Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman

Documents de travail
N° M2021-02 - Mars 2021

Laura CASTELL
Patrick SILLARD

Insee
Mesurer pour comprendre

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Série des documents de travail « Méthodologie Statistique »
de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2021/02

Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman

Laura CASTELL
Patrick SILLARD

Insee

Mars 2021

Les auteurs remercient Jérôme Accardo, Pascal Ardilly, Pauline Givord, Sylvie Lagarde, Stéfan Lollivier, Amandine Schreiber pour leur relecture attentive du document et leurs conseils précieux, ainsi que les participants aux séminaires de la DMCSI du 10 septembre et du 19 novembre 2020.

Direction de la méthodologie et de la coordination statistique et internationale
Département des Méthodes Statistiques -Timbre L001 -
88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : -DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman

Laura Castell et Patrick Sillard, 31 mars 2021

Résumé — **Ce document de travail a pour objectif de décrire les conditions dans lesquelles le biais de sélection lié à la non-réponse dans les enquêtes auprès des ménages peut être corrigé. Généralement, les méthodes de correction mises en œuvre font l'hypothèse d'un mécanisme de non-réponse ignorable. Cependant, lorsqu'il existe un problème de non-réponse endogène, alors le mécanisme de non-réponse n'est plus ignorable, et les estimateurs issus des méthodes de correction classiques sont biaisés.**
Pour corriger ce biais, nous proposons une pondération issue d'un modèle de Heckman. Ce modèle consiste à modéliser simultanément la participation et la variable d'intérêt que l'on cherche à estimer. L'identification du modèle est cependant conditionnée à un certain nombre d'hypothèses, comme l'existence d'un instrument, explicatif de la participation mais pas de la variable d'intérêt. Pour disposer d'un tel instrument, un protocole adapté avec des sous-échantillons indépendants peut être mis en place. Ce document détaille les conditions sous lesquelles ce type de protocole permet une estimation corrigée de la sélection endogène.

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

Dans un certain nombre d'enquêtes auprès des ménages, on peut suspecter un phénomène de sélection endogène conduisant à un biais de sélection non-ignorable. De fait, la participation est souvent liée à l'intérêt des enquêtés pour la thématique de l'enquête, notamment lorsqu'il s'agit d'enquêtes auto-administrées. Si cet intérêt pour la thématique influe, conditionnellement aux observables, sur la participation et sur les variables d'intérêt de l'enquête, on se trouve face à un problème de sélection endogène. Or, dans ce cas, les méthodes de correction classiques conduisent à des estimateurs biaisés. Ce biais est d'autant plus important que le taux de réponse est faible et que la variable omise est corrélée aux variables d'intérêt.
Nous proposons ici une méthode de correction de la sélection endogène à partir d'un modèle de Heckman. Ce modèle consiste à modéliser de manière simultanée la participation et la variable d'intérêt. La relation entre la participation et la variable d'intérêt est identifiable à condition de disposer d'un instrument. Les caractéristiques de cet instrument ainsi que les hypothèses du modèle conditionnent cependant l'estimation.
Pour commencer, nous rappelons le cadre d'analyse de l'estimation d'une variable d'intérêt dans le cas d'une enquête auprès des ménages réalisée par sondage (partie I). Nous détaillons ensuite les conditions sous lesquelles la sélection liée à la non-réponse ignorable d'une part (partie II) et non-ignorable d'autre part (partie III) peut conduire à une estimation biaisée. Dans la partie IV, nous présentons la méthode de correction de la sélection endogène proposée à partir du modèle d'Heckman et développons ses conditions d'application et, plus généralement, d'identification. Enfin, la partie V précise les conditions, relativement rares, dans lesquelles il est possible de séparer les erreurs de mesure, par exemple liées à un mode de collecte, et un biais de sélection endogène, dans le cadre du modèle de Heckman.

I. Notations et principes généraux

On note y la variable d'intérêt, collectable conceptuellement sur les individus d'une population $\mathscr{P}$. Chaque individu de la population est identifié par son indice $i \in \{1, \ldots, N\}$. On s'intéresse à la moyenne de cette variable dans la population :

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

Dans une enquête, μ ne peut pas être observé, car seuls certains individus i font effectivement l'objet d'une observation. On note s_i la variable aléatoire indicatrice valant 1 si l'individu i est échantillonné, et 0 sinon. Cette variable est donc une variable binaire. Le plan de sondage est une variable aléatoire vectorielle

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

On note $\mathbf{Z}$ un ensemble de variables $\mathbf{z}_i$ connues *ex-ante* sur toute la population $\mathscr{P}$ (car figurant dans la base de sondage). Sa loi $f_{\mathbf{Z}}$ est donc connue sur $\mathscr{P}$. Cette variable $\mathbf{Z}$ est utilisée par exemple pour stratifier le plan de sondage $\mathbf{s}$ ou pour équilibrer le premier degré dans le cas d'un plan de sondage à plusieurs degrés.
L'enquête permet de collecter la variable d'intérêt y_i (dont la forme vectorielle sur $\mathscr{P}$ est notée $\mathbf{y}$), ainsi que des variables caractéristiques des individus enquêtés $\mathbf{x}_i$ (dont la forme matricielle sur $\mathscr{P}$ est notée $\mathbf{X}$). Ces variables sont collectées uniquement pour les répondants mais existent pour l'ensemble de la population $\mathscr{P}$.

Le plan de sondage est déterminé sur la base des variables $\mathbf{Z}$. Par conséquent, le plan de sondage se caractérise par une loi $f_{\mathbf{s}|\mathbf{Z}}$. La loi de $\mathbf{Z}$ étant connue pour la population $\mathscr{P}$, on peut en déduire $\mathbb{E}(s_i)$, d'après (27) :

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

On note $\mathbb{E}(s_i) = \pi_i$ la probabilité d'inclusion de i dans l'échantillon (ou probabilité que i soit sélectionné).
L'estimateur d'Horvitz-Thompson de μ, fondé sur le plan de sondage $\mathbf{s}$ est classiquement :

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

Celui-ci est observable puisque y_i est observé dès que $s_i = 1$.

On va montrer que $\hat{\mu}$ estime sans biais μ [1]. Pour justifier que l'estimateur d'Horvitz-Thompson est convergent, il convient d'examiner $\mathbb{E}(\hat{\mu}|\mathbf{y})$ (et non pas $\mathbb{E}(\hat{\mu})$ dans l'absolu), et de montrer que cette espérance conditionnelle vaut μ tel que défini en (1).

Calculons donc :

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

Or on a vu que $\mathbf{s}$ ne dépend que de $\mathbf{Z}$, donc

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

C'est l'hypothèse (H1) de ce texte. Il est intéressant de s'arrêter un instant sur cette hypothèse, pour préciser le mécanisme sous-jacent. Little and Rubin (1987) indiquent que la distribution de l'échantillonnage ($\mathbf{s}$) étant déterminée **avant** que toute observation $\mathbf{y}$ ne soit réalisée, alors la distribution $f_{\mathbf{s}|\mathbf{y},\mathbf{Z}}$ ne peut pas dépendre de $\mathbf{y}$ qui est inconnue pour l'échantillonneur dans le cadre *ex-ante* dans lequel il fixe $\mathbf{s}$ et sa réalisation. Par conséquent, $f_{\mathbf{s}|\mathbf{y},\mathbf{Z}} \equiv f_{\mathbf{s}|\mathbf{Z}}$, conformément aux relations (25) et (29), qui s'écrit donc aussi : $\mathbf{s} \perp\!\!\!\perp \mathbf{y}|\mathbf{Z}$. Dawid (1979), qui a introduit cette notation, indique qu'intuitivement, cela revient à dire qu'étant donné $\mathbf{Z}$, une information complémentaire connue sur $\mathbf{y}$ n'altère en rien l'incertitude résiduelle sur $\mathbf{s}$.

Il en découle, d'après (2), la formule des espérances itérées (28) et la conséquence, sur les espérances conditionnelles, des relations d'orthogonalité entre variables (31), que :

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

Finalement,

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

On note que la condition sous laquelle ce résultat est obtenu est la connaissance de la loi de $\mathbf{Z}$ (i.e. sur $\mathscr{P}$) et de celle de $\mathbf{s}|\mathbf{Z}$ (relation (2)).

1. Pour cela, il convient de préciser le cadre de pensée de la théorie des sondages. Dans ce cadre, la variable $\mathbf{y}$ n'est pas une variable aléatoire. C'est un jeu de paramètres observables sur les personnes enquêtées. Néanmoins, cette grandeur peut être reliée, formellement, aux variables aléatoires caractéristiques de l'enquête, notamment du plan de sondage (i.e. la variable $\mathbf{s}$). Pour être précis, elle peut conditionner ces variables. Par conséquent, pour établir les conditions de convergence des estimateurs, il convient théoriquement de traiter $\mathbf{y}$ comme une variable aléatoire, même si c'est une hypothèse qui n'est pas nécessaire au cadre de la théorie des sondages. Une manière de sortir de ce débat de principe est de considérer que $\mathbf{y}$ est effectivement une variable aléatoire (par exemple issue d'une « super-population ») mais que tous les estimateurs fondés sur l'enquête sont conditionnés par $\mathbf{y}$. C'est ce que nous faisons dans ce texte.

II. Non-réponse ignorable : le modèle standard

La non-réponse dans les enquêtes est un processus complémentaire à la sélection dans l'échantillon mais qui fonctionne de manière analogue. En complément à la sélection caractérisée par le plan de sondage $\mathbf{s}$, la non-réponse est une variable binaire r_i qui vaut 1 si l'individu i, sélectionné dans l'enquête, répond à l'enquête, et 0 sinon. La réponse à l'enquête est donc caractérisée par le produit des variables $s_i r_i$. On dispose des variables $(\mathbf{y}, \mathbf{X})$ sur le seul champ $(\mathbf{sry}, \mathbf{srX})$. L'estimateur d'Horvitz-Thompson non corrigé

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

est biaisé puisque $\mathbb{E}(s_i r_i) \neq \pi_i$. On peut néanmoins dériver les hypothèses sous lesquelles il est possible de construire un estimateur d'Horvitz-thompson corrigé sans biais.

Assez naturellement, on cherche, pour cet estimateur corrigé, un estimateur de la forme :

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

où $\hat{\rho}_i$ jouerait le rôle d'un modèle de r_i, sous une forme et des conditions à préciser, de sorte que $\mathbb{E}(\hat{\mu}^1|\mathbf{y}) = \mu$. Nous allons examiner ces conditions.

Comme précédemment, le calcul de $\mathbb{E}(\hat{\mu}^1|\mathbf{y})$ va faire intervenir un conditionnement par la variable $\mathbf{Z}$, c'est-à-dire par les observables disponibles sur l'ensemble de l'échantillon $\mathbf{s}$. En effet, pour identifier un modèle de $\mathbf{r}$, on ne peut pas se contenter de travailler sur le seul observé, car dans ce cas, tous les r_i sont égaux à 1. Il n'y aurait donc aucun moyen d'identifier un tel modèle.

Par construction,

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

Pour les mêmes raisons que celles employées pour justifier l'hypothèse (H1), s_i est déterminé *ex-ante* par la connaissance de $\mathbf{Z}$. Dans ce contexte, la connaissance de r_i n'apporte rien sur celle de s_i (et réciproquement en vertu de la symétrie de la relation $\perp\!\!\!\perp$ – voir Annexe A). Cela se traduit par l'hypothèse suivante :

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

Sous cette hypothèse et en application de la relation (30), il est possible de séparer les contributions dans l'expression précédente :

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

A l'aide de l'hypothèse (H1), il vient :

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

Admettons que l'on soit capable de modéliser r_i par un estimateur $\hat{\rho}_i$, sans biais et convergent.

Pour fixer les idées, $\hat{\rho}_i$, dans ce schéma, est obtenu par régression (modèle à probabilité linéaire par exemple) de r_i

sur y_i et $\mathbf{z}_i$. Cette régression est problématique pour les deux raisons suivantes.

En premier lieu, cette régression fait intervenir *a priori* y_i comme explicative. Ceci est raisonnable car il est possible que la propension à répondre r_i soit expliquée par y_i, ou par une variable qui lui est corrélée. C'est, du moins dans le cas général, une hypothèse que l'on ne peut exclure. Or, l'échantillon observé, qui serait utilisé ici pour identifier le modèle $\hat{\rho}$, serait tel que pour tout i, $r_i = 1$, puisque y_i n'est observé que lorsque $r_i = 1$. Un tel modèle ne serait pas identifiable.

Ce constat réalisé, on peut envisager, en second lieu, de prédire r_i par un modèle tronqué, conditionnel à $\mathbf{Z}$. Mais si y_i intervient réellement comme explicative de r_i, alors y_i apparaît comme une variable omise du modèle, et dont l'absence conduit à un estimateur $\hat{\rho}_i$ baisé. En somme, le modèle $\hat{\rho}$ n'est identifiable et non biaisé que lorsque l'on fait la double hypothèse suivante :

(H3)
- $r_i \perp\!\!\!\perp y_i | \mathbf{Z}$
- $\hat{\rho}_i \equiv \rho(\mathbf{z}_i; \hat{\boldsymbol{\gamma}})$ où
  - ρ est une fonction connue et continue de $\mathbf{z}_i$ et $\hat{\boldsymbol{\gamma}}$
  - $\hat{\boldsymbol{\gamma}}$ est un estimateur convergent de paramètres inconnus $\boldsymbol{\gamma}^*$ (i.e. $\operatorname{plim} \hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^*$) ; $\hat{\boldsymbol{\gamma}}$ dépend de $\mathbf{y}$ et $\mathbf{Z}$
  - $\rho(\mathbf{z}_i; \boldsymbol{\gamma}^*) = \mathbb{E}(r_i | \mathbf{Z})$

Sous les hypothèses précédentes (H1), (H2) et (H3), l'estimateur d'Horvitz-Tompson corrigé $\hat{\mu}^1$ estime μ sans biais asymptotiquement (en $\hat{\boldsymbol{\gamma}}$). En effet,

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

Par hypothèse (H3), $\hat{\rho}_i^{-1}$ ne dépend que de $\mathbf{Z}$ et de $\mathbf{y}$, par l'intermédiaire de $\hat{\boldsymbol{\gamma}}$. Donc,

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

Puis, compte-tenu de (H3), $\operatorname{plim}(\hat{\rho}_i^{-1}) = [\mathbb{E}(r_i | \mathbf{Z})]^{-1}$. Il en découle que, asymptotiquement, $\mathbb{E}(\hat{\mu}^1 | \mathbf{y}) = \mu$.

Le développement précédent traite le cas de la non-réponse complètement ignorable (*Missing completely at random* ou MCAR) et de la non-réponse ignorable (*Missing at random* ou MAR) au sens de Little and Rubin (1987). Le premier correspond au cas où $\hat{\rho}_i \equiv \rho$, avec ρ une constante. La probabilité de répondre est indépendante de $\mathbf{Z}$ et de $\mathbf{y}$. Moyennant quoi, l'estimateur de Hajek (Tillé 2019), défini par :

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

estime μ sans biais asymptotiquement.

Le second correspond au cas où il est possible de construire un estimateur non biaisé de r_i à partir des observables $\mathbf{z}_i$, comme on l'a supposé en (H3). La probabilité de répondre est indépendante de $\mathbf{y}$, mais pas de $\mathbf{Z}$.

Revenons à présent sur le modèle de non-réponse $\hat{\rho}_i$. Comme tous les modèles, il repose sur l'identification de paramètres inconnus $\boldsymbol{\gamma}^*$. Comme l'indique l'hypothèse (H3), on peut écrire :

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

En pratique, (H3), comme (6) qui est analogue, peut toutefois être compliquée à justifier. Nous avons jusqu'à présent supposé qu'elle était vérifiée. Nous allons maintenant étudier quelques déviations par rapport à cette hypothèse.

III. Non-réponse non-ignorable

On propose dans ce paragraphe d'examiner ce qui se passe lorsqu'une variable, bien qu'explicative de la participation à l'enquête r_i, est omise dans l'expression de $\hat{\rho}_i$. Supposons que la variable ξ_i explique r_i mais qu'on omette cette dépendance dans la modélisation :

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

avec $\mathbb{E}(u_i | \mathbf{Z}) = 0$, tandis que le modèle appliqué est :

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

Dans tout ce paragraphe III, on adopte, pour gagner en simplicité et en lisibilité, le cadre de dépendance linéaire correspondant aux deux relations (7-8) précédentes.

Notons tout d'abord qu'il est possible que $\tilde{\rho}_i$ respecte l'hypothèse (H3), malgré l'omission de ξ_i. En effet, des deux relations précédentes, on tire :

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

Or, par hypothèse, $\mathbb{E}(u_i | \mathbf{z}_i) = 0$. Par conséquent, dès que $\mathbb{E}(\xi_i | \mathbf{z}_i) = 0$, il est possible d'obtenir, par exemple par régression linéaire de r_i sur $\mathbf{z}_i$, un estimateur $\tilde{\rho}_i$ vérifiant la deuxième condition de (H3), tout en omettant ξ_i dans le modèle. La première condition de l'hypothèse (H3) peut également être vérifiée en projetant cette condition sur ξ_i, à l'aide de la relation (7). Ainsi, toute variable ξ_i vérifiant :

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

ne pose pas de problème d'omission, c'est-à-dire qu'elle peut être omise dans le modèle de non-réponse sans que cela ne se traduise par un quelconque biais (asymptotique) de l'estimateur de Hajek.

Regardons à présent le cas d'une variable omise ne respectant pas l'une ou l'autre des deux conditions précédentes. Posons par exemple :

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

On peut imaginer deux types de problème :

- un problème d'endogénéité de ξ_i dans la modélisation de $\tilde{\rho}_i$ qui biaise l'estimation des coefficients du modèle (8). Ceci correspond au cas où $\theta \neq 0$ et $\vartheta = 0$ dans l'expression de ξ_i ci-dessus.
- un problème d'auto-sélection endogène dans lequel y_i est en même temps variable d'intérêt et explicative de la non-réponse. Ceci correspond au cas où $\theta = 0$ et $\vartheta \neq 0$.

Naturellement, les deux types de problème sont susceptibles de se superposer, mais leurs conséquences sont très différentes. Et pour la présentation, il est plus simple de les dissocier.

Considérons d'abord le cas où ξ_i est endogène dans le modèle de non-réponse (i.e. $\theta \neq 0$ et $\vartheta = 0$). C'est le cas si on estime $\tilde{\beta}$, dans le cadre d'un modèle à probabilité linéaire, par régression linéaire de r_i sur $\mathbf{z}_i$. Classiquement,

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

A l'aide de (7) et (9), on note que r_i s'écrit aussi :

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

En substituant cette dernière expression à r_i dans (10), et en notant que par hypothèses, $\mathbb{E}(u_i|\mathbf{z}_i) = 0$ et $\mathbb{E}(\upsilon_i|\mathbf{z}_i) = 0$, on conclut que :

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

Et donc :

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

Au final, l'endogénéité de ξ_i dans la modélisation de r_i n'est pas critique puisque, si elle biaise les coefficients de régression, elle ne biaise cependant pas les prédicteurs $\tilde{\rho}_i$ qui découlent de cette estimation, en tant qu'estimateurs des r_i. Dans ce cas, le mécanisme de non-réponse reste ignorable, malgré l'omission de ξ_i.

Considérons maintenant le cas d'une sélection endogène, où ξ_i dépend de y_i (i.e. $\theta = 0$ et $\vartheta \neq 0$). $\tilde{\rho}_i$, obtenu par régression de r_i sur $\mathbf{z}_i$ estime sans biais r_i, conditionnellement à $\mathbf{Z}$. Mais on ne peut pas identifier la dépendance de r_i à y_i par régression de r_i sur $(\mathbf{z}_i, y_i)$ puisque y_i n'est observée que pour les répondants, c'est-à-dire les i pour lesquels $r_i = 1$. Donc la régression identifiante pour $\tilde{\rho}_i$ se fonde sur les seuls $\mathbf{z}_i$. Si la seconde partie de l'hypothèse (H3) est valide, la première partie, en revanche, n'est plus vérifiée puisque r_i dépend de y_i par l'intermédiaire de ξ_i. Dans ce cas, le mécanisme de non-réponse est alors non-ignorable.

Dans le développement de $\mathbb{E}(\hat{\mu}^1|\mathbf{y})$, la première partie reste vraie sous les hypothèses (H1) et (H2) qui sont, elles, vérifiées. Par conséquent, sous ces hypothèses,

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

À l'aide de (7 – 9), on a alors :

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

Pour les mêmes raisons que celles qui conduisent à la relation (10), dans un modèle à probabilité linéaire,

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

On peut ici faire l'hypothèse complémentaire raisonnable que $u_i \perp\!\!\!\perp y_i|\mathbf{z}_i$, ce qui revient à considérer que toute la dépendance de r_i à y_i transite par ξ_i. Tout ceci est essentiellement formel puisqu'il s'agit là d'hypothèses sur la structure de r_i et la séparation additive de ses différentes composantes. Il découle de cette hypothèse complémentaire et de ce qui précède que, asymptotiquement :

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

Finalement (asymptotiquement),

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

Le biais lié à l'existence d'une sélection endogène est donc du signe de la dépendance de la participation à la variable d'intérêt (i.e. de ϑ) : si la participation croît avec la variable d'intérêt, $\hat{\mu}^1|\mathbf{y}$ est biaisé positivement ; ce dernier est biaisé négativement si la participation décroît avec la variable d'intérêt. Et toutes choses égales par ailleurs, le biais croît avec la variance de la variable d'intérêt sur $\mathscr{P}$.

IV. Correction de la non-réponse endogène

La correction de la non-réponse endogène est un problème connu de l'exploitation de données manquantes (Little and Rubin 1987). Il a fait l'objet de nombreux développements économétriques destinés à caler un modèle convenable de la variable d'intérêt (voir par exemple Boutchenik, Coudin, and Maillard (2019)). Ce faisant, ces développements sont intéressants aussi pour l'analyse d'enquête puisqu'ils permettent en particulier d'estimer $\mathbb{E}(y)$ sans biais (voir par exemple Ardilly (2006)). Deux grandes classes de méthodes se distinguent dans ce contexte. Les méthodes à variable latente de participation, comme dans les modèles d'Heckman (Heckman 1979). Plus récemment, ces méthodes ont fait l'objet de déclinaisons et d'approfondissements (Vella 1998, Galimard, Chevret, Curis, and Resche-Rigon 2018, Wing 2019), y compris vers des modélisations très peu paramétriques, par exemple dans le domaine de l'évaluation de traitement avec sélection endogène (voir par exemple l'article de Lee (2009)). Les méthodes de calage généralisé se sont

également développées, fondées sur des conditions d'identification différentes (voir par exemple l'article de Lesage, Haziza, and D'Haultfœuille (2019)).
La suite de ce texte est consacrée à la présentation de l'utilisation du modèle classique d'Heckman pour le traitement de la non-réponse endogène dans les enquêtes.

A. Le cadre du modèle de Heckman

Le modèle de Heckman a été popularisé par l'économètre du même nom (Heckman 1979). Il est aussi connu sous le nom de *Tobit II* (Wooldridge 2010, Cameron and Trivedi 2005). Il consiste à modéliser simultanément y_i et r_i sous la forme suivante :

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

r_i^* est une variable latente qu'on n'observe pas. On observe r_i et y_i lorsque $r_i = 1$. $(\mathbf{z}_i, \mathbf{w}_i)$ est observé pour tout i.
β, χ et ψ sont des paramètres inconnus. $(\epsilon_i^0, \epsilon_i^1)$ sont des aléas. Dans ce modèle, l'équation de participation (13-(iii)) repose sur une variable latente r_i^* (relation (13-(ii)) qui fait intervenir les explicatives de y_i (ici sur l'ensemble de l'échantillon, répondants et non répondants). Cette variable latente fait aussi intervenir des instruments $\mathbf{w}_i$, c'est-à-dire des variables qui expliquent la participation mais ne sont pas explicatives des y_i. On parle de conditions d'exclusion à leur endroit. Formellement, les conditions d'identification du modèle précédent sont :

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

Σ est une matrice de variance. C'est donc à travers cette matrice qu'est modélisée la formation simultanée de la variable expliquée y_i et de la participation r_i. En pratique et dans le modèle (13), tous les paramètres de Σ ne peuvent cependant pas être identifiés. Comme dans un modèle probit, il convient d'adopter une variance unitaire pour ϵ_i^0 puisque les coefficients de (13-(ii)) sont identifiables à un facteur multiplicatif près. Moyennant quoi, la forme proposée pour Σ ci-dessus est la plus générale possible dans le contexte du modèle de Heckman [2].

Notons que ce modèle porte sur l'ensemble de la population $\mathscr{P}$. Outre la sélection liée à l'échantillonnage ($s_i = 1$) dont on a montré qu'elle était sans conséquence sur l'espérance des estimateurs, on n'observe que les répondants, c'est-à-dire les i tels que $r_i = 1$. Il peut être intéressant d'étudier, dans ce modèle, comment se comporte $(\mathbf{ry})$. Calculons donc $\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 1)$. Sous les hypothèses précédentes, on a (voir Annexe A - deuxième partie) :

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

Il en découle que :

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

ϵ_i^0 est une variable gaussienne, donc l'expression de $\mathbb{E}(\epsilon_i^0|\epsilon_i^0 \geqslant -a)$ est une fonction $\lambda(a)$ connue, correspondant à l'inverse du ratio de Mills [3]. Finalement,

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

On montre de la même manière que [4] :

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

On note que :

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

où $\breve{r}_i^* = c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi$. Ainsi défini, $\breve{r}_i^*$ est assimilable pour le raisonnement à un prédicteur de r_i^*.

On observe d'abord, à partir de l'expression précédente, que si $\varrho = 0$, alors $\mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 1) = \mathbb{E}(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 0)$. Ainsi, l'endogénéité de la sélection survient [5], dans le modèle de Heckman, lorsque la corrélation des aléas $(\epsilon_i^0, \epsilon_i^1)$ est non nulle. Et réciproquement, il n'y a pas d'endogénéité de la sélection lorsque $\varrho = 0$.
Dans l'expression (16), le terme entre crochets est positif puisque la fonction λ l'est. Il en découle que si $\varrho > 0$, alors les y_i observés sont, toutes choses égales par ailleurs, plus grands, en moyenne, que les y_i non observés. A l'inverse, si $\varrho < 0$, alors les y_i observés sont, toutes choses égales par ailleurs, plus petits, en moyenne, que les y_i non observés.

La résolution du modèle de Heckman nous permet de construire deux estimateurs : l'un par imputation des y_i pour les non-répondants, l'autre par repondération.

Tout d'abord, la connaissance de $\mathbb{E}(y_i|\mathbf{z}_i, r_i = 0)$ nous permet de construire un nouvel estimateur de μ s'affranchissant de l'étape de sélection endogène r_i, par imputation des y_i des non-répondants, de la manière suivante :

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

Les paramètres $(\hat{c}^1, \hat{\beta}, \hat{\psi}, \hat{c}^0, \hat{\chi}, \hat{\varrho}, \hat{\sigma})$ sont estimés sur le fondement du modèle (13). Les méthodes pour les estimer sont évoquées au paragraphe suivant.

On peut également dériver un autre estimateur de μ en le fondant sur la probabilité d'inclusion conditionnelle issue du modèle de Heckman. Repartons pour cela de l'expression (11) :

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

2. et pour des aléas indépendants entre individus, c'est-à-dire : $\forall i \neq j$, $(\epsilon_i^0, \epsilon_i^1) \perp\!\!\!\perp (\epsilon_j^0, \epsilon_j^1)$.

3. dans la mesure où $\text{var}(\epsilon_i^0) = 1$. Rappelons que l'inverse du ratio de Mills est défini par : $\lambda(a) = \varphi(a)/\Phi(a)$ où φ est la densité de la loi normale centrée-réduite et Φ sa répartition. C'est une fonction strictement décroissante, positive, et qui a pour asymptotes $y = -x$ en $-\infty$ et $y = 0$ en $+\infty$.

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

5. On suppose que $\sigma > 0$.

On peut ici faire intervenir comme variable de conditionnement supplémentaire $\mathbf{w}_i$. Moyennant quoi, s'il est possible de construire un estimateur convergent $\tilde{\rho}_i$ de $\mathbb{E}[r_i|y_i, \mathbf{z}_i, \mathbf{w}_i]$, alors, comme précédemment (cf. §II), on pourra construire un estimateur asymtotiquement sans biais de $\mathbb{E}(\hat{\mu}^1|y)$.
Afin de construire $\tilde{\rho}_i$, calculons $\mathbb{E}(r_i|y_i, \mathbf{z}_i, \mathbf{w}_i)$ à partir du modèle (13). $(r_i|y_i, \mathbf{z}_i, \mathbf{w}_i)$ est une variable binaire, donc $\mathbb{E}(r_i|y_i, \mathbf{z}_i, \mathbf{w}_i) = \Pr(r_i^* \geqslant 0|y_i, \mathbf{z}_i, \mathbf{w}_i)$. Sous les hypothèses précédentes, la loi de $(r_i^*|y_i, \mathbf{z}_i, \mathbf{w}_i)$ est connue. En effet [6],

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

Puis $(\epsilon_i^0, \epsilon_i^1) \hookrightarrow \mathscr{N}\left(0, \Sigma = \begin{pmatrix} 1 & \varrho\sigma \\ \varrho\sigma & \sigma^2 \end{pmatrix}\right)$, donc la loi conditionnelle de $(\epsilon_i^0|\epsilon_i^1)$ est une loi normale. Il vient [7] :

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

Il en découle que :

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

On en déduit l'expression de $\tilde{\rho}_i$:

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

où, comme pour l'estimateur (17), les paramètres $(\hat{c}^1, \hat{\beta}, \hat{\psi}, \hat{c}^0, \hat{\chi}, \hat{\varrho}, \hat{\sigma})$ sont estimés sur le fondement du modèle (13). Les méthodes pour les estimer sont évoquées au paragraphe suivant.

(19) donne l'expression de $\tilde{\rho}_i$. On observe qu'il ne dépend que des variables y_i, $\mathbf{z}_i$ et $\mathbf{w}_i$ et de paramètres dont on peut construire des estimateurs convergents. Sous les hypothèses du modèle d'Heckman (13 – 14), ainsi défini, $\tilde{\rho}_i$ estime asymptotiquement sans biais $\mathbb{E}(r_i|y_i, \mathbf{z}_i, \mathbf{w_i})$. Il en découle, et pour les mêmes raisons que celles avancées au paragraphe II, que

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

avec $\tilde{\rho}_i$ défini à la relation (19), est un estimateur asymptotiquement sans biais de μ.

Il est également possible de traiter le cas de variables binaires avec un modèle de Heckman en utilisant une variable latente pour la variable d'intérêt. Le modèle modifié est le suivant :

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

Les conditions d'identification (i.e. d'exclusion) restent identiques au cas continu (14), à ceci près que l'on peut poser désormais $\sigma = 1$, car les coefficients de (21-(i)) sont désormais identifiés à un facteur multiplicatif près. On observe que :

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

Puis,

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

On en déduit finalement que [8] :

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

où Φ_2 désigne la répartition de la loi normale bivariée (voir note de bas de page n°8). Cette relation est celle préconisée par Galimard, Chevret, Curis, and Resche-Rigon (2018) pour imputer les valeurs prédites pour les non-répondants, conformément à la démarche proposée pour l'estimateur (17). Concrètement, les auteurs précédents proposent d'imputer la réponse des non-répondants en tirant cette réponse dans une loi de Bernouilli de paramètre $\mathbb{P}(y_i = 1|\mathbf{z}_i, \mathbf{w}_i, r_i = 0)$, tel que donné à l'expression (22).

Comme dans le cas continu, il est possible de construire un estimateur de Hajek corrigé analogue à (20) en utilisant les probabilités de réponses, conditionnelles à y_i. Ces probabilités conditionnelles valent :

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

et

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

Ces probabilités permettent de calculer les poids à affecter à chaque observation dans un estimateur de Hajek, selon que la variable d'intérêt y_i vaut 0 ou 1.

6. $\mathscr{L}$ désigne la loi de la variable aléatoire argument.

7. Si $\begin{pmatrix} X \\ Y \end{pmatrix} \hookrightarrow \mathscr{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_X\sigma_Y\varrho \\ \sigma_X\sigma_Y\varrho & \sigma_Y^2 \end{pmatrix}\right)$ est un vecteur bivarié de variables normales , alors $\mathscr{L}(X|Y = y) = \mathscr{N}\left(\mu_X + \varrho\sigma_X(y - \mu_Y)/\sigma_Y; \sigma_X^2(1 - \varrho^2)\right)$.

8. Si $\begin{pmatrix} X \\ Y \end{pmatrix} \hookrightarrow \mathscr{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}\right)$ est un vecteur bivarié de variables normales , alors $\mathbb{P}(X \geqslant 0, Y \geqslant 0) = \Phi_2(\mu_X, \mu_Y; \varrho)$, $\mathbb{P}(X \geqslant 0, Y \leqslant 0) = \Phi_2(\mu_X, -\mu_Y; -\varrho)$, $\mathbb{P}(X \leqslant 0, Y \geqslant 0) = \Phi_2(-\mu_X, \mu_Y; -\varrho)$ et $\mathbb{P}(X \leqslant 0, Y \leqslant 0) = \Phi_2(-\mu_X, -\mu_Y; \varrho)$, où $\Phi_2(x, y; \varrho)$ désigne la fonction de répartition de la loi normale bivariée centrée-réduite de corrélation ϱ.

B. Estimation et discussion du modèle de Heckman

Le modèle de Heckman continu (13) peut s'estimer de deux manières différentes (Cameron and Trivedi 2005) :

- en deux étapes en déterminant d'abord le modèle Probit associé à la relation (iii) ce qui conduit aux estimateurs sans biais des coefficients $(\hat{c}^0, \hat{\beta}, \hat{\psi})$; puis en régressant y_i sur les variables $\left[1, \mathbf{z}_i, \lambda(\hat{c}^0 + \mathbf{z}_i\hat{\beta} + \mathbf{w}_i\hat{\psi})\right]$ ce qui conduit aux estimateurs sans biais des coefficients $(\hat{c}^1, \hat{\chi}, \hat{\varrho})$.
- par maximum de vraisemblance, la vraisemblance associée à (13) étant exprimable analytiquement (Cameron and Trivedi 2005).

Cette deuxième solution est plus efficace au plan statistique. En revanche, la première est plus rapide en temps de calcul (voir annexe C). Et il n'est pas nécessaire de supposer que ϵ_i^1 est normale pour que l'estimation soit sans biais. La méthode en deux étapes est donc moins paramétrique que la méthode par maximum de vraisemblance en une étape. Dans le cas d'une variable d'intérêt binaire (21), seule l'estimation par maximum de vraisemblance est praticable (voir par exemple Galimard, Chevret, Curis, and Resche-Rigon (2018)). L'expression analytique, dans le cas binaire, est donnée par Cameron and Trivedi (2005).

Le package R sampleSelection permet d'estimer de manière très commode un modèle de Heckman et les différents estimateurs développés dans la partie précédente. Des résultats de simulations sont présentés en annexe C. Il montrent, pour un exemple de variable simulée de revenu, avec sélection endogène pour les hauts revenus, que l'estimateur de Hajek peut-être fortement biaisé et qu'un estimateur de Heckman permet, en l'espèce, de corriger ce biais.

Le modèle (13) est théoriquement identifiable sans instruments $\mathbf{w}_i$ dans la relation (ii). Il en est de même pour le modèle (21). Cependant, la présence des instruments $\mathbf{w}_i$ est une condition *sine qua non* de la bonne convergence des méthodes d'estimation. En l'absence d'instruments expliquant la participation sans affecter la variable d'intérêt, dans le cas continu, il y a quasi-colinéarité dans la relation (i) entre l'inverse du ratio de Mills $\lambda(\hat{c}^0 + \mathbf{z}_i\hat{\beta})$ et $\mathbf{z}_i$, de sorte que l'estimation du modèle de Heckman a toutes les chances de ne pas converger, notamment dans le cas d'une estimation par maximum de vraisemblance. Des phénomènes analogues surviennent dans le cas discret. Un protocole d'enquête adapté peut permettre de disposer d'un tel instrument.

Il est utile de noter que l'instrument $\mathbf{w}$ de l'équation de sélection (13-ii) ou (21-ii) implique que le principe de monotonie de l'instrument (Imbens and Angrist 1994, Vytlacil 2002) soit vérifié par le protocole d'enquête retenu. Voyons ce point de plus près lorsque l'instrument est une variable binaire.

Au vu de ces deux équations de sélection, $r_i^*(\mathbf{w}_i = 1) - r_i^*(\mathbf{w}_i = 0) = \psi$, pour tout i. Or ψ est une constante, soit positive, soit négative. Supposons par exemple que $\psi \geqslant 0$. Alors, tout individu participant à l'enquête en l'absence d'instrument (c'est-à-dire tel que $\mathbf{w}_i = 0$ et $r_i^* \geqslant 0$) aurait nécessairement participé à l'enquête en présence de l'instrument (c'est-à-dire si, au lieu de $\mathbf{w}_i = 0$, la valeur de l'instrument le concernant avait été $\mathbf{w}_i = 1$). ψ étant soit positive, soit négative, la participation à l'enquête est, pour tout i, soit croissante avec l'instrument, soit décroissante: elle est donc monotone. Le point important est que cette propriété s'entend *toutes choses égales par ailleurs* : ce n'est pas seulement en moyenne que les individus doivent participer davantage selon qu'ils sont dans un des groupes définis par l'instrument, mais pour tout individu.

La nécessaire monotonie de l'instrument emporte donc des conséquences sur les conditions de protocole permettant de disposer d'un instrument sous lesquelles un modèle de Heckman peut s'appliquer. En pratique, il convient donc que le protocole retenu permette de justifier que les individus du groupe dont le taux de réponse est le plus faible et qui participent effectivement à l'enquête auraient aussi participé, s'ils avaient bénéficié du protocole alternatif, tel que caractérisé par l'instrument.

De tels instruments sont mobilisables dans le cadre d'enquêtes aléatoires lorsque, par exemple, deux sous-échantillons, administrés selon deux protocoles de collecte différents, ont été sélectionnés, l'un des protocoles donnant lieu à une participation plus élevée que l'autre. Dans ces conditions, la réunion des deux échantillons forme un seul échantillon aléatoire et l'indicatrice d'appartenance à un sous-échantillon constitue un instrument. En effet, dans la mesure où la participation diffère entre les deux protocoles, la variable indicatrice d'appartenance à l'un des deux sous-échantillons est explicative de la participation (du fait de la sélection aléatoire d'un échantillon par rapport à l'autre - mathématiquement si $\mathbf{s}^1$ et $\mathbf{s}^2$ sont les plans de sondage des deux sous-échantillons 1 et 2, alors $\mathbf{s}^1 \perp\!\!\!\perp \mathbf{s}^2$), tandis que par construction, elle n'explique pas y_i.

Par exemple, des incitations renforçant la participation d'un sous-échantillon de personnes, sélectionnées aléatoirement dans un échantillon plus vaste de personnes enquêtées, permettent de construire une variable indicatrice instrument pour le modèle de Heckman (Wing 2019). Il peut s'agir d'incitations financières ou d'efforts de relance plus importants par exemple.

Il est également possible, dans certaines circonstances, d'utiliser des indicatrices d'enquêteurs (Bärnighausen, Bor, Wandira-Kazibwe, and Canning 2011) ou des variables caractérisant le rang d'appel pour une enquête téléphonique (Behaghel, Crépon, Gurgand, and Le Barbanchon 2015).

Les différences de protocole entre les sous-échantillons peuvent également tenir à l'utilisation de modes de collecte différents en fonction des sous-échantillons, certains modes de collecte permettant d'obtenir des taux de participation plus importants que d'autres. Cependant, on sait qu'en général les enquêtés ne participent pas tous de la même manière selon le mode de collecte proposé. L'effet de l'instrument sur la participation n'est alors pas uniforme. En revanche, une solution qui répond à l'hypothèse de monotonie est de mettre en œuvre des protocoles « emboîtés ». En d'autres termes, le protocole alternatif, permettant un meilleur taux de réponse, doit inclure le (ou les) mode(s) de collecte du protocole de référence, de sorte que l'on puisse effectivement affirmer que si une personne à qui le protocole de référence a été appliqué s'était vue proposer le protocole alternatif, elle aurait nécessairement répondu. Par exemple, une situation

dans laquelle un des groupes se verrait attribuer un mode de collecte par téléphone, et l'autre par internet, ne vérifierait pas *a priori* cette propriété car rien ne dit qu'une personne qui participe sur internet aurait participé si on lui avait proposé de répondre par téléphone (et réciproquement). En revanche, un protocole dans lequel tous les enquêtés se voient proposer une réponse par internet mais, qu'en outre, un sous-échantillon aléatoire se voit proposer de répondre aussi par téléphone, vérifie d'emblée les conditions suffisantes à l'application d'un modèle de Heckman.

Des extensions du modèle de Heckman, plus récentes, sont développées dans la littérature, notamment de façon à s'affranchir des hypothèses fortement paramétriques du modèle de Heckman. Pour une présentation générale récente du cadre de ces modèles non paramétriques, le lecteur est invité à se reporter à Tchetgen Tchetgen and Wirth (2017). Auparavant, des écarts au modèle de Heckman, par usage de distributions alternatives ou de distributions empiriques ont fait l'objet de plusieurs articles. Un exemple est celui de Martins (2001). On peut aussi se reporter aux références indiquées par Boutchenik, Coudin, and Maillard (2019), dans un contexte différent.

C. L'identification de la sélection endogène

Dans cette partie, on essaie d'approfondir la compréhension du mécanisme de sélection endogène et de son identification. La relation (11) montre que la clé du traitement du problème de sélection endogène est d'identifier la relation qui existe entre r_i et y_i. Nous allons détailler dans ce paragraphe les hypothèses faites dans le modèle d'Heckman sur la forme de cette relation et dresser quelques pistes pour prendre en compte des formes de relation un peu plus générales.

On se place donc ici dans les conditions d'application du modèle d'Heckman (13), conditions que nous désignerons ci-après par condition [9] ou modèle *heckit*, terme usité dans la littérature économétrique (Greene 2003, par exemple). Nous supposerons dans ce cadre que y_i est continu, le cas où y_i est une variable binaire se généralisant à partir du cas continu. On supposera en outre, et sauf mention contraire, que l'instrument w_i est binaire.
Indépendamment des hypothèses paramétriques sur lesquelles il repose, la particularité du modèle *heckit* est de supposer l'existence d'une variable latente r_i^* qui ordonne les participations individuelles à l'enquête (i.e. à réalisation d'aléa ϵ_i^0 fixée). La forme retenue et l'existence de la variable latente, hormis sa linéarité, traduisent de manière assez logique le comportement individuel consistant à décider de participer ou non à l'enquête [10]. Du reste, la littérature scientifique sur les effets de sélection prend, la plupart du temps, l'existence de cette variable latente pour acquise, sans la remettre en cause, au motif qu'elle traduit dans l'essentiel de sa généralité, le comportement individuel à l'œuvre (Vytlacil 2002, Lee 2009).

Dans un modèle à variable latente, la participation de l'individu i découle directement de cette variable : si celle-ci est positive ou nulle, l'individu participe, sinon il ne participe pas. Dans le modèle *heckit*, la variable $\bar{r}_i^* = c^0 + \mathbf{z}_i\beta + \epsilon_i^0$ peut être vue comme une propension individuelle à participer à l'enquête, sous l'effort moyen de collecte, conditionnellement aux caractéristiques individuelles $\mathbf{z}_i$. La propension individuelle moyenne à participer sous l'effort moyen de collecte est ici caractérisée par $c^0 + \mathbf{z}_i\beta$. L'ajout d'un instrument – binaire en l'occurrence – traduit un effort de collecte accru pour le sous-ensemble des individus tels que $w_i = 1$. Et pour ceux-là, la participation est augmentée selon une valorisation de l'effort de collecte, en termes de participation, s'établissant à la valeur ψ (pour les notations se référer au modèle 13).

En pratique, connaître le lien entre y_i et r_i revient, soit à identifier la forme fonctionnelle qui lie ces deux variables, soit à identifier celle qui lie les variables y_i et r_i^*. Comme ces deux variables sont liées (c'est-à-dire que leurs aléas ne sont pas indépendants), l'observation du lien entre y_i et r_i ou r_i^* nécessite un instrument. L'instrument permet de « faire bouger » les deux variables de manière exogène. Et on peut ainsi identifier, au moins en partie, la relation qui relie ces fonctions.

Pour bien comprendre les conditions d'identification de cette relation, regardons plus précisément ce qui se passe dans le plan $(y_i, \bar{r}_i^*)$ où $\bar{r}_i^*$ est la propension individuelle à participer sous l'effort moyen de collecte et conditionnellement aux caractéristiques individuelles $\mathbf{z}_i$. Notons que cette propension individuelle n'est pas observée.

Une analyse similaire, développée en annexe B, peut être menée dans le plan $(y_i, \bar{\pi}_i)$, où $\bar{\pi}_i = \Pr(\bar{r}_i = 1) = \Pr(\bar{r}_i^* \geqslant 0)$ est la probabilité de participation sous l'effort moyen de collecte, probabilité qui est observée.

Seule est ici développée l'analyse dans le plan $(y_i, \bar{r}_i^*)$. Dans ce plan, tous les individus sont ordonnés par leur propension à participer $\bar{r}_i^* = c^0 + \mathbf{z}_i\beta + \epsilon_i^0$ sous l'effort moyen de collecte et conditionnellement à leurs caractéristiques individuelles $\mathbf{z}_i$. On suppose un instrument binaire $\mathbf{w}_i \equiv w_i \in \{0, 1\}$. A l'aide de l'instrument w_i, on peut écrire d'après (15) :

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

la dernière ligne correspondant au premier ordre du développement de Taylor de la fonction λ [11], calculée au point $c^0 + \mathbf{z}_i\beta$. Puis,

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

9. On note que ces conditions sont plus générales que celles du modèle d'Heckman, puisque nous ne faisons pas, à ce stade, d'hypothèse sur la distribution jointe des aléas. En somme, le modèle *heckit* se limite aux équations (13) et à l'existence d'une loi jointe des $(\epsilon_i^0, \epsilon_i^1)$ quelconque.

10. En termes de comportement, la forme de cette variable latente traduit assez naturellement l'existence d'une propension individuelle à participer à l'enquête considérée. Cette propension va différer selon les individus et elle va, très généralement, dépendre des caractéristiques individuelles et du contexte personnel de l'individu, ainsi que des caractéristiques de l'enquête. Il est aussi vraisemblable que des aléas, indépendants des conditions, jouent effectivement sur la participation. C'est le cas, par exemple, dans une enquête téléphonique, lorsque les appels-contacts de l'enquêté sont passés à son domicile alors que celui-ci en est absent.

11. approximation valable pour ψ petit. Il est utile de mentionner ici que la fonction λ est l'inverse du ratio de Mills dans le cas gaussien bivarié. Plus généralement, ici, on peut faire l'hypothèse qu'il s'agit d'une fonction quelconque, dépendant de la distribution jointe des aléas, sans qu'il soit nécessaire de la spécifier plus avant, les seules hypothèses importantes étant que $\mathbb{E}(\epsilon_i^1|\epsilon_i^0) = \varrho\sigma\epsilon_i^0$ et symétriquement, $\mathbb{E}(\epsilon_i^0|\epsilon_i^1) = \varrho\,\epsilon_i^1/\sigma$. C'est d'ailleurs une des pistes utilisées par les économètres pour généraliser, à des distribution bivariées quelconques des aléas, le raisonnement d'Heckman (Greene 2003, par exemple).

Il en découle que :

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

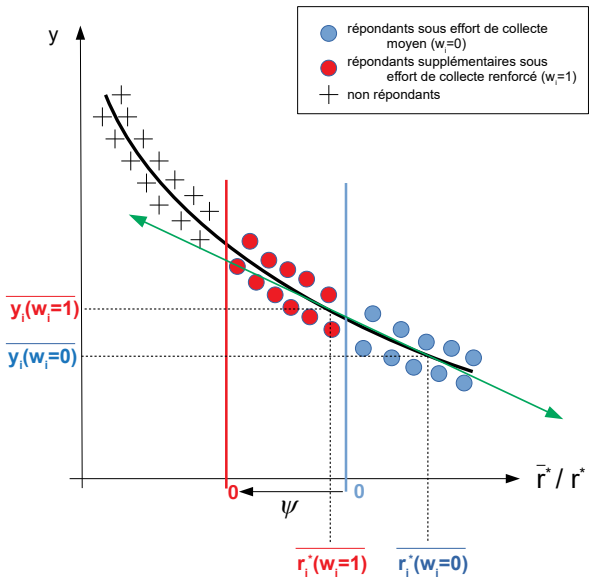
A l'aide des relations précédentes, il est possible de préciser la situation telle qu'elle se dessine dans le plan $(y_i, \bar{r}_i^*)$. Dans ce contexte, il est utile de noter que le modèle *heckit* n'explicite pas la relation $y(\bar{r}^*)$. Cependant, la relation (23) donne le coefficient directeur de la tangente à la courbe $y(\bar{r}^*)$, au point moyen entre les deux efforts de collecte.

On peut noter que l'expression de la pente de la courbe $y(\bar{r}^*)$ en ce point ne dépend pas de ψ, c'est-à-dire de la valorisation de l'effort de collecte supplémentaire en termes de surcroît de participation. Ceci confirme le rôle de ψ qui est assimilable ici à un paramètre exogène (au sens des fonctions mathématiques dépendant d'un paramètre) pour les deux fonctions implicites $\mathbb{E}(y_i|\mathbf{z}_i, \psi)$ et $\mathbb{E}(r_i^*|\mathbf{z}_i, \psi)$ de ψ. Comme ψ intervient de manière linéaire dans ces deux quantités (conformément au lien postulé entre les deux aléas des équations d'outcome et de participation), il disparaît au premier ordre dans la modélisation de $y(\bar{r}^*)$ [12].

La figure 1 montre un cas général dans lequel la relation entre y et $\bar{r}^*$ est décroissante. Cela correspondrait à une situation dans laquelle les hauts revenus seraient réticents à répondre à l'enquête, tandis que les bas revenus répondraient plus volontiers, par exemple. Dans ces circonstances, si un revenu est plus élevé que la moyenne, la probabilité que la personne considérée participe à l'enquête est plus faible, toutes choses égales par ailleurs. Donc dans le modèle *heckit*, la corrélation ϱ est négative.
Dans le plan (y, r^*), les individus sont ordonnés selon leur propension à participer sous l'effort moyen $\bar{r}_i^*$. Participent, sous l'effort moyen, tous ceux tels que $\bar{r}_i^* \geqslant 0$. Ces individus sont figurés en cercles bleus dans la figure 1. Puis, lorsque l'instrument est mis en œuvre, c'est-à-dire que l'effort de collecte est accru, l'origine des $\bar{r}_i^*$ se décale vers la gauche d'une quantité ψ (ici positive traduisant un surcroît de participation du fait de l'accroissement de l'effort de collecte), de sorte que participent désormais ceux dont la propension est telle que $\bar{r}_i^* + \psi \geqslant 0$. Aux répondants précédents, à l'aide de cet effort de collecte renforcé, s'ajoutent les répondants correspondant aux cercles rouges dans la figure 1. La différence de moyennes des y sur ces deux ensembles de répondants (point bleus d'un côté, et rouges et bleus de l'autre), rapportée à la différence des moyennes des r_i^* sur ces mêmes ensembles, donne la pente locale de la courbe $y(\bar{r}^*)$. C'est la dérivé de cette fonction. Elle est figurée en vert dans la figure 1.

Comme expliqué plus haut, l'instrument et son coefficient associé ψ jouent le rôle de paramètre pour les deux fonctions liées, y et $\bar{r}^*$. Ce paramètre permet d'identifier le lien. Ce dernier est *de facto* connu *localement*. En d'autres termes, si la fonction $y(\bar{r}^*)$ n'est pas linéaire, l'approximation donnée par le modèle d'Heckman n'est valable qu'au voisinage du taux de participation considéré. Si par exemple, l'effort de collecte renforcé (correspondant à l'instrument) augmente de 10 points un taux de réponse d'origine faible, disons de 30%, alors il est peu probable que l'approximation linéaire obtenue au voisinage d'une participation à 30-40% soit encore valide pour les non-répondants à hautes valeurs d'outcome, ainsi que le montre la figure 1.

Fig. 1. Courbe $y(r^*)$, répondants et non-répondants sous deux protocoles de collecte emboîtés



Note : $\overline{r_i^*(w_i = 0)}$ correspond à la valeur moyenne de r^* pour les répondants sous l'effort de collecte moyen (i.e. tel que $w_i = 0$). $\overline{r_i^*(w_i = 1)}$ correspond à la valeur moyenne de r^* pour les répondants sous l'effort de collecte renforcé (i.e. tel que $w_i = 1$). Formellement, dans le protocole renforcé, les individus répondants sont ceux qui répondent au protocole moyen (les points bleus), complétés de ceux qui répondent, du fait du surcroît d'effort de collecte (les points rouges). De la même manière, $\overline{y_i(w_i = 0)}$ est la valeur moyenne de y pour les répondants sous l'effort de collecte moyen et $\overline{y_i(w_i = 1)}$ est la valeur moyenne de y pour les répondants sous l'effort de collecte renforcé. Les individus représentés par des croix ne sont jamais observés; on sait seulement qu'ils ne participent ni sous l'effort moyen de collecte, ni sous le protocole renforcé. La courbe en noir est la vraie fonction $y(r^*)$, non observable. La droite en vert est la tangente à cette courbe, observée grâce à l'instrument (cf. texte).

En pratique, si on dispose d'un instrument binaire, la seule connaissance qu'on peut acquérir du lien entre y et $\bar{r}^*$ est locale. Il est possible d'acquérir une connaissance plus étendue (i.e. sur un support de y et de $\bar{r}^*$ plus large par rapport à leurs supports réels) en mettant en œuvre plusieurs protocoles permettant d'atteindre des niveaux de participation différenciés, du plus bas au plus haut. Un tel enchaînement de protocoles, appliqués sur des échantillons indépendants peut faire appel à une modélisation de la sélection identique au protocole de base, la seule différence étant l'application d'un instrument différent pour chaque sous-échantillon associé à un protocole donné. Par exemple si on met en œuvre trois protocoles associés à trois sous-échantillons $\left(G^{(k)}\right)_{k\in\{0,1,2\}}$, la participation croissant avec le numéro de l'échantillon, alors par rapport au modèle (13), seule l'équation (ii) est modifiée

12. C'est aussi la traduction statistique du fait que ce paramètre est associé à une variable instrumentale. Dans ce contexte, au plan différentiel, on a : $\Delta\mathbb{E}(y_i|r_i = 1, \mathbf{z}_i)/\Delta r_i^* = (\Delta\mathbb{E}(y_i|r_i = 1, \mathbf{z}_i, \psi)/\Delta\psi) \times (\Delta r_i^*(\psi)/\Delta\psi)^{-1}$.

en :

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

où $w_i^{(1)} = 1$ si $i \in G^{(1)} \cup G^{(2)}$ et $w_i^{(2)} = 1$ si $i \in G^{(2)}$. Dans ce modèle, ϱ, caractéristique de la sélection endogène, est le même pour tous les protocoles et l'éventuelle non-linéarité de la relation entre y et $\bar{r}^*$ passe par le coefficient ψ associé aux instruments du sous-échantillon $G^{(1)}$. En effet, à la relation (23) se substitue désormais deux relations (on note $\mathbf{w}_i = (w_i^{(1)}, w_i^{(2)})$) :

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

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

Si on veut pouvoir décrire plus aisément une non-linéarité de la fonction $y(\bar{r}^*)$, il est possible d'associer aux deux couples de protocoles $(G^{(0)}, G^{(1)})$ et $(G^{(1)}, G^{(2)})$, deux modes de sélection différents. Cela peut se faire en estimant deux modèles *heckit* séparément. Cela peut aussi se faire avec un seul modèle et deux équations de sélection différentes (Vella 1998, Ogundimu and Hutton 2016). Moyennant quoi, les coefficients ϱ apparaissant dans (24-(a – b)) sont estimés comme étant deux coefficients distincts.

V. Séparer les erreurs de mesure liées au mode de collecte et le biais de sélection endogène

Lorsqu'on utilise le modèle de Heckman avec comme variable instrumentale des différences de protocoles impliquant plusieurs modes de collecte, on fait jusqu'ici l'hypothèse qu'il n'existe pas d'effets de mesure sur la variable d'intérêt modélisée.

S'il y a une erreur de mesure, c'est-à-dire que les enquêtés répondent différemment, selon le mode de collecte, à la question à laquelle est associée la variable d'intérêt y_i, sans pour autant que cet effet résulte d'une auto-sélection des enquêtés, alors, dans certaines circonstances, on peut théoriquement identifier ces effets propres au mode. En effet, au plan de la modélisation de Heckman, un effet de mesure explique la variable y_i, tandis qu'il ne joue aucun rôle dans l'équation de sélection. Explicitons plus en détail ce point.

Supposons qu'il y ait $J+1$ modes de collecte d'enquête, notés $j \in \{0, 1, \ldots, J\}$ et que le mode 0 soit le mode de référence par rapport auquel on détermine l'erreur de mesure associée aux modes alternatifs $j \in \{1, \ldots, J\}$. Le modèle (13) peut alors être écrit :

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

où m_i désigne le mode[13] avec lequel l'individu i répond à l'enquête. Si i est non-répondant, alors m_i n'est pas observé. Sa valeur est donc conventionnelle et peut par exemple être supposée égale à 0, sans conséquence sur l'identification du modèle. Dans ce modèle, les α_j peuvent être identifiés lorsque la variable m_i n'est pas colinéaire aux instruments $\mathbf{w}_i$. L'hypothèse identifiante est donc que le mode n'explique pas le choix (ou n'est pas liée à celui-ci) de participer ou non à l'enquête.

De ce point de vue, il faut donc qu'une condition d'exclusion sur m_i soit vérifiée dans l'équation de sélection. C'est donc une contrainte forte qui nécessite, pour être vérifiée, de mettre en place un protocole très spécifique. Dans le cas général, il y a tout lieu de penser que le mode influe également sur la participation. Le moyen le plus naturel d'assurer l'indépendance du mode et de la participation est de réserver un sous-échantillon sur lequel on s'assure de la participation, préalablement à l'affectation randomisée à un mode de collecte. Ce faisant, on s'assure que la variable indicatrice du mode ne joue que dans y, et non dans la participation. Cependant, en pratique, un tel protocole est rarement mis en oeuvre car il est coûteux de s'assurer de la participation avant d'affecter un mode de collecte et qu'il n'est pas possible de s'assurer de l'absence de non-réponse une fois le mode de collecte affecté, malgré l'accord préalable.

Sous cette hypothèse où m_i est exogène vis-à-vis de l'équation de sélection, les $(\alpha_j)_{j\in\{1,\ldots,J\}}$, estimés par maximum de vraisemblance, estiment l'erreur de mesure associée aux modes alternatifs, par rapport au mode de référence.

Cette méthode est valable pour le cas continu comme pour le cas discret.

A l'inverse, si on introduit le mode m_i dans l'équation d'outcome sans s'être assuré préalablement que les conditions d'exclusion sont effectivement remplies, il est extrêmement probable que m_i soit endogène dans cette équation. En effet, le mode a sans doute une influence sur la participation. La non-inclusion de m_i dans l'équation de participation rejette le mode dans l'aléa de cette équation, lequel est par hypothèse corrélé avec celui de l'équation d'outcome. Par conséquent m_i, variable explicative de y_i, est vraisemblablement corrélée avec l'aléa de l'équation d'outcome[14]. Donc les coefficients de régression associés aux effets de mesure dans l'équation d'outcome sont biaisés. Et, à la différence du raisonnement tenu au paragraphe III où l'on s'intéressait à la prédiction du modèle, ce sont ici les coefficients associés aux effets de mesure qui nous intéressent.

Par conséquent, sans protocole *ad hoc* qui garantisse que les conditions d'exclusion sont remplies, il n'est pas possible d'identifier un effet de mesure dans l'équation d'outcome, en même temps qu'une sélection endogène.

On peut aller plus loin dans le raisonnement en montrant que s'il existe une erreur de mesure en même temps qu'un mécanisme de sélection endogène, rien n'est identifiable si le surcroît de participation résulte de l'ajout d'un mode de collecte supplémentaire : dans ce cas, ni l'erreur de mesure, ni la sélection endogène ne sont identifiables. En effet, si une erreur de mesure existe en même temps qu'un mécanisme de sélection endogène, alors la non-prise en compte, dans l'équation d'outcome, de l'erreur de mesure renvoie celle-ci

13. m_i prend donc ses valeurs dans $\{0, \ldots, J\}$ selon le mode de collecte associé à l'individu i.

14. sous l'hypothèse où une sélection endogène est suspectée, par exemple si on estime, en même temps qu'on estime l'effet de mesure, un coefficient de corrélation dans un modèle probit bivarié à la Heckman.

dans l'aléa ϵ_i^1, puis dans ϵ_i^0 *via* la corrélation qui existe entre les aléas du fait de la sélection endogène. Si à ce stade, les instruments de r_i^* sont liés au mode de collecte [15], alors les conditions d'exclusion du modèle d'Heckman ne sont plus vérifiées, en particulier dans l'équation de participation.

En synthèse :

— pour traiter de la sélection endogène, on peut combiner les modes de collecte pour un sous-échantillon sélectionné aléatoirement, mais ce mécanisme ne corrige la sélection endogène que sous l'hypothèse d'absence d'erreur de mesure liée au mode ;

— et dans le même temps, on ne peut pas corriger d'une erreur de mesure qui surviendrait simultanément à la sélection endogène, si les conditions d'exclusion nécessaires à l'identification des erreurs de mesure dans l'équation d'outcome ne sont pas vérifiées.

Par conséquent, la combinaison de modes de collecte pour traiter de la sélection endogène comporte un risque d'échec réel si une erreur de mesure associée au mode de collecte est suspectée. Dans ce cas, un instrument fondé sur des incitations accrues à participer, appliquées sur un sous-échantillon aléatoire, mais indépendant du mode de collecte (par exemple des incitations financières ou des efforts de relance significatifs supplémentaires), devra être privilégié.

15. Ce serait le cas lorsque la participation renforcée découle de l'ajout, au mode de collecte de référence, d'un mode alternatif.

REFERENCES

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

BÄRNIGHAUSEN, T., J. BOR, S. WANDIRA-KAZIBWE, AND D. CANNING (2011): "Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models," Epidemiology, pp. 27–35.

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. Quelques variations sur les lois conditionnelles

Dans cette annexe, on redémontre quelques résultats utiles sur les probabilités et espérances conditionnelles. Le but est de rappeler ces résultats au lecteur et de le familiariser avec des raisonnements qu'on retrouve au long du texte. Dans une première partie, on revient sur le conditionnement et l'orthogonalité conditionnelle de variables. Une seconde partie dérive des résultats sur le modèle normal bivarié, sous-jacents au modèle d'Heckman.

Sauf mention explicite, toutes les variables ici considérées sont vectorielles.
Toute variable aléatoire $\mathbf{y}$ est caractérisée par une densité de probabilité $f_{\mathbf{y}}(u)$. Celle-ci vérifie par définition : $\mathbb{P}(\mathbf{y} < y) = \int_{\{u<y\}} f_{\mathbf{y}}(u)du$. Deux variables $(\mathbf{y}, \mathbf{z})$ ont une loi jointe $f_{\mathbf{y},\mathbf{z}}(u, v)$. $\mathbf{y}|\mathbf{z}$ est une variable aléatoire (on dit "$\mathbf{y}$ sachant $\mathbf{z}$"), dont la densité (appelée "loi conditionnelle de "$\mathbf{y}$ sachant $\mathbf{z}$") vaut :

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

Les variables $\mathbf{y}$ et $\mathbf{z}$ sont indépendantes si et seulement si leur densité jointe est le produit de leurs densités marginales : $f_{\mathbf{y},\mathbf{z}}(u, v) = f_{\mathbf{y}}(u)f_{\mathbf{z}}(v)$. On note dans ce cas $\mathbf{y} \perp\!\!\!\perp \mathbf{z}$. Cette relation est symétrique, c'est-à-dire que $(\mathbf{y} \perp\!\!\!\perp \mathbf{z} \Leftrightarrow \mathbf{z} \perp\!\!\!\perp \mathbf{y})$. De manière équivalente,

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

Ce que l'on peut aussi écrire plus simplement :

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

On note que si $\mathbf{y} \perp\!\!\!\perp \mathbf{z}$, alors les variables $\mathbf{y}$ et $\mathbf{z}$ ne sont pas corrélées (i.e. ont une corrélation nulle). En effet, supposons sans perte de généralité $\mathbb{E}(\mathbf{y}) = \mathbb{E}(\mathbf{z}) = 0$ et calculons

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

On note au passage que

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

L'espérance d'une variable $\mathbf{y}$ est $\mathbb{E}(\mathbf{y}) = \int u f_{\mathbf{y}}(u)du$. On peut définir, de manière cohérente, l'espérance conditionnelle de $\mathbf{y}|\mathbf{z}$ par :

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

Là encore, il est d'usage d'adopter la notation simplifiée suivante :

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

On déduit de ce qui précède que

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

On note que l'expression $\mathbb{E}(\mathbf{y}|\mathbf{z})$ définit une variable aléatoire fonction de la variable aléatoire de $\mathbf{z}$. Par conséquent,

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

On en déduit la formule des espérances conditionnelles itérées, toujours vérifiée :

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

En revanche, pour pouvoir être calculée, il convient de connaître *in extenso* l'espérance conditionnelle $\mathbb{E}(\mathbf{y}|\mathbf{z})$ **et** la loi de la variable $\mathbf{z}$. On peut enchaîner les espérances conditionnelles par itération :

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

Il se peut enfin, que l'indépendance de variables soit conditionnelle à une autre variable. Par exemple, les variables $\mathbf{x}|\mathbf{z}$ et $\mathbf{y}|\mathbf{z}$ peuvent être indépendantes tandis que $\mathbf{x}$ et $\mathbf{y}$ ne le sont pas. Cette propriété prend son sens par rapport à ce qui précède. Par extension des notations précédentes :

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

Sous cette hypothèse, il est possible de séparer les calculs d'espérance. Soit une fonction séparable $g(\mathbf{x}, \mathbf{y}) = g_1(\mathbf{x})g_2(\mathbf{y})$ quelconque. On peut calculer son espérance conditionnellement à $\mathbf{z}$. Par hypothèse, $\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z}$ donc

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

Et, en particulier, en prenant $g(x, y) = xy$, on a :

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

Cette formule peut être vue comme une généralisation de la relation (26).
On remarque, grâce à cette dernière expression, que l'application de la formule des espérances conditionnelles itérées n'est pas immédiate. En effet, par les espérances itérées (relation 27), on a $\mathbb{E}\left[\mathbb{E}(\mathbf{xy}|\mathbf{z})\right] = \mathbb{E}(\mathbf{xy})$. En revanche, $\mathbb{E}\left[\mathbb{E}(\mathbf{x}|\mathbf{z})\ \mathbb{E}(\mathbf{y}|\mathbf{z})\right] \neq \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{y})$. En effet, pour que cette dernière propriété soit vraie, il faudrait que les variables $\mathbb{E}(\mathbf{x}|\mathbf{z})$ et $\mathbb{E}(\mathbf{y}|\mathbf{z})$ soient indépendantes, donc en particulier indépendantes de la variable $\mathbf{z}$ par rapport à laquelle les deux espérances conditionnelles sont ici déterminées.

Enfin, il peut être utile de noter que

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

En effet,

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

d'après (29), d'où le résultat annoncé.

Pour fixer les idées, considérons trois variables aléatoires, x, y et z. Supposons que, (a, b, α, β) étant un jeu de constantes,

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

avec $e \hookrightarrow \mathcal{N}(0, \sigma_e^2)$, $\varepsilon \hookrightarrow \mathcal{N}(0, \sigma_\varepsilon^2)$ et $z \hookrightarrow \mathcal{N}(\eta, \sigma^2)$, où $\mathcal{N}(\eta, \sigma^2)$ désigne la densité de la loi normale d'espérance η et de variance σ^2. On suppose en outre que $e \perp\!\!\!\perp z$ et $\varepsilon \perp\!\!\!\perp z$. Ces éléments nous permettent de préciser la loi du vecteur aléatoire $(x, y)'$, conditionnellement à z :

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

où ρ désigne la corrélation des variables e et ε.
Classiquement, compte-tenu de ce qui précède, la densité de la variable vectorielle $(x, y)|z$ est le produit des densités marginales de $x|z$ et $y|z$ si, et seulement si, $\rho = 0$. Au final, dans le cas présent, les équivalences suivantes sont vérifiées :

$$\rho = 0 \iff e \perp\!\!\!\perp \varepsilon \iff x \perp\!\!\!\perp y|z \qquad (34)$$

On peut établir les résultats suivants :

— $\mathbb{E}(x|z) = a + bz$ sans restriction. Dans les mêmes conditions, $\mathbb{E}(y|z) = \alpha + \beta z$.

— $\mathbb{E}(xy|z) = (a+bz)(\alpha+\beta z) = \mathbb{E}(x|z)\mathbb{E}(y|z)$ n'est vérifié que si, et seulement si, $x \perp\!\!\!\perp y|z$. En effet,

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

d'après (32) et (33). L'application des équivalences (34) permet de conclure.

— A l'aide de l'expression précédente, on peut aussi calculer, grâce à la formule des espérances conditionnelles itérées :

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

Puis, $\mathbb{E}(x) = a + b\eta$ et de même, $\mathbb{E}(y) = \alpha + \beta\eta$. Il en découle que :

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

Ainsi, la covariance de x et y est nulle si, et seulement si, x ou y est indépendant de z (i.e. $b = 0$ ou $\beta = 0$) ou z est certaine (i.e. $\sigma = 0$) et, simultanément à l'une ou l'autre des conditions précédentes, $\rho = 0$.

— Un autre point utile à observer, à ce stade, est que $\mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y)$ diffère de $\mathbb{E}\left[\mathbb{E}(xy|z) - \mathbb{E}(x|z)\mathbb{E}(y|z)\right]$. En effet, si on a bien $\mathbb{E}\left[\mathbb{E}(xy|z)\right] = \mathbb{E}(xy)$, il n'en est pas de même pour la deuxième composante de l'expression. En effet,

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

qui diffère évidemment de $\mathbb{E}(x)\mathbb{E}(y) = (a+b\eta)(\alpha+\beta\eta)$, sauf quand z est une variable certaine ou que x ou y est indépendant de z. Ceci vient du fait que la formule des espérances conditionnelles itérées ne s'applique pas à un produit de variables : l'espérance totale est (seulement) une application linéaire. On retrouve ici une propriété évoquée à l'avant dernier paragraphe de la partie précédente.

Avec le modèle de dépendance précédent, il est possible de poursuivre dans l'explicitation des lois et espérances conditionnelles. Ainsi, on peut déterminer la loi non conditionnelle du vecteur $(x, y)'$:

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

Il est donc possible de calculer $\mathbb{E}(x|y)$. On sait que pour toutes variables normales centrées-réduites u et v, de corrélation c, $\mathbb{E}(u|v) = cv$. On en déduit que :

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

On peut aussi calculer $\mathbb{E}(x|y, z)$. Partant de l'expression (32), on a :

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

Or $\mathbb{E}(e|y, z) = \mathbb{E}(e|\varepsilon = y - \alpha - \beta z)$. Puis, par hypothèses,

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

Or, comme précédemment, $\mathbb{E}(e|\varepsilon) = \rho\frac{\sigma_e}{\sigma_\varepsilon}\varepsilon$. Donc

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

Et finalement

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

Les résultats (35) et (38) sont intéressants. On peut noter les points suivants :

— La relation (38) montre que si $\rho = 0$, alors $\mathbb{E}(x|y, z)$ ne dépend pas de y. En d'autres termes, dans ce cas, $\mathbb{E}(x|y, z) = \mathbb{E}(x|z) = a + bz$. Pour autant, si $\rho = 0$, $\mathbb{E}(x|y)$ dépend quand même de y pour la partie qui « transite » par z. En effet, de (35) il vient :

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

Et il serait trop rapide, partant de (36), de passer de $\mathbb{E}\left[\mathbb{E}(x|y, z)|y\right] = a+b\mathbb{E}(z|y)+\mathbb{E}(e|y)$ à $\mathbb{E}(x|y) = a+b\eta$ qui serait, bien-sûr, faux, car :

1) d'une part, $\mathbb{E}(z|y) \neq \eta$; on peut montrer que $\mathbb{E}(z|y) = \eta + \frac{\beta\sigma^2}{\sigma_\varepsilon^2+\beta^2\sigma^2}(y - (\alpha + \beta\eta))$ qui dépend donc de y ;
2) et d'autre part, $\mathbb{E}(e|y) \neq 0$, ceci en vertu de la relation (37) et de ce qui précède.

— La relation (35) montre que même si x et y ne sont pas engendrées par une même variable z (i.e. $b = 0$ ou $\beta = 0$), la corrélation des aléas e et ε génère une dépendance conditionnelle puisque $\mathbb{E}(x|y)$ dépend effectivement de y.

— Dans le même temps, même si $\rho = 0$, l'espérance de x conditionnellement à y dépend de y, par le truchement de la variable z (point déjà mentionné plus haut).

— A partir de la relation (38), on retrouve $\mathbb{E}(x|z) = a+bz$. En effet, partant de (38), on a :

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

Or $\mathbb{E}(y|z) = \alpha + \beta z$, d'où le résultat.

B. Analyse des conditions d'identification de la sélection endogène dans le plan $(y_i, \bar{\pi}_i)$

Dans ce plan, les variables sont observables. Contrairement au cas présenté au paragraphe IV-C, les individus ne sont pas ordonnés sur l'axe des π puisque pour une même probabilité de participation, des individus peuvent ou non participer à l'enquête (selon la réalisation de l'aléa ϵ_i^0 apparaissant dans l'équation de participation). Dans ce plan, la relation

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

vue précédemment est toujours vraie. Puis, d'après la relation (18), on a, en utilisant un développement de Taylor au premier ordre :

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

où φ est la densité de la loi normale centrée-réduite. Il en découle que :

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

Comme précédemment, ψ joue le rôle d'un paramètre pour les deux fonctions $\mathbb{E}(y_i|r_i = 1, \mathbf{z}_i, \psi)$ et $\pi_i(\psi)$. La situation correspondante est présentée à la figure 2. L'expression précédente correspond à la pente de la droite représentée en vert dans la figure. Cette droite, en tant qu'approximation de la courbe $y(\bar{\pi})$ au voisinage de $(\overline{\pi(w=0)}, \overline{y(w=0)})$ est observable puisque les valeurs moyennes de y et π le sont. En revanche, comme précédemment, la tangente précédente ne donne qu'une approximation de la courbe pour les taux de collecte observés. Dans l'exemple de la figure, l'approximation obtenue pour des taux de participation faibles comporte des risques de ne pas être valide pour des taux plus élevés. La discussion du paragraphe IV-C sur le plan $(y_i, \bar{r}_i^*)$ s'applique également dans le cas considéré ici.

C. Simulations à l'aide du package R sampleSelection

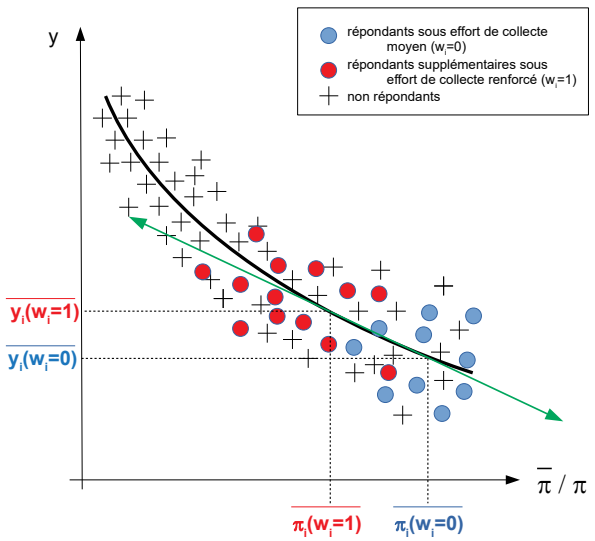
Dans cette annexe, on présente des simulations réalisées à partir d'observations synthétiques donnant lieu à sélection endogène et les résultats obtenus à l'aide de modèles de Heckman ajustés sur ces observations.

Tous les résultats présentés ici sont obtenus à partir du programme R NRC-Heck-model.R [16]. Ce programme fait appel au package sampleSelection [17] (Toomet and Henningsen 2008).

16. Accessible sous GitHub à l'adresse suivante : `https://github.com/InseeFrLab/NRC-heck-model`

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

Fig. 2. Courbe $y(\pi)$, répondants et non-répondants sous deux protocoles de collecte emboîtés



Note : $\overline{\pi_i(w_i = 0)}$ correspond à la valeur moyenne de π pour les répondants sous l'effort de collecte moyen (i.e. tel que $w_i = 0$). $\overline{\pi_i(w_i = 1)}$ correspond à la valeur moyenne de π pour les répondants sous l'effort de collecte renforcé (i.e. tel que $w_i = 1$). Formellement, dans le protocole renforcé, les individus répondants sont ceux qui répondent au protocole moyen (les points bleus), complétés de ceux qui répondent, du fait du surcroît d'effort de collecte (les points rouges). De la même manière, $\overline{y_i(w_i = 0)}$ est la valeur moyenne de y pour les répondants sous l'effort de collecte moyen. $\overline{y_i(w_i = 1)}$ est la valeur moyenne de y pour les répondants sous l'effort de collecte renforcé. Les individus représentés par des croix ne sont jamais observés; on sait seulement qu'ils ne participent ni sous l'effort moyen de collecte, ni sous le protocole renforcé. La courbe en noir est la vraie fonction $y(\pi)$, non observable. La droite en vert est la tangente à cette courbe, observée grâce à l'instrument (cf. texte).

1) Construction de la population synthétique et autosélection: On construit une population de 10 000 individus dont le revenu dépend de trois variables exogènes x_1, x_2, x_3, chacune de ces variables étant tirée dans une loi uniforme $\mathscr{U}$, sur respectivement $[2, 5]$, $[0, 2]$, et $[0, 1]$. Le revenu est obtenu par la relation suivante :

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

où ϵ_i est tirée dans une loi normale $\mathscr{N}(0, 2^2)$.

On en déduit que la moyenne vraie des revenus est de 7.75 et l'écart-type vrai de la moyenne arithmétique d'un échantillon de 10 000 individus iid est de 0.0277.

Une distribution empirique d'un vecteur de 10 000 revenus $(y_i)_{i\in\{1,\ldots,10\ 000\}}$ est donnée à la figure 3. La moyenne simulée associée au tirage de ces 10 000 individus est de 7.74. L'écart-type de la moyenne est de 0.0274.

Un mécanisme de sélection endogène est simulé, sur la base du revenu précédent, pour les 10 000 individus simulés, conformément à la relation suivante, de façon à ce que la

Fig. 3. Histogramme du revenu simulé

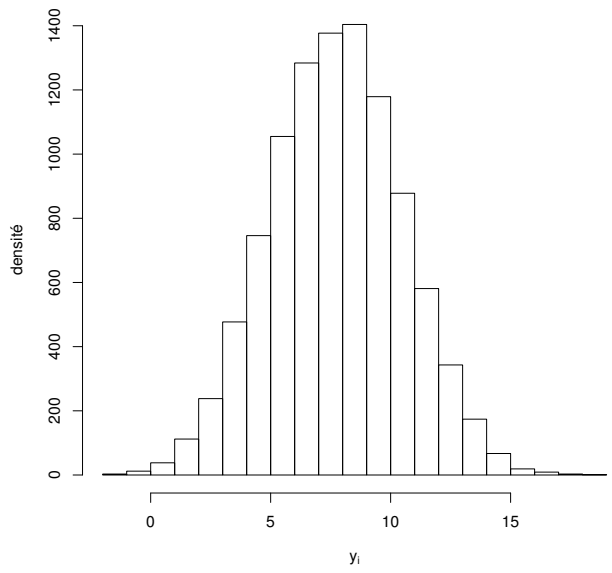
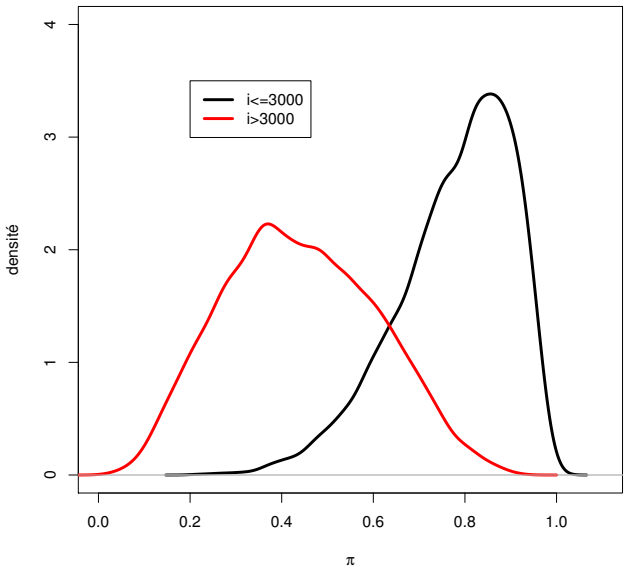


Fig. 4. Distribution des probabilités de réponse simulées



Note : distribution des probabilités de réponse, en noir pour les 3000 premiers individus de l'échantillon, c'est-à-dire ceux pour lesquels la probabilité de participer est renforcée (en lien avec le terme instrumental en $0.2 \times \mathbb{1}(i \leqslant 3000)$ dans l'expression (39), et en rouge, pour les 7000 individus restants. En application de (40), $\pi_i = \Phi\left[(-0.4 + (\max(y) - y_i)/30 + 0.2 \times \mathbb{1}(i \leqslant 3000))/0.2\right]$. Densités estimées par méthode à noyau Gaussien (Silverman 1986).

participation décroît avec le revenu :

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

où $\max(y)$ désigne le maximum observé sur les $(y_i)_{i \in \{1,\ldots,10\ 000\}}$ et ν_i est un aléa tiré dans une loi normale $\mathcal{N}(0, 0.2^2)$. Le nombre de répondants $n_r = \sum_i r_i$; la moyenne des n_r vaut 4 550 et, dans le tirage simulé, elle vaut 4 518. D'après (40), les 3 000 premiers individus, classés de 1 à 10 000, ont leur variable latente de participation renforcée, par rapport aux 7 000 individus suivants. Cette indicatrice d'appartenance aux 3 000 premiers individus est conforme à la définition d'un instrument puisqu'elle explique la participation accrue de ces 3 000 individus, sans que cet instrument ne joue dans la formation du revenu. La figure 4 donne le tracé des distributions des probabilités de participation simulées, selon que les individus sont dans le groupe à participation renforcée (auquel est associé l'instrument), ou non.

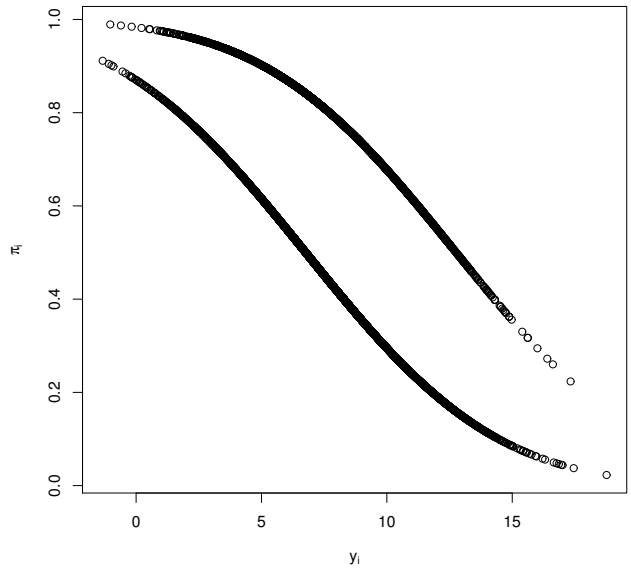
La figure 5 donne le tracé de la probabilité de réponse en fonction du revenu pour l'ensemble des 10 000 individus. On note, comme attendu, la décroissance de la fonction obtenue et la séparation des deux groupes, selon que l'individu est dans le groupe des 3 000 individus affectés par le surcroît de probabilité de réponse à laquelle sera associée la variable instrumentale, ou bien que l'individu est parmi les 7 000 restants.

Cette figure peut être comparée à la figure 2, sous réserve de deux aménagements. En premier, il convient d'échanger abscisses et ordonnées des deux figures. En second, la courbe tracée dans la figure 2 se réfère, en abscisse, à la probabilité de réponse $\bar{\pi}$ sous l'effort moyen de collecte, c'est-à-dire à valeur de l'instrument nulle. *Stricto sensu*, l'instrument, lorsqu'il est non nul, augmente la probabilité de participer. Il décale donc la courbe de probabilité d'inclusion sous l'effort de collecte accru, vers le haut, à y donné, dans le système d'axes de la figure 5. C'est précisément ce qu'on observe dans cette dernière figure : on y représente, dans le même graphique, les probabilités non instrumentées (i.e. sous l'effort de collecte moyen) et les probabilités instrumentées (i.e. sous l'effort de collecte accru). Dans la figure 2, on a choisi, au contraire, de conserver, en représentation, la probabilité sous l'effort de collecte moyen, en figurant les points supplémentaires de répondants, obtenus du fait de l'instrumentation, à l'aide d'une couleur différente (rouge en l'occurrence) de celle des répondants de l'échantillon sans instrumentation (bleu).

2) Estimateurs et variance: A partir de la simulation de population précédente, on étudie l'espérance et la variance de différents estimateurs portant, d'une part sur la population simulée en l'absence de non-réponse, et d'autre part sur les seuls répondants, en lien avec la variable de participation simulée. Les estimateurs sur les répondants sont soit des estimateurs de Hajek – biaisés par construction – , soit des estimateurs d'Horvitz-Thompson fondés sur des probabilités d'inclusion issues de modèles d'Heckman en une ou deux étapes, obtenus par repondération des répondants, conformément à la relation (19), ou par imputation des réponses des non-répondants, conformément à la relation (17). Différents modes d'estimation de variance sont utilisés : par génération multiple de la population simulée, ou par bootstrap sur une simulation particulière de la population des 10 000 individus, l'estimation par bootstrap étant naturellement la seule par simulation pouvant être utilisée, en pratique, lorsque l'on travaille sur un échantillon réel.

La table I donne les résultats des différentes simulations

Fig. 5. Tracé des probabilités de réponse simulées en fonction du revenu



Note : En application de (40), $\pi_i = \Phi\left[(-0.4 + (\max(y) - y_i)/30 + 0.2 \times \mathbb{1}(i \leqslant 3000))/0.2\right]$. La courbe supérieure (en haut à droite) correspond aux individus à probabilité de participer renforcée par l'instrument.

réalisées. Plusieurs points méritent d'être soulignés concernant les résultats présentés dans cette table.

- On observe comme prévu le caractère fortement biaisé du revenu estimé sur les répondants (cf. colonne « Moyenne » de la table I : comparer la ligne (b) à la ligne (a)). L'écart entre la valeur vraie (7.74) et la valeur estimée sur les répondants (6.81) est très nettement plus large que les intervalles de confiance portant sur les estimateurs (de demi-amplitude 0.1 point).
- L'estimateur de Hajek utilisant les probabilités d'inclusion vraies des répondants, c'est-à-dire celles calculées à l'aide de la vraie distribution de la variable latente, comme à la figure 4, est sans biais (ligne (c)). Cet estimateur est en revanche plus incertain que celui portant sur l'ensemble de la population (ligne (a)). Il est aussi plus incertain que l'estimateur de Hajek sur les répondants (ligne (b)), en raison de la dispersion accrue des poids (écart-type de 0.0515 contre 0.0383).
- Les estimateurs de Heckman (lignes (d-g)), obtenus, en une ou deux étapes, par repondération des répondants ou par imputation des non-répondants, sont non-biaisés, au regard de leurs intervalles de confiance. Les écart-types associés sont plus élevés que l'écart-type de référence pour la moyenne des revenus sur la population : le rapport des écart-types est environ de 3, par rapport à une situation sans non-réponse. Ainsi – et c'est naturel – la non-réponse endogène et son traitement ont un coût se traduisant par une perte de précision.
- Les estimateurs d'écart-type sont eux-mêmes sujets à imprécision. La comparaison des colonnes (2) et (3) par rapport à la colonne (1) de la table I pour les estimateurs des lignes (a-b) donne la signature de cette imprécision. Compte-tenu de l'ordre de grandeur de cette imprécision, les estimateurs bootstrap et ceux obtenus par simulation de population sont compatibles.
- Sous la réserve de l'imprécision des estimateurs d'écart-types obtenus par simulation, l'estimateur de Heckman par repondération est plus précis (20% environ) que celui obtenu par imputation des non-répondants (comparer les lignes (d) et (e) du tableau, puis (f) et (g)). Ceci est vrai, que les estimateurs d'Heckman soient fondés sur un modèle en une étape ou en deux étapes.
- Les estimateurs fondés sur des modèles d'Heckman en une étape sont plus précis (5 à 10%) que les estimateurs fondés sur des modèles en deux étapes, ceci que la correction soit réalisée par repondération (comparer les lignes (d) et (f)) ou par imputation (comparer les lignes (e) et (g)). Ce résultat est lié à l'efficacité accrue de l'estimateur par maximum de vraisemblance (i.e. en une étape) par rapport à l'estimateur en deux étapes.
- S'agissant du temps de calcul des boucles bootstrap, l'expérience montre que l'écart-type converge assez lentement. Les résultats sont stables à partir d'un nombre de boucles bootstrap supérieur à 10 000. Aussi, le calcul est assez long puisque l'estimateur de Heckman nécessite, à chaque boucle bootstrap, de réaliser l'optimisation d'une vraisemblance. Cette vraisemblance est plus complexe dans le cas en une étape, donc plus longue à calculer, que dans le cas en deux étapes. Bien que l'estimateur en une étape soit plus efficace, dans ce contexte, l'estimateur en deux étapes peut lui être préféré puisque le calcul de l'estimateur de Hajek corrigé par modèle d'Heckman en une étape dure, dans le cas présent, 2 fois plus longtemps.

TABLE I

ESTIMATEURS SIMULÉS ET VARIANCES ASSOCIÉES

	Estimateur	Effectif	Moyenne	Ecart-type			$\hat{\varrho}$	Remarque
				(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 = (proba. vraie d'inclusion)$^{-1}$
(d)	$\hat{\mu}_{\text{Heckman-1Et.}}$	4518 / 10000	7.75		0.0676	0.0722	-0.363	repondération (†)
(e)	$\hat{\mu}_{\text{Heckman-1Et.}}$	4518 / 10000	7.77		0.0878	0.0848		imputation
(f)	$\hat{\mu}_{\text{Heckman-2Et.}}$	4518 / 10000	7.75		0.0733	0.0792	-0.368	repondération (†)
(g)	$\hat{\mu}_{\text{Heckman-2Et.}}$	4518 / 10000	7.78		0.0927	0.0881		imputation

Note : n = 10 000 ; dans les équations, les notations font référence aux relations (39) et (40). La colonne « Moyenne » est la velaur estimée pour une génération particulière de la population synthétique, la même que celle ayant servi aux figures 3, 4 et 5. (1) : application de la formule de variance analytique ; (2) : variance par bootstrap (20 000 boucles) dans un tirage particulier de la population de référence simulée, le même que celui correspondant aux lignes (a) à (c) du tableau ; (3) : variance par génération multiple (2 000 simulations) de la population simulée (assimilable à une variance simulée vraie) – dans ce cas, l'effectif de répondants est en moyenne de 4 550 individus ; (4) coefficient de corrélation des aléas dans le modèle de Heckman (cf. relation (13), par exemple), l'écart-type issu de l'estimation par maximum de vraisemblance s'élevant à 0.0460 ; tester l'absence de sélection endogène revient à tester la nullité de ce coefficient ; (†) : winsorisée pour les probabilités d'inclusion prédites inférieures à 0.1, lesquelles sont donc retraitées pour être saturées à ce niveau.

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD,
X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS
K. MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS
K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR
R. LE SAOUT
B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD
M. GUILLERM

M2016/06 : Econométrie spatiale : une introduction pratique
P. GIVORD
R. LE SAOUT

M2016/07 : La gestion de la confidentialité pour les données individuelles
M. BERGEAT

M2016/08 : Exploitation de l'enquête expérimentale Logement internet-papier
T. RAZAFINDROVONA

M2017/01: Exploitation de l'enquête expérimentale Qualité de vie au travail
T. RAZAFINDROVONA

M2018/01: Estimation avec le score de propension sous R
S. QUANTIN

M2018/02: Modèles semi-paramétriques de survie en temps continu sous R
S. QUANTIN

M2019/01 : Les méthodes de décomposition appliquées à l'analyse des inégalités
B. BOUTCHENIK
E. COUDIN
S. MAILLARD

M2020/01 : L'économétrie en grande dimension
J. L'HOUR

M2021/01 : R Tools for JDemetra+ - Seasonal adjustment made easier
A. SMYK
A. TCHANG

M2021/02 : Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman
L. CASTELL
P. SILLARD