{\sqrt{1 - \varrho^2}}\right)\end{aligned} \quad (18)$$

We deduce the expression of $\tilde{\rho}_i$:

$$\tilde{\rho}_i = \Phi\left(\frac{\hat{c}^0 + \mathbf{z}_i\hat{\beta} + \mathbf{w}_i\hat{\psi} + \frac{\hat{\varrho}}{\hat{\sigma}}(y_i - \hat{c}^1 - \mathbf{z}_i\hat{\chi})}{\sqrt{1 - \hat{\varrho}^2}}\right) \quad (19)$$

where, as for the estimator (17), the parameters $(\hat{c}^1, \hat{\beta}, \hat{\psi}, \hat{c}^0, \hat{\chi}, \hat{\varrho}, \hat{\sigma})$ are estimated under the model (13). The methods for estimating them are discussed in the following paragraph.

(19) gives the expression of $\tilde{\rho}_i$. We observe that it depends only on the variables y_i, $\mathbf{z}_i$ and $\mathbf{w}_i$ and on parameters for

2. and for independent residuals between individuals, i.e.: $\forall i \neq j$, $(\epsilon_i^0, \epsilon_i^1) \perp\!\!\!\perp (\epsilon_j^0, \epsilon_j^1)$.

3. as long as $\text{var}(\epsilon_i^0) = 1$. Let us recall that the inverse of the Mills' ratio is defined by: $\lambda(a) = \varphi(a)/\Phi(a)$ where φ is the density of the zero-mean and unit-variance normal law and Φ is its cumulative distribution. It is a strictly decreasing function, positive, and has asymptotes $y = -x$ in $-\infty$ and $y = 0$ in $+\infty$.

4. $\mathbb{E}(\epsilon_i^0|\epsilon_i^0 \leqslant a) = -\lambda(a)$ (Cameron and Trivedi 2005, p. 540).

5. $\mathscr{L}$ denotes the law of the random variable in argument.

6. If $\begin{pmatrix} X \\ Y \end{pmatrix} \hookrightarrow \mathscr{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_X\sigma_Y\varrho \\ \sigma_X\sigma_Y\varrho & \sigma_Y^2 \end{pmatrix}\right)$ is a bivariate vector of normal random variables, then $\mathscr{L}(X|Y = y) = \mathscr{N}\left(\mu_X + \varrho\sigma_X(y - \mu_Y)/\sigma_Y; \sigma_X^2(1 - \varrho^2)\right)$.

which we can construct convergent estimators. Under the hypotheses of the Heckman model (13 – 14), thus defined, $\tilde{\rho}_i$ estimates asymptotically unbiased $\mathbb{E}(r_i|y_i, \mathbf{z}_i, \mathbf{w_i})$. It follows, and for the same reasons as those given in paragraph II, that

$$\hat{\mu}_H^{1H} = \sum_{i=1}^{N} \frac{y_i}{\tilde{\rho}_i \pi_i} s_i r_i \Bigg/ \sum_{i=1}^{N} \frac{1}{\tilde{\rho}_i \pi_i} s_i r_i \tag{20}$$

with $\tilde{\rho}_i$ defined in the relation (19), is an asymptotically unbiased estimator of μ.

It is also possible to treat the case of binary variables with a Heckman model using a latent variable for the variable of interest. The modified model is the following:

$$\begin{cases} (0) & y_i & = & \mathbb{1}(y_i^* \geqslant 0) \\ (\text{i}) & y_i^* & = & c^1 + \mathbf{z}_i\chi + \epsilon_i^1 \\ (\text{ii}) & r_i^* & = & c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0 \\ (\text{iii}) & r_i & = & \mathbb{1}(r_i^* \geqslant 0) \end{cases} \tag{21}$$

The conditions of identification (i.e. exclusion) remain identical to the continuous case (14), except that we can henceforth pose $\sigma = 1$, because the coefficients of (21-(i)) are henceforth identified down to a multiplicative factor. We observe that:

$$\mathscr{L}(y_i^*, r_i^*|\mathbf{z}_i, \mathbf{w}_i) = \mathscr{N}\left[\begin{pmatrix} c^1 + \mathbf{z}_i\chi \\ c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi \end{pmatrix}, \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}\right]$$

Then,

$$\begin{aligned} \mathbb{P}(y_i = 1|\mathbf{z}_i, \mathbf{w}_i, r_i = 0) &= \mathbb{P}(y_i^* \geqslant 0|\mathbf{z}_i, \mathbf{w}_i, r_i^* \leqslant 0) \\ &= \frac{\mathbb{P}(y_i^* \geqslant 0, r_i^* \leqslant 0|\mathbf{z}_i, \mathbf{w}_i)}{\mathbb{P}(r_i^* \leqslant 0|\mathbf{z}_i, \mathbf{w}_i)} \end{aligned}$$

Finally, it can be deduced that [7] :

$$\begin{aligned} &\mathbb{P}(y_i = 1|\mathbf{z}_i, \mathbf{w}_i, r_i = 0) \\ &\quad = \frac{\Phi_2(c^1 + \mathbf{z}_i\chi, -(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi); -\varrho)}{\Phi(-(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi))} \end{aligned} \tag{22}$$

where Φ_2 denotes the distribution of the bivariate normal distribution (see footnote n.7). This relation is the one recommended by Galimard, Chevret, Curis, and Resche-Rigon (2018) to impute the predicted values for the non-respondents, in accordance with the approach proposed for the estimator (17). Concretely, the previous authors propose to impute the response of the non-respondents by drawing this response in a Bernouilli distribution of parameter $\mathbb{P}(y_i = 1|\mathbf{z}_i, \mathbf{w}_i, r_i = 0)$, as given in the expression (22).

As in the continuous case, it is possible to construct a corrected Hajek estimator analogous to (20) using the response probabilities, conditional on y_i. These conditional probabilities are:

$$\begin{aligned} &\mathbb{P}(r_i = 1|\mathbf{z}_i, \mathbf{w}_i, y_i = 1) \\ &\quad = \frac{\mathbb{P}(r_i^* \geqslant 0, y_i^* \geqslant 0|\mathbf{z}_i, \mathbf{w}_i)}{\mathbb{P}(y_i^* \geqslant 0|\mathbf{z}_i, \mathbf{w}_i)} \\ &\quad = \frac{\Phi_2(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi, c^1 + \mathbf{z}_i\chi; \varrho)}{\Phi(c^1 + \mathbf{z}_i\chi)} \end{aligned}$$

and

$$\begin{aligned} &\mathbb{P}(r_i = 1|\mathbf{z}_i, \mathbf{w}_i, y_i = 0) \\ &\quad = \frac{\mathbb{P}(r_i^* \geqslant 0, y_i^* \leqslant 0|\mathbf{z}_i, \mathbf{w}_i)}{\mathbb{P}(y_i^* \leqslant 0|\mathbf{z}_i, \mathbf{w}_i)} \\ &\quad = \frac{\Phi_2(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi, -(c^1 + \mathbf{z}_i\chi); -\varrho)}{\Phi(-(c^1 + \mathbf{z}_i\chi))} \end{aligned}$$

These probabilities are used to calculate the weights to be assigned to each observation in a Hajek estimator, depending on whether the variable of interest y_i is 0 or 1.

B. Estimation and discussion of the Heckman model

The continuous Heckman model (13) can be estimated in two different ways (Cameron and Trivedi 2005) :

- in two steps by first determining the Probit model associated with the relation (iii) which leads to unbiased estimators of the coefficients $(\hat{c}^0, \hat{\beta}, \hat{\psi})$; then by regressing y_i on the variables $\left[1, \mathbf{z}_i, \lambda(\hat{c}^0 + \mathbf{z}_i\hat{\beta} + \mathbf{w}_i\hat{\psi})\right]$ which leads to the unbiased estimators of the coefficients $(\hat{c}^1, \hat{\chi}, \hat{\varrho})$.
- by maximum likelihood, the likelihood associated with (13) being expressible analytically (Cameron and Trivedi 2005).

This second solution is more efficient statistically. On the other hand, the first one is faster in computation time (see Appendix C). And it is not necessary to assume that ϵ_i^1 is normal for the estimate to be unbiased. The two-step method is therefore less parametric than the one-step maximum likelihood method. In the case of a binary variable of interest (21), only the maximum likelihood estimation is feasible (see for example Galimard, Chevret, Curis, and Resche-Rigon (2018)). The analytical expression, in the binary case, is given by Cameron and Trivedi (2005).

The package R sampleSelection allows to estimate in a very convenient way a Heckman model and the various estimators developed in the preceding part. Simulation results are presented in appendix C. They show, for an example of simulated income variable, with endogenous selection for high incomes, that the Hajek estimator can be strongly biased and that a Heckman estimator allows, in this case, to correct this bias.

The model (13) is theoretically identifiable without instruments $\mathbf{w}_i$ in the relation (ii). The same is true for the model (21). However, the presence of the $\mathbf{w}_i$ instruments is a condition *sine qua non* for the good convergence of the estimation methods. In the absence of instruments explaining participation without affecting the variable of interest, in the continuous case, there is quasi-linearity in the relationship (i) between the inverse of the Mills ratio $\lambda(\hat{c}^0 + \mathbf{z}_i\hat{\beta})$ and $\mathbf{z}_i$, so that the estimation of the Heckman model is likely to fail

7. If $\begin{pmatrix} X \\ Y \end{pmatrix} \hookrightarrow \mathscr{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}\right)$ is a bivariate vector of normal variables, then $\mathbb{P}(X \geqslant 0, Y \geqslant 0) = \Phi_2(\mu_X, \mu_Y; \varrho)$, $\mathbb{P}(X \geqslant 0, Y \leqslant 0) = \Phi_2(\mu_X, -\mu_Y; -\varrho)$, $\mathbb{P}(X \leqslant 0, Y \geqslant 0) = \Phi_2(-\mu_X, \mu_Y; -\varrho)$ et $\mathbb{P}(X \leqslant 0, Y \leqslant 0) = \Phi_2(-\mu_X, -\mu_Y; \varrho)$, where $\Phi_2(x, y; \varrho)$ denotes the distribution function of the zero-mean and unit-variance bivariate normal law, with correlation ϱ.

to converge, especially in the case of maximum likelihood estimation. Similar phenomena occur in the discrete case. A suitable survey protocol can provide such an instrument.

It is useful to note that the instrument $\mathbf{w}$ of the selection equation (13-ii) or (21-ii) implies that the monotonicity principle of the instrument (Imbens and Angrist 1994, Vytlacil 2002) is verified by the survey protocol chosen. Let us take a closer look at this point when the instrument is a binary variable. Considering these two selection equations, $r_i^*(\mathbf{w}_i = 1) - r_i^*(\mathbf{w}_i = 0) = \psi$, for all i. Now ψ is a constant, either positive or negative. Let us suppose for example that ψ is positive. Then, any individual participating in the survey in the absence of an instrument (i.e. such that $\mathbf{w}_i = 0$ and $r_i^* \geqslant 0$) would necessarily have participated in the survey in the presence of the instrument (i.e. if, instead of $\mathbf{w}_i = 0$, the value of the instrument concerning him had been $\mathbf{w}_i = 1$). Since ψ is either positive or negative, participation in the survey is, for any i, either increasing with the instrument or decreasing: it is therefore monotonous. The important point is that this property is to be understood "all other things being equal": it is not only on average that individuals must participate more according to whether they are in one of the groups defined by the instrument, but for each individual.

The necessary monotonicity of the instrument therefore has consequences for the protocol conditions under which a Heckman model can be applied. In practice, it is therefore necessary that the selected protocol makes it possible to justify that individuals in the group with the lowest response rate who actually participate in the survey would also have participated if they had benefited from the alternative protocol, as characterised by the instrument.

Such instruments can be used in random surveys when, for example, two sub-samples, administered according to two different collection protocols, have been selected, one of which results in a higher participation than the other. In these circumstances, the two samples are combined to form a single random sample and the subsample indicator is an instrument. Indeed, insofar as participation differs between the two protocols, the indicator variable of belonging to one of the two subsamples is explanatory of participation (due to the random selection of one sample over the other - mathematically if $\mathbf{s}^1$ and $\mathbf{s}^2$ are the sampling designs of the two subsamples 1 and 2, then $\mathbf{s}^1 \perp\!\!\!\perp \mathbf{s}^2$), while by construction it does not explain y_i.

For example, incentives that increase the participation of a subsample of people, randomly selected from a larger sample, allow the construction of an instrument indicator variable for the Heckman model (Wing 2019). These may be financial incentives or greater stimulus efforts, for example.
It is also possible, under certain circumstances, to use interviewer indicators or variables characterising the call rank for a telephone survey (Behaghel, Crépon, Gurgand, and Le Barbanchon 2015).

Differences in protocol between sub-samples may also be due to the use of different modes of data collection for different sub-samples, with some collection modes achieving higher participation rates than others. However, it is known that in general, respondents do not all participate in the same way depending on the collection mode proposed. The effect of the instrument on participation is therefore not uniform. On the other hand, a solution that meets the monotony hypothesis is to implement "nested" protocols. In other words, the alternative protocol, allowing for a better response rate, must include the collection mode(s) of the reference protocol, so that it can effectively be said that if a person to whom the reference protocol was applied had been offered the alternative protocol, he or she would necessarily have responded. For example, a design in which one group was assigned to telephone and the other to internet would not verify this property because there is no evidence that a person participating on the internet would have participated if offered the telephone for responding (and vice versa). However, a protocol in which all respondents are offered a response via the internet but, in addition, a random sub-sample is offered to respond by telephone, immediately verifies the sufficient conditions for the application of a Heckman model.

More recent extensions of the Heckman model have been developed in the literature, in particular in order to get rid of the strongly parametric assumptions of the Heckman model. For a recent general presentation of the framework of these non-parametric models, the reader is invited to refer to Tchetgen Tchetgen and Wirth (2017). Previously, deviations from the Heckman model, by using alternative or empirical distributions, have been the subject of several articles. One example is that of Martins (2001). One can also refer to the references indicated by Boutchenik, Coudin, and Maillard (2019), in a different context.

C. Identifying endogenous selection

In this part, we try to deepen the understanding of the endogenous selection mechanism and its identification. The relation (11) shows that the key to dealing with the endogenous selection problem is to identify the relationship between r_i and y_i. In this paragraph we will detail the assumptions made in Heckman's model about the form of this relationship and outline some ways of taking into account more general forms of relationship.

We assume here that the conditions under which Heckman's model (13) is convergent are met, conditions which we shall hereafter designate by condition [8] or *heckit* model, a term used in the econometric literature (Greene 2003). We will assume in this framework that y_i is continuous, the case where y_i is a binary variable being generalized from the continuous case. We will also assume, unless otherwise stated, that the instrument w_i is binary.
Independently of the parametric hypotheses on which it is based, the particularity of the model is to suppose the existence of a latent variable r_i^* which orders the individual willingness to participate in the survey (i.e. with a given realization of the random residual ϵ_i^0). The form adopted and the existence of the latent variable, apart from its linearity, translate in a fairly logical way the individual behaviour

8. It should be noted that these conditions are more general than those of Heckman's model, since we are not, at this stage, making any hypothesis on the joint distribution of the residuals. In sum, the *heckit* model is limited to the equations (13) and to the existence of a joint distribution of $(\epsilon_i^0, \epsilon_i^1)$ of any kind.

consisting in deciding whether or not to participate in the survey [9]. Moreover, the scientific literature on selection effects most often takes the existence of this latent variable for granted, without questioning it, on the grounds that it reflects, in its essential generality, the individual behaviour at work (Vytlacil 2002, Lee 2009).

In a model with a latent variable, the participation of the individual i follows directly from this variable : if this one is positive or null, the individual participates, otherwise he does not participate. In the *heckit* model, the variable $\bar{r}_i^* = c^0 + \mathbf{z}_i$ can be seen as an individual propensity to participate in the survey, under the average collection effort, conditional on the individual characteristics $\mathbf{z}_i$. The average individual propensity to participate under the average collection effort is here characterised by $c^0 + \mathbf{z}_i$. The addition of an instrument – binary in this case – in the model stands for an increased collection effort for the subset of individuals such that $w_i = 1$. For these individuals, participation is increased according to a valuation of the collection effort, in terms of participation, which is set at the value ψ (for the notations refer to the model 13).

In practice, knowing the link between y_i and r_i means either identifying the functional form that links these two variables, or identifying the one that links the variables y_i and r_i^*. As these two variables are linked (i.e. their residuals are not independent), the observation of the link between y_i and r_i or r_i^* requires an instrument. The instrument makes it possible to "move" the two variables exogenously. And we can thus identify, at least in part, the relationship between these functions.

To understand the conditions for identifying this relationship, let us look more precisely at what happens in the $(y_i, \bar{r}_i^*)$ plane, where $\bar{r}_i^*$ is the individual propensity to participate under the average collection effort and conditional to the individual characteristics $\mathbf{z}_i$. Note that this individual propensity is not observed.

A similar analysis, developed in the appendix B, can be carried out in the plane $(y_i, \bar{\pi}_i)$, where $\bar{\pi}_i = \Pr(\bar{r}_i = 1) = \Pr(\bar{r}_i^* \geqslant 0)$ is the probability of participation under the average collection effort, this probability being actually observed.

Only the analysis in the $(y_i, \bar{r}_i^*)$ plane is developed here. In this plane, all the individuals are ordered by their propensity to participate $\bar{r}_i^* = c^0 + \mathbf{z}_i\beta + \epsilon_i^0$ under the average collection effort and conditionally to their individual characteristics $\mathbf{z}_i$. We assume a binary instrument $\mathbf{w}_i \equiv w_i \in \{0, 1\}$. Using the instrument w_i, we can write according to (15) :

$$\begin{aligned}\mathbb{E}(y_i|\mathbf{z}_i, w_i = 1, r_i = 1) - \mathbb{E}(y_i|\mathbf{z}_i, w_i = 0, r_i = 1) \\ = \varrho\sigma\left(\lambda(c^0 + \mathbf{z}_i\beta + \psi) - \lambda(c^0 + \mathbf{z}_i\beta)\right) \\ \approx \varrho\sigma\lambda'(c^0 + \mathbf{z}_i\beta)\psi\end{aligned}$$

The last line corresponding to the first order of the Taylor development of the function λ [10], calculated at the point $c^0 + \mathbf{z}_i\beta$, λ' being the first derivative of function λ. Then,

$$r_i^*(w_i = 1|\mathbf{z}_i, \epsilon_i^0) - r_i^*(w_i = 0|\mathbf{z}_i, \epsilon_i^0) = \psi$$

It follows that :

$$\frac{\mathbb{E}(y_i|\mathbf{z}_i, w_i = 1, r_i = 1) - \mathbb{E}(y_i|\mathbf{z}_i, w_i = 0, r_i = 1)}{r_i^*(w_i = 1|\mathbf{z}_i, \epsilon_i^0) - r_i^*(w_i = 0|\mathbf{z}_i, \epsilon_i^0)} \approx \varrho\sigma\lambda'(c^0 + \mathbf{z}_i\beta) \tag{23}$$

With the help of the above relations, it is possible to specify the situation as it stands in the $(y_i, \bar{r}_i^*)$ plane. In this context, it is useful to note that the *heckit* model does not make explicit the relation $y(\bar{r}^*)$. However, the relation (23) gives the directing coefficient of the tangent to the curve $y(\bar{r}^*)$, at the mean point between the two collection efforts.

It can be noted that the expression of the slope of the $y(\bar{r}^*)$ curve at this point does not depend on ψ, i.e. the valuation of the additional collection effort in terms of additional participation. This confirms the role of ψ which can be assimilated here to an exogenous parameter (in the sense of mathematical functions depending on a parameter) for the two implicit functions $\mathbb{E}(y_i|\mathbf{z}_i, \psi)$ and $\mathbb{E}(r_i^*|\mathbf{z}_i, \psi)$ of ψ. As ψ is involved in a linear way in these two quantities (in accordance with the postulated link between the two residuals of the outcome and participation equations), it disappears at the first order of magnitude, in the modelling of $y(\bar{r}^*)$ [11].

Figure 1 shows a general case in which the relationship between y and $\bar{r}^*$ is decreasing. This would correspond to a situation in which high incomes would be reluctant to respond to the survey, while low income would respond more willingly, for example. In these circumstances, if an income is higher than average, the probability that the considered person will participate in the survey is lower, all other things being equal. So in the *heckit* model, the correlation is negative.
In the (y, r^*) plane, individuals are ordered according to their propensity to participate under the average effort $\bar{r}_i^*$. All those with $\bar{r}_i^* \geqslant 0$ participate under the average effort. These individuals are shown in blue circles in figure 1. Then, when the instrument is implemented, i.e. the collection effort is increased, the origin of the $\bar{r}_i^*$ shifts to the left by an amount ψ (here positive, reflecting an increase in participation due to the increase in collection effort), so that those whose propensity is such that $\bar{r}_i^* + \psi \geqslant 0$ now participate. To the previous respondents, using this increased collection effort, are added the respondents corresponding to the red circles in Figure 1. The difference in the means of y on these two sets of respondents (blue points on one side, and red and blue on the

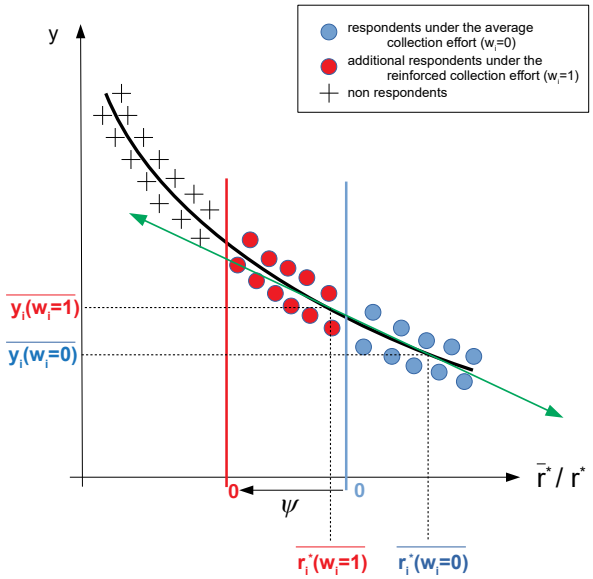
9. In terms of behaviour, the form of this latent variable translates quite naturally the existence of an individual propensity to participate in the survey considered. This propensity will differ from one individual to another and will, very generally, depend on the individual's characteristics and personal context, as well as the characteristics of the survey. It is also likely that residuals, independent of the conditions, do affect participation. This is the case, for example, in a telephone survey, when contact calls are made to the respondent's home when he or she is not there.

10. approximation valid when ψ is small. It is worth mentioning here that the λ function is the inverse of the Mills ratio in the bivariate Gaussian case. More generally, it can be assumed here that it is any function, depending on the joint distribution of the residuals, without the need to specify it further, the only important hypotheses being that $\mathbb{E}(\epsilon_i^1|\epsilon_i^0) = \varrho\sigma\epsilon_i^0$ and symmetrically, $\mathbb{E}(\epsilon_i^0|\epsilon_i^1) = \varrho\ \epsilon_i^1/\sigma$. By the way, this is one of the approaches used by econometricians to generalise Heckman's reasoning to any bivariate distribution of residuals (Greene 2003, for example).

11. This is also the statistical consequence of the fact that this parameter is associated with an instrumental variable. In this context, at the the differential point of view, we have : $\Delta\mathbb{E}(y_i|r_i = 1, \mathbf{z}_i)/\Delta r_i^* = (\Delta\mathbb{E}(y_i|r_i = 1, \mathbf{z}_i, \psi)/\Delta\psi) \times (\Delta r_i^*(\psi)/\Delta\psi)^{-1}$.

other), related to the difference in the means of r_i^* on these same sets, gives the local slope of the $y(\bar{r}^*)$ curve. This is the derivative of this function. It is shown in green in figure 1.

Fig. 1. Curve $y(r^*)$, respondents and non-respondents under two nested collection protocols



Note : $\overline{r_i^*(w_i = 0)}$ corresponds to the average value of r^* for respondents under the average collection effort (i.e. such that $w_i = 0$). $\overline{r_i^*(w_i = 1)}$ is the average value of r^* for respondents under the enhanced collection effort (i.e. such that $w_i = 1$). Formally, in the enhanced protocol, the responding individuals are those who respond to the average protocol (the blue dots), plus those who respond due to the extra collection effort (the red dots). Similarly, $\overline{y_i(w_i = 0)}$ is the average value of y for respondents under the average collection effort and $\overline{y_i(w_i = 1)}$ is the average value of y for respondents under the enhanced collection effort. The individuals represented by crosses are never observed; we only know that they do not participate under either the average collection effort or the enhanced protocol. The black curve is the true $y(r^*)$ function, which is not observable. The green line is the tangent to this curve, observed with the instrument (see text).

As explained above, the instrument and its associated coefficient ψ act as a parameter for the two related functions, y and $\bar{r}^*$. This parameter allows to identify the link. The latter is *de facto* known *locally*. In other words, if the function $y(\bar{r}^*)$ is not linear, the approximation given by the Heckman model is only valid in the vicinity of the participation rate considered. If, for example, the enhanced collection effort (corresponding to the instrument) increases a low original response rate by 10 points, say 30%, then it is unlikely that the linear approximation obtained in the vicinity of a 30-40% participation rate is still valid for non-respondents with high outcome values, as shown in Figure 1.

In practice, if one has a binary instrument, the only knowledge one can get of the link between y and $\bar{r}^*$ is a local one. It is possible to acquire a more extensive knowledge (i.e. on a larger support of y and $\bar{r}^*$ than their real supports) by implementing several protocols allowing to reach differentiated levels of participation, from the lowest to the highest. Such a sequence of protocols, applied on independent samples, can use a selection model identical to the basic protocol, the only difference being the application of a different instrument for each sub-sample associated with a given protocol. For example, if three protocols associated with three sub-samples are implemented $\left(G^{(k)}\right)_{k\in\{0,1,2\}}$, with participation increasing with the sample number, then compared to the model (13), only equation (ii) is modified in:

$$r_i^* = c^0 + z_i\beta + w_i^{(1)}\psi + w_i^{(2)}\kappa + \epsilon_i^0$$

where $w_i^{(1)} = 1$ if $i \in G^{(1)} \cup G^{(2)}$ and $w_i^{(2)} = 1$ if $i \in G^{(2)}$. In this model, ϱ, characteristic of the endogenous selection, is the same for all the protocols and the possible non-linearity of the relation between y and $\bar{r}^*$ goes through the coefficient ψ associated with the instruments of the subsample $G^{(1)}$. In fact, relation (23) is now replaced by two relations (we note $\mathbf{w}_i = (w_i^{(1)}, w_i^{(2)})$) :

$$\text{(a)}\quad \frac{\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w_i} = (1,0), r_i = 1) - \mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w_i} = (0,0), r_i = 1)}{r_i^*(\mathbf{w_i} = (1,0)|\mathbf{z}_i) - r_i^*(\mathbf{w_i} = (1,0)|\mathbf{z}_i)} \approx \varrho\sigma\lambda'(c^0 + \mathbf{z}_i\beta)$$

$$\text{(b)}\quad \frac{\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w_i} = (1,1), r_i = 1) - \mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w_i} = (1,0), r_i = 1)}{r_i^*(\mathbf{w_i} = (1,1)|\mathbf{z}_i) - r_i^*(\mathbf{w_i} = (1,0)|\mathbf{z}_i)} \approx \varrho\sigma\lambda'(c^0 + \mathbf{z}_i\beta + \psi) \tag{24}$$

If we want to be able to describe more easily a non-linearity of the function $y(\bar{r}^*)$, it is possible to associate to the two pairs of protocols $(G^{(0)}, G^{(1)})$ and $(G^{(1)}, G^{(2)})$, two different selection modes. This can be done by estimating two models separately. It can also be done with a single model and two different selection equations (Vella 1998, Ogundimu and Hutton 2016). In this way, the coefficients ϱ appearing in (24-(a − b)) are estimated as two distinct coefficients.

V. Separating measurement error due to mode of data collection and endogenous selection bias

When using the Heckman model with differences in protocols involving several modes of data collection as an instrumental variable, it has so far been assumed that there are no measurement effects on the modelled variable of interest.
If there is a measurement error, i.e. respondents respond differently, depending on the mode of data collection, to the question to which the variable of interest y_i is associated, without this effect being the result of self-selection by the respondents, then, under certain circumstances, it is theoretically possible to identify these mode-specific effects. Indeed, in terms of Heckman's modelling, a measurement effect explains the variable y_i, while it plays no role in the selection equation. Let us explain this point in more detail.

Let us suppose that there are $J+1$ modes of data collection, noted $j \in \{0, 1, \ldots, J\}$ and that mode 0 is the reference mode in relation to which we determine the measurement error associated with the alternative modes $j \in \{1, \ldots, J\}$. The model (13) can then be written:

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

where m_i stands for the mode [12] with whom the individual i answers the survey. If i is non-respondent, then m_i is not observed. Its value is thus conventional and can for example be assumed to be equal to 0, without consequence on the identification of the model. In this model, the α_j can be identified when the variable m_i is not collinear with the $\mathbf{w}_i$ instruments. The identifying assumption is therefore that the mode does not explain (or is not related to) the choice of whether to participate in the survey.

From this point of view, an exclusion condition on m_i must be verified in the selection equation. This is therefore a strong constraint that requires a very specific protocol to be verified. In the general case, there is every reason to believe that the mode also affects participation. The most natural way to ensure independence of mode and participation is to reserve a sub-sample on which participation is ensured, prior to a randomised assignment to a collection mode. This ensures that the mode indicator variable only affects y, not participation. However, in practice, such a protocol is rarely implemented because it is costly to ensure participation before assigning a collection mode and it is not possible to ensure the absence of non-response once the collection mode has been assigned, despite prior agreement.

Under this hypothesis, where m_i is exogenous with respect to the selection equation, the $(\alpha_j)_{j\in\{1,\dots,J\}}$, estimated by maximum likelihood, estimate the measurement error associated with the alternative modes, compared to the reference mode. This method holds for both the continuous and the discrete case.

Conversely, if one introduces the mode m_i into the outcome equation without first ensuring that the exclusion conditions are indeed met, it is extremely likely that m_i is endogenous in this equation. Indeed, the mode probably influences participation. The non-inclusion of m_i in the participation equation rejects the mode in the residual of this equation, which is by hypothesis correlated with that of the outcome equation. Therefore m_i, the explanatory variable of y_i, is likely to be correlated with the residual in the outcome equation [13]. Thus the regression coefficients associated with the measurement effects in the outcome equation are biased. And, unlike the reasoning in paragraph III, where we were interested in the prediction of the model, it is the coefficients associated with the measurement effects that we are interested in here.

Therefore, without an ad hoc protocol to ensure that the exclusion conditions are met, it is not possible to identify a measurement effect in the outcome equation, together with endogenous selection.

We can go further in the reasoning by showing that if a measurement error exists at the same time as an endogenous selection mechanism, nothing is identifiable if the increase in participation results from the addition of an additional collection mode: in this case, neither the measurement error nor the endogenous selection is identifiable. Indeed, if a measurement error exists at the same time as an endogenous selection mechanism, then the non-inclusion of the measurement error in the outcome equation returns it to the ϵ_i^1 residual, and then to ϵ_i^0 *via* the correlation that exists between the residuals due to endogenous selection. If at this stage the r_i^* instruments are linked to the mode of collection [14], then the exclusion conditions of the Heckman model no longer hold, in particular in the participation equation.

In summary:

— to deal with endogenous selection, one can combine modes of data collection for a randomly selected subsample, but this mechanism only corrects for endogenous selection under the assumption that there is no mode-related measurement error;

— and at the same time, one cannot correct for a measurement error that would occur simultaneously with the endogenous selection, if the exclusion conditions necessary to identify measurement errors in the outcome equation do not hold.

Therefore, combining modes of data collection to address endogenous selection carries a real risk of failure if measurement error associated with the mode is suspected. In this case, an instrument based on increased incentives to participate, applied to a random sub-sample, but independent of the mode of data collection (e.g. financial incentives or additional significant follow-up efforts), should be preferred.

12. m_i therefore takes its values in $\{0,\dots,J\}$ according to the collection mode associated with the individual i.

13. Under the assumption that endogenous selection is suspected, for example if a correlation coefficient in a Heckman bivariate probit model is estimated at the same time as the measurement effect.

14. This would be the case when the increased participation results from the addition of an alternative mode to the reference mode of collection.

REFERENCES

ARDILLY, P. (2006): Les techniques de sondage. Technip.

BEHAGHEL, L., B. CRÉPON, M. GURGAND, AND T. LE BARBANCHON (2015): "Please call again: Correcting non-response bias in treatment effect models," Review of Economics and Statistics, 97, 1070–1080.

BOUTCHENIK, B., E. COUDIN, AND S. MAILLARD (2019): "Les méthodes de décomposition appliquées à l'analyse des inégalités," Documents de travail méthodologiques de l'INSEE, (M2019/01).

CAMERON, A. C., AND P. K. TRIVEDI (2005): Microeconometrics: methods and applications. Cambridge university press.

DAWID, A. P. (1979): "Conditional independence in statistical theory," Journal of the Royal Statistical Society: Series B (Methodological), 41(1), 1–15.

GALIMARD, J.-E., S. CHEVRET, E. CURIS, AND M. RESCHE-RIGON (2018): "Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors," BMC Medical Research Methodology, 18(1).

GREENE, W. H. (2003): Econometric analysis. Prentice Hall, 5th edn.

HECKMAN, J. J. (1979): "Sample selection bias as a specification error," Econometrica, 47(1), 153–161.

IMBENS, G., AND J. ANGRIST (1994): "Estimation and identification of local average treatment effects," Econometrica, 62, 467–475.

LEE, D. S. (2009): "Training, wages, and sample selection: Estimating sharp bounds on treatment effects," The Review of Economic Studies, 76(3), 1071–1102.

LESAGE, É., D. HAZIZA, AND X. D'HAULTFŒUILLE (2019): "A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys," Journal of the American Statistical Association, 114(526), 906–915.

LITTLE, R. J., AND D. B. RUBIN (1987): Statistical analysis with missing data. John Wiley & Sons.

MARTINS, M. F. O. (2001): "Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal," Journal of Applied Econometrics, 16(1), 23–39.

OGUNDIMU, E. O., AND J. L. HUTTON (2016): "A unified approach to multilevel sample selection models," Communications in Statistics-Theory and Methods, 45(9), 2592–2611.

SILVERMAN, B. W. (1986): Density estimation for statistics and data analysis, vol. 26. Chapman & Hall.

TCHETGEN TCHETGEN, E. J., AND K. E. WIRTH (2017): "A general instrumental variable framework for regression analysis with outcome missing not at random," Biometrics, 73(4), 1123–1131.

TILLÉ, Y. (2019): Théorie des sondages : Échantillonnage et estimation en populations finies. Dunod, 2nd edn.

TOOMET, O., AND A. HENNINGSEN (2008): "Sample Selection Models in R: Package sampleSelection," Journal of Statistical Software, 27(7).

VELLA, F. (1998): "Estimating models with sample selection bias: a survey," Journal of Human Resources, pp. 127–169.

VYTLACIL, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," Econometrica, 70(1), 331–341.

WING, C. (2019): "What Can Instrumental Variables Tell Us About Nonresponse In Household Surveys and Political Polls?," Political Analysis, 27(3), 320–338.

WOOLDRIDGE, J. M. (2010): Econometric analysis of cross section and panel data. MIT press.

APPENDIX

A. Some variations on conditional laws

In this appendix, some useful results on probabilities and conditional expectations are shown. The aim is to remind the reader of these results and to familiarise him or her with the reasoning used throughout the text. In the first part, we return to the conditioning and conditional orthogonality of variables. A second part derives the results on the bivariate normal model, underlying the Heckman model.

Unless explicitly stated, all variables considered here are vectors.

Any random variable $\mathbf{y}$ is characterised by a probability density $f_{\mathbf{y}}(u)$. This one verifies by definition : $\mathbb{P}(\mathbf{y} < y) = \int_{\mathbf{u}<y} f_{\mathbf{y}}(u)du$. Two variables $(\mathbf{y}, \mathbf{z})$ have a joint distribution $f_{\mathbf{y},\mathbf{z}}(u, v)$. $\mathbf{y}$ is a random variable (we say "$\mathbf{y}$ knowing $\mathbf{z}$"), whose density (called "conditional law of $\mathbf{y}$ knowing $\mathbf{z}$") is:

$$f_{\mathbf{y}|\mathbf{z}}(u|\mathbf{z} = v) = \frac{f_{\mathbf{y},\mathbf{z}}(u, v)}{f_{\mathbf{z}}(v)}$$

The variables $\mathbf{y}$ and $\mathbf{z}$ are independent if and only if their joint density is the product of their marginal densities: $f_{\mathbf{y},\mathbf{z}}(u, v) = f_{\mathbf{y}}(u)f_{\mathbf{z}}(v)$. In this case we note $\mathbf{y} \perp\!\!\!\perp \mathbf{z}$. This relation is symetrical, i.e. $(\mathbf{y} \perp\!\!\!\perp \mathbf{z} \Leftrightarrow \mathbf{z} \perp\!\!\!\perp \mathbf{y})$. Equivalently,

$$\mathbf{y} \perp\!\!\!\perp \mathbf{z} \Leftrightarrow \left(\forall v \ , \ f_{\mathbf{y}|\mathbf{z}}(u|\mathbf{z} = v) = f_{\mathbf{y}}(u)\right)$$

This can also be written more simply :

$$\mathbf{y} \perp\!\!\!\perp \mathbf{z} \Leftrightarrow f_{\mathbf{y}|\mathbf{z}}(u|\mathbf{z}) = f_{\mathbf{y}}(u) \tag{25}$$

We note that if $\mathbf{y} \perp\!\!\!\perp \mathbf{z}$, then the variables $\mathbf{y}$ and $\mathbf{z}$ are not correlated (i.e. have a zero correlation). Indeed, let us suppose without loss of generality that $\mathbb{E}(\mathbf{y}) = E(\mathbf{z}) = 0$ and let us calculate

$$\begin{aligned} \mathbb{E}(\mathbf{yz}) &= \int uv f_{\mathbf{yz}}(u, v)dudv \\ &= \int u f_{\mathbf{y}}(u)du \int v f_{\mathbf{z}}(v)dv \\ &= 0 \end{aligned}$$

We note incidentally that

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \Rightarrow \mathbb{E}(\mathbf{xy}) = \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{y}) \tag{26}$$

The expectation of a variable $\mathbf{y}$ is $\mathbb{E}(\mathbf{y}) = \int u f_{\mathbf{y}}(u)du$. We can define, in a coherent way, the conditional expectation of $\mathbf{y}$ by :

$$\mathbb{E}(\mathbf{y}|\mathbf{z} = z) = \int u f_{\mathbf{y}|\mathbf{z}}(u|\mathbf{z} = z)du$$

Here again, it is customary to adopt the following simplified notation:

$$\mathbb{E}(\mathbf{y}|\mathbf{z}) = \int u f_{\mathbf{y}|\mathbf{z}}(u|\mathbf{z})du$$

From the above, it can be deduced that

$$\mathbf{y} \perp\!\!\!\perp \mathbf{z} \Rightarrow \mathbb{E}(\mathbf{y}|\mathbf{z}) = \mathbb{E}(\mathbf{y})$$

We note that the expression $E(\mathbf{y}|\mathbf{z})$ defines a random variable which is a function of the random variable of $\mathbf{z}$. Therefore,

$$\mathbb{E}(\mathbf{y}) = \int \mathbb{E}(\mathbf{y}|\mathbf{z} = v) f_{\mathbf{z}}(v)dv$$

The formula for iterated conditional expectations, which is always verified, is deduced:

$$\mathbb{E}(\mathbf{y}) = \mathbb{E}\left(\mathbb{E}(\mathbf{y}|\mathbf{z})\right) \tag{27}$$

On the other hand, to be able to be calculated, it is necessary to know *in extenso* the conditional expectation $\mathbb{E}(\mathbf{y}|\mathbf{z})$ and the law of the variable $\mathbf{z}$. We can chain the conditional expectations by iteration:

$$\mathbb{E}(\mathbf{y}|\mathbf{z}) = \mathbb{E}\left[\,\mathbb{E}(\mathbf{y}|\mathbf{x},\mathbf{z})|\,\mathbf{z}\right] \tag{28}$$

Finally, it is possible that the independence of variables is conditional on another variable. For example, the variables $\mathbf{x}|\mathbf{z}$ and $\mathbf{y}|\mathbf{z}$ can be independent whereas $\mathbf{x}$ and $\mathbf{y}$ are not. This property makes sense in relation to the above. By extension of the previous notations:

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} \Leftrightarrow \left[f_{\mathbf{x},\mathbf{y}|\mathbf{z}}(x,y|\mathbf{z}=z) = f_{\mathbf{x}|\mathbf{z}}(x|\mathbf{z}=z) f_{\mathbf{y}|\mathbf{z}}(y|\mathbf{z}=z)\right] \tag{29}$$

Under this assumption, it is possible to separate the expectation calculations. Let any separable function $g(\mathbf{x},\mathbf{y}) = g_1(\mathbf{x})g_2(\mathbf{y})$. We can calculate its expectation conditional on $\mathbf{z}$. By hypothesis, $\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z}$ therefore

$$\begin{aligned}\mathbb{E}(g(\mathbf{x},\mathbf{y})|\mathbf{z}) &= \int g_1(u)g_2(v)f_{\mathbf{x}|\mathbf{z}}(u|\mathbf{z})f_{\mathbf{y}|\mathbf{z}}(v|\mathbf{z})dudv \\ &= \int g_1(u)f_{\mathbf{x}|\mathbf{z}}(u|\mathbf{z})du \int g_2(v)f_{\mathbf{y}|\mathbf{z}}(v|\mathbf{z})dv \\ &= \mathbb{E}(g_1(\mathbf{x})|\mathbf{z})\ \mathbb{E}(g_2(\mathbf{x})|\mathbf{z})\end{aligned}$$

And, in particular, taking $g(x,y) = xy$, we have :

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} \Rightarrow \mathbb{E}(\mathbf{xy}|\mathbf{z}) = \mathbb{E}(\mathbf{x}|\mathbf{z})\ \mathbb{E}(\mathbf{y}|\mathbf{z}) \tag{30}$$

This formula can be seen as a generalization of the relation (26).

We notice, thanks to this last expression, that the application of the formula of the iterated conditional expectations is not immediate. Indeed, by the iterated expectations (relation 27), we have $\mathbb{E}\left[\mathbb{E}(\mathbf{xy}|\mathbf{z})\right] = \mathbb{E}(\mathbf{xy})$. On the other hand, $\mathbb{E}\left[\mathbb{E}(\mathbf{x}|\mathbf{z})\ \mathbb{E}(\mathbf{y}|\mathbf{z})\right] \neq \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{y})$. Indeed, for this last property to be true, the variables $\mathbb{E}(\mathbf{x}|\mathbf{z})$ and $\mathbb{E}(\mathbf{y}|\mathbf{z})$ would have to be independent, and therefore in particular independent of the variable $\mathbf{z}$ with respect to which the two conditional expectations are determined here.

Finally, it may be useful to note that

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} \Rightarrow \mathbb{E}(\mathbf{x}|\mathbf{y},\mathbf{z}) = \mathbb{E}(\mathbf{x}|\mathbf{z}) \tag{31}$$

Indeed,

$$\begin{aligned}f_{\mathbf{x}|\mathbf{y},\mathbf{z}}(u|\mathbf{y},\mathbf{z}) &= f_{\mathbf{x},\mathbf{y}|\mathbf{z}}(u,v|\mathbf{z}) \big/ f_{\mathbf{y}|\mathbf{z}}(v|\mathbf{z}) \\ &= f_{\mathbf{x}|\mathbf{z}}(u|\mathbf{z}) f_{\mathbf{y}|\mathbf{z}}(v|\mathbf{z}) \big/ f_{\mathbf{y}|\mathbf{z}}(v|\mathbf{z}) \\ &= f_{\mathbf{x}|\mathbf{z}}(u|\mathbf{z})\end{aligned}$$

according to (29), hence the announced result.

To fix ideas, let us consider three random variables, x, y and z. Suppose that (a,b,α,β) is a set of constants,

$$\left\{\begin{array}{rcl} x|z &=& a + bz + e \\ y|z &=& \alpha + \beta z + \varepsilon \end{array}\right. \tag{32}$$

with $e \hookrightarrow \mathcal{N}(0,\sigma_e^2)$, $\varepsilon \hookrightarrow \mathcal{N}(0,\sigma_\varepsilon^2)$ and $z \hookrightarrow \mathcal{N}(\eta,\sigma^2)$, where $\mathcal{N}(\eta,\sigma^2)$ denotes the normal distribution with expectation η and variance σ^2. It is further assumed that $e \perp\!\!\!\perp z$ and $\varepsilon \perp\!\!\!\perp z$. These elements allow us to specify the distribution of the random vector $(x,y)'$, conditional to z:

$$(x,y)'|z \hookrightarrow \mathcal{N}\left[\begin{pmatrix} a+bz \\ \alpha+\beta z \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & \rho\sigma_e\sigma_\varepsilon \\ \rho\sigma_e\sigma_\varepsilon & \sigma_\varepsilon^2 \end{pmatrix}\right] \tag{33}$$

where ρ denotes the correlation of the variables e and ε. Classically, given what precedes, the density of the vector variable $(x,y)|z$ is the product of the marginal densities of $x|z$ and $y|z$ if, and only if, $\rho = 0$. Finally, in the present case, the following equivalences hold:

$$\rho = 0 \ \Leftrightarrow\ e \perp\!\!\!\perp \varepsilon \ \Leftrightarrow\ x \perp\!\!\!\perp y|z \tag{34}$$

The following results can be established:

— $\mathbb{E}(x|z) = a + bz$ without restriction. Under the same conditions, $\mathbb{E}(y|z) = \alpha + \beta z$.

— $\mathbb{E}(xy|z) = (a+bz)(\alpha+\beta z) = \mathbb{E}(x|z)\mathbb{E}(y|z)$ is verified if, and only if, $x \perp\!\!\!\perp y|z$. Indeed,

$$\begin{aligned}\mathbb{E}(xy|z) &= \mathbb{E}((a+bz+e)(\alpha+\beta z+\varepsilon)|z) \\ &= (a+bz)(\alpha+\beta z) + \mathbb{E}(e\varepsilon|z) \\ &= (a+bz)(\alpha+\beta z) + \rho\sigma_e\sigma_\varepsilon\end{aligned}$$

according to (32) and (33).The application of equivalences (34) allows us to conclude.

— Using the previous expression, we can also calculate, thanks to the formula of iterated conditional expectations:

$$\begin{aligned}\mathbb{E}(xy) &= \mathbb{E}\left[\mathbb{E}(xy|z)\right] \\ &= (a+b\eta)(\alpha+\beta\eta) + b\beta\sigma^2 + \rho\sigma_e\sigma_\varepsilon\end{aligned}$$

And, $\mathbb{E}(x) = a + b\eta$ and similarly, $\mathbb{E}(y) = \alpha + \beta\eta$. It follows that:

$$\mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y) = b\beta\sigma^2 + \rho\sigma_e\sigma_\varepsilon$$

Thus, the covariance of x and y is zero if, and only if, x or y is independent of z (i.e. $b = 0$ or $\beta = 0$) or z is certain (i.e. $\sigma = 0$) and, simultaneously to one or the other of the preceding conditions, $\rho = 0$.

— Another useful point to note at this stage is that $\mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y)$ is different from $\mathbb{E}\left[\mathbb{E}(xy|z) - \mathbb{E}(x|z)\mathbb{E}(y|z)\right]$. Indeed, if we have $\mathbb{E}\left[\mathbb{E}(xy|z)\right] = \mathbb{E}(xy)$, it is not true for the second component of the expression. This is because,

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}(x|z)\mathbb{E}(y|z)\right] &= \mathbb{E}\left[(a+bz)(\alpha+\beta z)\right] \\ &= (a+b\eta)(\alpha+\beta\eta) + b\beta\sigma^2\end{aligned}$$

which is obviously different from $\mathbb{E}(x)\mathbb{E}(y) = (a + b\eta)(\alpha + \beta\eta)$, except when z is a certain variable or x or y is independent of z. This comes from the fact that the formula of conditional iterated expectations does not hold for a product of variables : the total expectation is (only) a linear application. We find here a property mentioned in the penultimate paragraph of the previous section.

With the previous dependency model, it is possible to continue to explain the conditional laws and expectations. Thus,

we can determine the unconditional law of the vector $(x, y)'$:

$$(x, y)' \hookrightarrow \mathcal{N}\left[\begin{pmatrix} a + b\eta \\ \alpha + \beta\eta \end{pmatrix}, \begin{pmatrix} \sigma_e^2 + b^2\sigma^2 & \rho\sigma_e\sigma_\varepsilon + b\beta\sigma^2 \\ \rho\sigma_e\sigma_\varepsilon + b\beta\sigma^2 & \sigma_\varepsilon^2 + \beta^2\sigma^2 \end{pmatrix}\right]$$

It is thus possible to calculate $\mathbb{E}(x|y)$. We know that for all normal variables u and v, of correlation c, $\mathbb{E}(u|v) = cv$. We deduce that :

$$\mathbb{E}(x|y) = a + b\eta + \frac{\rho\sigma_e\sigma_\varepsilon + b\beta\sigma^2}{\sigma_\varepsilon^2 + \beta^2\sigma^2}(y - \alpha - \beta\eta) \tag{35}$$

We can also calculate $\mathbb{E}(x|y, z)$. Starting from the expression (32), we have :

$$\mathbb{E}(x|y, z) = a + bz + \mathbb{E}(e|y, z) \tag{36}$$

Now $\mathbb{E}(e|y, z) = \mathbb{E}(e|\varepsilon = y - \alpha - \beta z)$. Then, by hypothesis,

$$(e, \varepsilon)' \hookrightarrow \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & \rho\sigma_e\sigma_\varepsilon \\ \rho\sigma_e\sigma_\varepsilon & \sigma_\varepsilon^2 \end{pmatrix}\right)$$

Now, as before, $\mathbb{E}(e|\varepsilon) = \rho\frac{\sigma_e}{\sigma_\varepsilon}\varepsilon$. Then

$$\mathbb{E}(e|y, z) = \rho\frac{\sigma_e}{\sigma_\varepsilon}(y - \alpha - \beta z) \tag{37}$$

And finally,

$$\mathbb{E}(x|y, z) = a - \rho\frac{\sigma_e}{\sigma_\varepsilon}\alpha + (b - \rho\beta\frac{\sigma_e}{\sigma_\varepsilon})z + \rho\frac{\sigma_e}{\sigma_\varepsilon}y \tag{38}$$

The results (35) and (38) are interesting. We can note the following points:

— Relation (38) shows that if $\rho = 0$, then $\mathbb{E}(x|y, z)$ does not depend on y. In other words, in this case, $\mathbb{E}(x|y, z) = \mathbb{E}(x|z) = a + bz$. However, if $\rho = 0$, $\mathbb{E}(x|y)$ still depends on y for the part which "goes through" z. Indeed, from (35) comes :

$$\mathbb{E}(x|y; \rho = 0) = a + b\eta + \frac{b\beta\sigma^2}{\sigma_\varepsilon^2 + \beta^2\sigma^2}(y - (\alpha + \beta\eta))$$

And it would be too quick, starting from (36), to pass from $\mathbb{E}[\mathbb{E}(x|y, z)|y] = a + b\mathbb{E}(z|y) + \mathbb{E}(e|y)$ to $\mathbb{E}(x|y) = a + b\eta$ which would be, of course, false, because:

1) on the one hand, $\mathbb{E}(z|y) \neq \eta$; we can show that $\mathbb{E}(z|y) = \eta + \frac{\beta\sigma^2}{\sigma_\varepsilon^2 + \beta^2\sigma^2}(y - (\alpha + \beta\eta))$ which therefore depends on y;
2) and on the other hand, $\mathbb{E}(e|y) \neq 0$, this by virtue of the relation (37) and of what precedes.

— The relation (35) shows that even if x and y are not generated by the same variable z (i.e. $b = 0$ or $\beta = 0$), the correlation of the residuals e and ε generates a conditional dependence since $\mathbb{E}(x|y)$ effectively depends on y.

— At the same time, even if $\rho = 0$, the expectation of x conditional on y depends on y, through the variable z (point already mentioned above).

— From relation (38), we have again $\mathbb{E}(x|z) = a + bz$. Indeed, starting from (38), we have:

$$\begin{aligned} \mathbb{E}(x|z) &= \mathbb{E}[\,\mathbb{E}(x|y, z)|\,z] \\ &= a - \rho\frac{\sigma_e}{\sigma_\varepsilon}\alpha + \left(b - \rho\beta\frac{\sigma_e}{\sigma_\varepsilon}\right)z + \rho\frac{\sigma_e}{\sigma_\varepsilon}\mathbb{E}(y|z) \end{aligned}$$

Now $\mathbb{E}(y|z) = \alpha + \beta z$, hence the result.

B. Analysis of the conditions for identifying endogenous selection in the $(y_i, \bar{\pi}_i)$ plane

In this plan, the variables are observable. Contrary to the case presented in paragraph IV-C, the individuals are not ordered on the π axis since for the same probability of participation, individuals can or cannot participate in the survey (according to the realization of the ϵ_i^0 residual appearing in the participation equation). In this plane, the relation

$$\begin{aligned} \mathbb{E}(y_i|\mathbf{z}_i, w_i = 1, r_i = 1) - \mathbb{E}(y_i|\mathbf{z}_i, w_i = 0, r_i = 1) \\ \approx \varrho\sigma\lambda'(c^0 + \mathbf{z}_i\beta)\psi \end{aligned}$$

seen previously is still true. Then, according to the relation (18), we have, using a Taylor development to the first order :

$$\begin{aligned} \pi_i(w_i = 1|\mathbf{z}_i, \epsilon_i^1) - \pi_i(w_i = 0|\mathbf{z}_i, \epsilon_i^1) \\ \approx \varphi\left(\frac{c^0 + \mathbf{z}_i\beta + \frac{\varrho}{\sigma}\epsilon_i^1}{\sqrt{1 - \varrho^2}}\right)\frac{\psi}{\sqrt{1 - \varrho^2}} \end{aligned}$$

where φ is the density of the zero-mean and unit-variance normal law. It follows that :

$$\begin{aligned} \frac{\mathbb{E}(y_i|\mathbf{z}_i, w_i = 1, r_i = 1) - \mathbb{E}(y_i|\mathbf{z}_i, w_i = 0, r_i = 1)}{\pi_i(w_i = 1|\mathbf{z}_i, \epsilon_i^1) - \pi_i(w_i = 0|\mathbf{z}_i, \epsilon_i^1)} \approx \\ \varrho\sigma\sqrt{1 - \varrho^2}\left[\varphi\left(\frac{c^0 + \mathbf{z}_i\beta + \frac{\varrho}{\sigma}\epsilon_i^1}{\sqrt{1 - \varrho^2}}\right)\right]^{-1}\lambda'(c^0 + \mathbf{z}_i\beta) \end{aligned}$$

As before, ψ plays the role of a parameter for the two functions $\mathbb{E}(y_i|r_i = 1, \mathbf{z}_i, \psi)$ and $\pi_i(\psi)$. The corresponding situation is presented in figure 2. The previous expression corresponds to the slope of the straight line shown in green in the figure. This straight line, as an approximation of the $y(\bar{\pi})$ curve in the vicinity of $(\overline{\pi(w = 0)}, \overline{y(w = 0)})$ is observable since the average values of y and π are. On the other hand, as before, the previous tangent gives only an approximation of the curve for observed collection rates. In the example of the figure, the approximation obtained for low participation rates may not be valid for higher rates. The discussion in paragraph IV-C on the $(y_i, \bar{r}_i^*)$ plane also applies in the case considered here.

C. Simulations using the R package sampleSelection

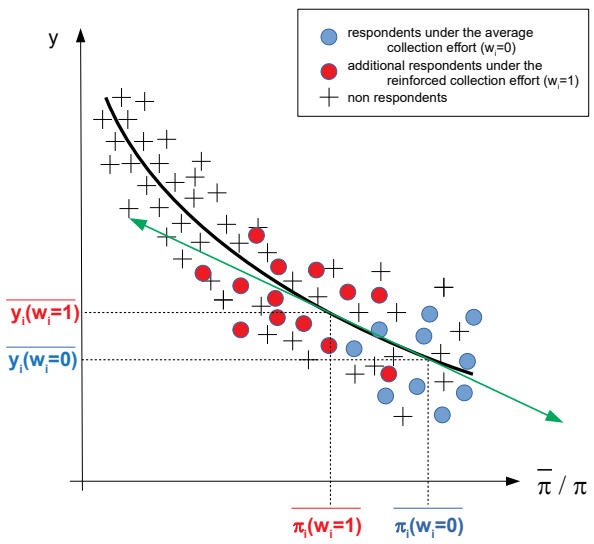
In this appendix, we present simulations based on synthetic observations giving rise to endogenous selection and the results obtained using Heckman models fitted to these observations.

All the results presented here are obtained from the R code NRC-Heck-model.R [15]. This program uses the sampleSelection [16] package (Toomet and Henningsen 2008).

15. Accessible under GitHub at the following link : `https://github.com/InseeFrLab/NRC-heck-model`

16. `https://cran.r-project.org/package=sampleSelection`

Fig. 2. Curve $y(\pi)$, respondents and non-respondents under two nested collection protocols



Note : $\overline{\pi_i(w_i = 0)}$ corresponds to the average value of π for respondents under the average collection effort (i.e. such that $w_i = 0$). $\overline{\pi_i(w_i = 1)}$ corresponds to the average value of π for respondents under the reinforced collection effort (i.e. such that $w_i = 1$). Formally, in the reinforced protocol, the responding individuals are those who respond to the average protocol (the blue dots), plus those who respond due to the extra collection effort (the red dots). Similarly, $\overline{y_i(w_i = 0)}$ is the average value of y for respondents under the average collection effort. $\overline{y_i(w_i = 1)}$ is the average value of y for respondents under the reinforced collection effort. The individuals represented by crosses are never observed; we only know that they do not participate under either the average collection effort or the reinforced protocol. The black curve is the true $y(\pi)$ function, which is not observable. The green line is the tangent to this curve, observed with the instrument (see text).

1) Construction of the synthetic population and self-selection: We build a population of 10 000 individuals whose income depends on three exogenous variables x_1, x_2, x_3, each of these variables being drawn in a uniform distribution $\mathscr{U}$, on $[2, 5]$, $[0, 2]$, and $[0, 1]$ respectively. The income is obtained by the following relation:

$$y = 2 \times x_1 + 1 \times x_2 - 0.5 \times x_3 + \epsilon \tag{39}$$

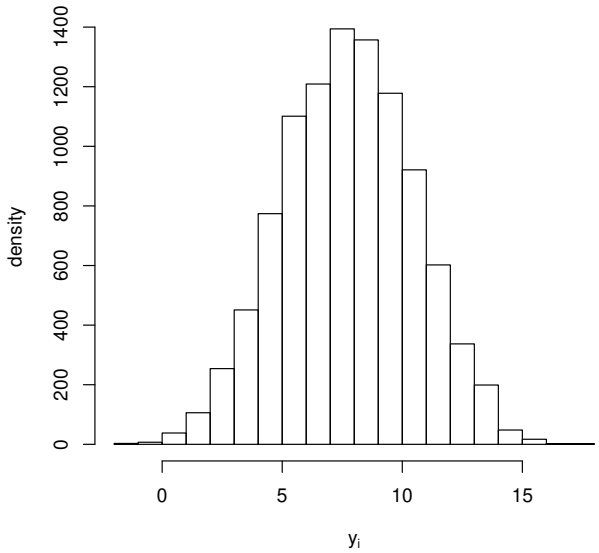
where ϵ_i is drawn from a normal distribution $\mathscr{N}(0, 2^2)$.

We deduce that the true mean income is 7.75 and the true standard deviation of the arithmetic mean of a sample of 10 000 independent and identically distributed (iid) individuals is 0.0277.

An empirical distribution of a vector of 10 000 incomes $(y_i)_{i\in\{1,\dots,10\ 000\}}$ is given in figure 3. The simulated mean associated with the drawing of these 10 000 individuals is 7.74. The standard deviation of the mean is 0.0274.

An endogenous selection mechanism is simulated, based on the previous income, for the 10 000 simulated individuals, according to the following relationship, so that participation

Fig. 3. Histogram of simulated income



decreases with income:

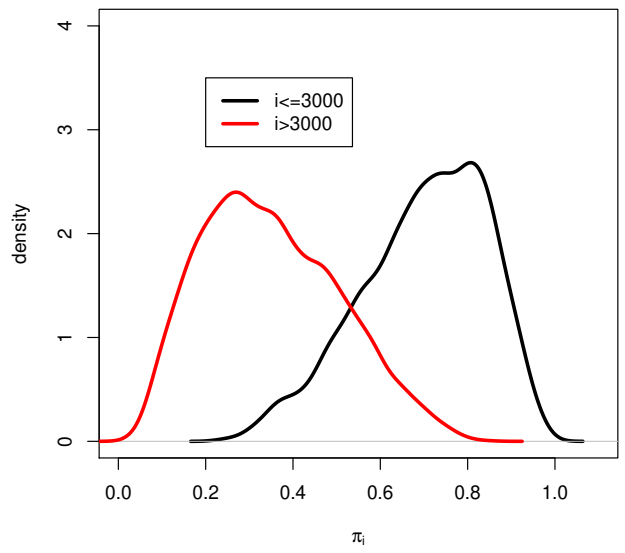
$$\begin{cases} r_i^* &= -0.4 + (\max(y) - y_i)/30 + 0.2 \times \mathbb{1}(i \leqslant 3000) + \nu_i \\ r_i &= \mathbb{1}(r_i^* \geqslant 0) \end{cases} \tag{40}$$

where $\max(y)$ denotes the maximum observed on the $(y_i)_{i\in\{1,\dots,10\ 000\}}$ and ν_i is a random variable drawn in a normal distribution $\mathscr{N}(0, 0.2^2)$. The number of respondents $n_r = \sum_i r_i$; the mean of n_r is 4 550 and, in the simulated draw, it is 4 518. According to (40), the first 3 000 individuals, ranked from 1 to 10 000, have their latent participation variable reinforced, compared to the 7 000 following individuals. This indicator of belonging to the first 3 000 individuals is consistent with the definition of an instrument since it explains the increased participation of these 3 000 individuals, without playing a role in income formation. Figure 4 plots the distributions of simulated participation probabilities, depending on whether individuals are in the group with increased participation (to which the instrument is associated) or not.

Figure 5 shows the probability of response as a function of income for all 10 000 individuals. We note, as expected, the decrease in the function obtained and the separation of the two groups, depending on whether the individual is in the group of 3 000 individuals affected by the increased probability of response to which the instrumental variable will be associated, or whether the individual is among the 7 000 remaining.

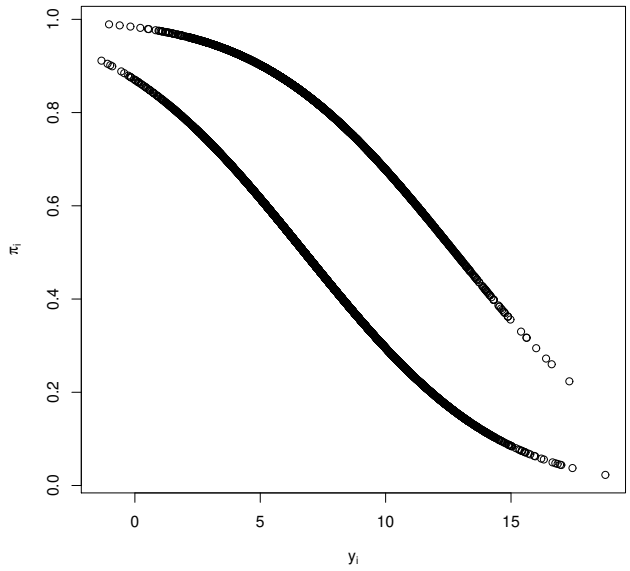
This figure can be compared to the figure 2, subject to two adjustments. Firstly, the abscissa and ordinate of the two figures must be exchanged. Secondly, the curve drawn in figure 2 refers, on the abscissa, to the probability of response $\bar{\pi}$ under the average collection effort, i.e. at zero instrument value. Strictly speaking, the instrument, when it is non-zero, increases the probability of participating. It therefore shifts the inclusion probability curve upwards under the increased collection effort, at a given y, in the system of axes in Figure 5. This is precisely what is observed in the

Fig. 4. Distribution of simulated response probabilities



Note : distribution of the response probabilities, in black for the first 3000 individuals in the sample, i.e. those for whom the probability of participating is reinforced (in connection with the instrumental term in $0.2 \times \mathbb{1}(i \leqslant 3000)$ in the expression (39), and in red, for the 7000 remaining individuals. By applying (40), $\pi_i = \Phi\left[(-0.4 + (\max(y) - y_i)/30 + 0.2 \times \mathbb{1}(i \leqslant 3000))/0.2\right]$. Densities estimated by Gaussian kernel method (Silverman 1986).

Fig. 5. Plot of simulated response probabilities as a function of income



Note : In application of (40), $\pi_i = \Phi\left[(-0.4 + (\max(y) - y_i)/30 + 0.2 \times \mathbb{1}(i \leqslant 3000))/0.2\right]$. The upper curve (top right) corresponds to individuals whose probability of participation is enhanced by the instrument.

latter figure: the uninstrumented probabilities (i.e. under the average collection effort) and the instrumented probabilities (i.e. under the reinforced collection effort) are represented in the same graph. In Figure 2, we have chosen, on the contrary, to keep the probability under the average collection effort, by showing the additional points of respondents, obtained because of the instrumentation, using a different colour (red in this case) from that of respondents in the sample without instrumentation (blue).

2) Estimators and variance: From the previous population simulation, we study the expectation and variance of different estimators for the simulated population in the absence of non-response, on the one hand, and for the respondents alone, on the other, in relation to the simulated participation variable. The estimators for the respondents are either Hajek estimators – biased by construction – or Horvitz-Thompson estimators based on inclusion probabilities from one- or two-stage Heckman models, obtained by reweighting the respondents, in accordance with relation (19), or by imputing the responses of non-respondents, in accordance with relation (17). Different ways of variance estimation are used: by multiple generation of the simulated population, or by bootstrapping on a particular simulation of the population of 10 000 individuals, the bootstrap estimation being naturally the only one by simulation that can be used, in practice, when working on a real sample.

The table I gives the results of the different simulations carried out. Several points should be underlined concerning the results presented in this table.

— As expected, we observe the strong bias of the estimated income on respondents (see column "Average" of the table I : compare row (b) to row (a)). The difference between the true value (7.74) and the estimated value on respondents (6.81) is much wider than the confidence intervals on the estimators (of 0.1 point of semi-amplitude).

— The Hajek estimator using the true inclusion probabilities of the respondents, i.e. those calculated using the true distribution of the latent variable, as in Figure 4, is unbiased (row (c)). This estimator is, however, more uncertain than the one for the whole population (row (a)). It is also more uncertain than the Hajek estimator for respondents (row (b)), due to the increased dispersion of the weights (standard deviation of 0.0515 versus 0.0383).

— The Heckman estimators (rows (d-g)), obtained in one or two steps by reweighting respondents or imputing non-respondents, are unbiased with respect to their confidence intervals. The associated standard deviations are higher than the reference standard deviation for mean of population income: the ratio of standard deviations is about 3, compared to a situation without non-response. Thus – and this is natural – endogenous non-response and its treatment have a cost in terms of loss of precision.

— The standard deviation estimators are themselves subject to imprecision. The comparison of columns (2) and (3) with column (1) of the table I for the estimators in rows (a-b) gives the signature of this uncertainty. Given the order of magnitude of this uncertainty, the bootstrap estimators and those obtained by population simulation are compatible.

— Provided that the uncertainty of the standard deviation estimators obtained by simulation is small, the Heckman estimator by reweighting is more accurate (about 20%) than the one obtained by imputing non-respondents (compare rows (d) and (e) of the table, then (f) and (g)). This is true whether the Heckman estimators are based on a one-step or two-step model.
— Estimators based on one-step Heckman models are more accurate (5 to 10%) than estimators based on two-step models, whether the correction is made by reweighting (comparing lines (d) and (f)) or by imputation (comparing lines (e) and (g)). This result is related to the higher efficiency of the maximum likelihood (i.e. one-step) estimator compared to the two-step estimator.
— Regarding the computation time of the bootstrap loops, the experiment shows that the standard deviation converges rather slowly. The results are stable from a number of bootstrap loops higher than 10 000. Also, the calculation is rather long since the Heckman estimator requires, at each bootstrap loop, to perform a likelihood optimization. This likelihood is more complex in the one-step case, and therefore longer to calculate, than in the two-step case. Although the one-step estimator is more efficient, in this context the two-step estimator can be preferred since the calculation of the one-step Heckman-corrected Hajek estimator takes, in this case, twice as long.

TABLE I

SIMULATED ESTIMATES AND ASSOCIATED VARIANCES

	Estimate	Size	Average	Standard-deviation			$\hat{\varrho}$	Remark
				(1)	(2)	(3)	(4)	
(a)	$\hat{\mu} = \frac{1}{n}\sum y_i$	10000	7.74	0.0273	0.0274	0.0270		
(b)	$\hat{\mu} = \frac{1}{n_r}\sum r_i y_i$	4518	6.81	0.0383	0.0384	0.0390		
(c)	$\hat{\mu} = \sum \alpha_i r_i y_i$	4518	7.68	0.0515				α_i = (real probability of inclusion)$^{-1}$
(d)	$\hat{\mu}_{\text{Heckman-1St.}}$	4518 / 10000	7.75		0.0676	0.0722	-0.363	reweighting (†)
(e)	$\hat{\mu}_{\text{Heckman-1St.}}$	4518 / 10000	7.77		0.0878	0.0848		imputation
(f)	$\hat{\mu}_{\text{Heckman-2St.}}$	4518 / 10000	7.75		0.0733	0.0792	-0.368	reweighting (†)
(g)	$\hat{\mu}_{\text{Heckman-2St.}}$	4518 / 10000	7.78		0.0927	0.0881		imputation

Note : n = 10 000 ; in the equations, the notations refer to the relations (39) and (40). The column "Average " is the estimated value for a particular draw of the synthetic population, the same as that used in figures 3, 4 and 5. (1): application of the analytical variance formula; (2): variance by bootstrap (20 000 loops) in a particular draw of the simulated reference population, the same as the one corresponding to rows (a) to (c) of the table; (3): variance computed through multiple generation (2 000 simulations) of the population (comparable to a true simulated variance) – in this case, the number of respondents is on average 4 550 individuals; (4) correlation coefficient of the residuals in the Heckman model (cf. relation (13), for example), the standard deviation resulting from the maximum likelihood estimation being 0.0460 ; testing the absence of endogenous selection is equivalent to testing the nullity of this coefficient ; (†) : winsorised for the predicted inclusion probabilities lower than 0.1, which are thus reprocessed in order to be saturated at this level.