Journées de méthodologie statistique de l'Insee 2022

Le traitement du biais de sélection endogène dans les enquêtes multimodes par modèle de Heckman

Laura Castell, Patrick Sillard (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

laura.castell@insee.fr ; patrick.sillard@insee.fr

Mots-clés. (6 maximum) : sélection endogène, multimode, instrument, modèle d'Heckman.

Domaines. Non-réponse ; multimode.

Résumé

Dans la plupart des enquêtes auprès des ménages, les méthodes de correction de la non-réponse mises en œuvre corrigent la sélection sur observables. Elles font en effet l'hypothèse que le fait de répondre à l'enquête – lié pour partie à l'intérêt pour la thématique de l'enquête – est un mécanisme ignorable (missing at random, ou MAR), c'est-à-dire que la non-réponse est indépendante des variables d'intérêt, conditionnellement aux caractéristiques (observées dans la base d'enquête) prises en compte dans le modèle de correction.

Dans un certain nombre d'enquêtes auprès des ménages, on peut suspecter un phénomène de sélection endogène, c'est-à-dire que des déterminants des variables d'intérêt – tels que l'intérêt pour la thématique de l'enquête – influent également sur la participation, conditionnellement aux caractéristiques observables. Dans ce cas, le mécanisme de non-réponse est non-ignorable et les méthodes usuelles de correction de la non-réponse conduisent à des estimations biaisées. 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.

Dans le cadre du développement de la collecte par Internet dans les enquêtes auprès des ménages, cette problématique de la sélection endogène devient a priori plus prégnante. De fait, le mécanisme de réponse par Internet nécessite une démarche plus pro-active que des modes de collecte intermédiés par un enquêteur. Il est donc plus probable que la non-réponse fasse davantage l'objet d'une auto-sélection, très liée à l'intérêt pour la thématique de l'enquête, et que cet intérêt ne soit pas entièrement corrélé à des caractéristiques observables. Ce biais éventuel risque d'être d'autant plus marqué que les taux de réponse des enquêtes par Internet restent relativement faibles.

Dans cet article, nous étudions les conditions dans lesquelles une correction de la sélection endogène peut être réalisée et en particulier quelle configuration d'enquête autorise la mise en œuvre d'une correction.

Les méthodes examinées sont basées sur le modèle d'Heckman et nécessitent de disposer d'instruments issus d'un protocole de collecte adapté en amont.

Usuellement, en économétrie, la correction de la sélection endogène est réalisée par imputation de la variable d'intérêt pour les non-répondants. De manière similaire, on peut réaliser cette correction par repondération en utilisant la probabilité d'inclusion conditionnelle du modèle. Cette méthode inédite correspond davantage aux usages de correction de la non-réponse dans la production d'enquêtes auprès des ménages dans la statistique publique. C'est celle qui est développée dans l'article.

La mise en œuvre d'une telle méthode suppose l'existence d'un instrument expliquant la participation et non la variable d'intérêt, et vérifiant le principe de monotonie.

De tels instruments peuvent être créés en adaptant les protocoles d'enquêtes en amont de la collecte. Une façon de procéder consiste à décomposer l'échantillon d'enquête en plusieurs sous-échantillons tirés aléatoirement et administrés selon des protocoles différents impliquant des taux de participation distincts. L'utilisation de différents modes de collecte fait partie des pistes d'adaptation de protocoles pour se retrouver dans une telle situation instrumentale.

Au-delà de la correction de la sélection endogène, les effets de mode peuvent survenir en lien avec l'existence d'erreur de mesure liées au mode de réponse lui-même. Dans ce cas, erreur de mesure et sélection endogène se confondent et séparer les deux peut se révéler compliqué.

L'article discute différents cas de figure et pistes envisageables pour traiter des effets de mode (sélection endogène et erreur de mesure) dans des enquêtes mobilisant des protocoles multimodes.

Abstract en Anglais

In most household surveys, methods for correcting non-response assume an ignorable (i.e. missing at random) mechanism. However, when there is an endogenous non-response problem, then the non-response mechanism is no longer ignorable, and the estimators derived from conventional correction methods are biased. In the context of the development of Internet collection in household surveys, this problem of endogenous selection becomes more prevalent.

To correct this bias, we propose a weighting based on a Heckman model. This method requires instruments from an adapted collection protocol. The use of independent sub-samples with different modes of data collection is a good way to adapt the protocol in order to provide such an instrument. This paper details the conditions under which this type of protocol allows an estimate corrected for endogenous selection.

Introduction

La participation à une enquête dépend de nombreux facteurs, à la fois liés aux caractéristiques de l'enquête (commanditaire, modes de collecte, thématique, durée, etc.) mais aussi aux caractéristiques individuelles des enquêtés, dont leur intérêt pour la thématique de l'enquête (Groves *et alii*, 2000). Dans la plupart des enquêtes auprès des ménages, la participation est supposée indépendante des variables d'intérêt de l'enquête, conditionnellement à des observables. Cependant, si ce n'est pas le cas et que l'intérêt pour la thématique de l'enquête influe, conditionnellement aux observables, sur la participation et les variables d'intérêt, on se trouve alors face à un problème de sélection endogène. Dans ce cas, les méthodes de correction de la non-réponse usuelles sont biaisées.

Cet article propose d'utiliser le modèle de Heckman pour identifier et corriger cette sélection endogène. Cette démarche nécessite de disposer d'un protocole adapté répondant aux critères du modèle. En pratique, l'utilisation de protocoles multimodes peut permettre de mettre en œuvre cette démarche mais sous certaines hypothèses, qui seront présentées. Cet article présente de façon synthétique la problématique de la sélection endogène dans les enquêtes auprès des

ménages et les pistes envisagées pour l'identifier et la corriger. Il est issu d'un document de travail (Castell et Sillard, 2021) qui présente des démonstrations plus complètes.

1 La sélection endogène dans les enquêtes auprès des ménages

Dans la plupart des enquêtes auprès des ménages, les modèles de correction de la non-réponse font l'hypothèse que la participation est indépendante des variables d'intérêt de l'enquête conditionnellement aux caractéristiques observables mobilisées dans les modèles de correction. Cette hypothèse est acceptable dans de nombreuses enquêtes pour lesquelles les déterminants de la participation, et notamment l'intérêt pour la thématique de l'enquête, sont fortement corrélés à des caractéristiques observables classiques, comme l'âge, le sexe, le revenu, etc.

Cependant, dans un certain nombre d'enquêtes, cette hypothèse peut s'avérer forte et l'intérêt pour la thématique de l'enquête peut influer à la fois sur la participation et sur les variables d'intérêt, y compris après contrôle des caractéristiques observables. On se trouve alors face à un phénomène de sélection endogène.

Ce problème émerge avec plus de prégnance face à l'évolution des enquêtes auprès des ménages vers des protocoles multimodes, et surtout l'usage de modes de collecte auto-administrés comme Internet. De fait, la démarche de participation par Internet est davantage pro-active qu'avec des modes intermédiés et un contact avec un enquêteur cherchant à convaincre l'ensemble des individus, augmentant ainsi le risque d'une sélection endogène.

1.1 Le modèle de correction de la non-réponse sur observables

Soit y une variable d'intérêt dont on cherche à estimer la moyenne μ sur une population P. Dans une enquête, cette moyenne n'est pas observée puisque seuls certains individus i sont échantillonnés. On note s_i l'indicatrice valant 1 si l'individu i est échantillonné, 0 sinon. La probabilité d'inclusion dans l'échantillon de l'individu i, notée π_i, est connue et peut dépendre d'un certain nombre de variables $\mathbf{Z}$ connues pour l'ensemble de la population P.

Par ailleurs, parmi les individus échantillonnés, seuls certains acceptent de répondre à l'enquête. On note r_i l'indicatrice valant 1 si l'individu i accepte de répondre à l'enquête, 0 sinon. La réponse à l'enquête est donc caractérisée par le produit $s_i r_i$.

Pour disposer d'une estimation sans biais de la moyenne de y sur la population P, on cherche un estimateur de la forme :

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

où $\hat{\rho}_i$ jouerait le rôle d'un modèle de r_i, de sorte que $E(\hat{\mu}^1|\mathbf{y}) = \mu$.

Pour que cette modélisation conduise à une estimation sans biais, elle doit respecter les hypothèses suivantes :
$H1$ - L'échantillonnage ne dépend que des variables $\mathbf{Z}$: $\mathbf{s} \perp\!\!\!\perp \mathbf{y}|\mathbf{Z}$
$H2$ - L'échantillonnage ne dépend que des variables $\mathbf{Z}$, il n'est pas affecté par la participation r_i puisqu'il est déterminé *ex ante* : $s_i \perp\!\!\!\perp r_i|(\mathbf{y}, \mathbf{Z})$
$H3$ - La participation est indépendante de la variable d'intérêt conditionnellement aux observables $\mathbf{Z}$: $r_i \perp\!\!\!\perp y_i|\mathbf{Z}$; et la modélisation $\hat{\rho}_i$ à partir des observables $\mathbf{Z}$ estime sans biais la participation r_i.

Sous ces trois hypothèses, $\hat{\mu}^1$ est un estimateur sans biais. Il est alors possible de modéliser la participation à partir de caractéristiques observables $\mathbf{Z}$ pour corriger la non-réponse totale.

C'est ce qui est fait par les méthodes usuelles de correction de la non-réponse dans les enquêtes auprès des ménages.

Si les deux premières hypothèses sont vérifiées avec un échantillonnage aléatoire dont les spécifications sont connues, ce n'est pas forcément le cas de la troisième hypothèse. Celle-ci correspond au cas où la non-réponse est ignorable (MAR, ou *Missing at random*), c'est-à-dire qu'elle est indépendante de la variable d'intérêt, conditionnellement aux observables.

1.2 Le problème de la sélection endogène

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 :

$$r_i = c + \mathbf{z}_i\beta + \xi_i + u_i \tag{2}$$

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

$$\tilde{\rho}_i = \tilde{c} + \mathbf{z}_i\tilde{\beta} + \tilde{u}_i \tag{3}$$

Deux situations peuvent poser problème. Tout d'abord, supposons que la variable omise ξ_i est endogène et dépend des variables $\mathbf{z}_i$. Dans ce cas, les coefficients du modèle 3 sont biaisés. Cependant, l'estimation de la probabilité de participer $\tilde{\rho}_i$ quant à elle n'est pas biaisée. Le mécanisme de non-réponse reste ignorable, malgré l'omission de cette variable.

Supposons maintenant que la sélection est endogène, c'est-à-dire que la variable omise ξ_i dépend de la variable d'intérêt y_i :

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

Dans ce cas, la première partie de l'hypothèse $H3$ n'est plus vérifiée puisque r_i dépend de y_i, par l'intermédiaire de ξ_i. Le mécanisme de non-réponse est dit non ignorable (MNAR, ou *Missing not at random*). Contrôler par les observables $\mathbf{Z}$ ne suffit pas à rendre indépendante la relation entre la participation et la variable d'intérêt.

Sous les autres hypothèses, on peut trouver qu'asymptotiquement :

$$E(\hat{\mu}^1|\mathbf{y}) = \mu + \vartheta\frac{1}{N}\sum_{i=1}^{N} y_i^2 E\left[\, 1/\tilde{\rho}_i(\mathbf{z}_i) \,|\, y_i\right] \tag{4}$$

Le biais lié à l'existence d'une sélection endogène est donc du signe de la corrélation avec la variable d'intérêt (ϑ). Il est d'autant plus important que cette corrélation est forte, que le taux de réponse est faible et que la variance de la variable d'intérêt sur P est élevée.

Le problème de la sélection endogène est donc particulièrement préoccupant lorsque les déterminants de la participation sont peu corrélés à des caractéristiques observables classiques et que les taux de réponse à l'enquête sont faibles. Il faut alors être particulièrement vigilant à cette problématique pour les enquêtes avec un protocole peu incitatif, qui mobilise les personnes les plus volontaires et intéressées par la thématique de l'enquête, et avec des taux de réponse faibles. C'est particulièrement le cas d'enquêtes réalisées essentiellement par Internet, qui suppose une démarche plus pro-active de la part des enquêtés. Les enquêtes qui traitent d'une thématique mal expliquée par des caractéristiques observables peuvent également être sujettes à cette problématique, comme c'est le cas de l'enquête EpiCov menée par la Drees et l'Inserm portant sur la pandémie de la Covid-19 (Castell *et alii*, à paraître).

2 Correction de la sélection endogène par modèle de Heckman

La correction de la non-réponse endogène est un problème bien connu dans la littérature sur les données manquantes. Plusieurs méthodes, plus ou moins paramétriques, existent. Ici, nous proposons d'utiliser le modèle de Heckman. Des modélisations avec d'autres distributions pour la corrélation des aléas ou des modélisations moins paramétriques existent et peuvent éventuellement être mobilisées pour traiter ce problème (Tchetgen Tchetgen et Wirth, 2017).

2.1 Le modèle de Heckman

Le modèle de Heckman (Heckman, 1979) consiste à modéliser simultanément la variable d'intérêt y_i et la participation r_i sous la forme suivante :

$$\left\{\begin{array}{lrcl} \text{(i)} & y_i & = & c^1 + \mathbf{z}_i\chi + \epsilon_i^1 \\ \text{(ii)} & r_i^* & = & c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0 \\ \text{(iii)} & r_i & = & \mathbb{1}(r_i^* \geqslant 0) \end{array}\right. \tag{5}$$

r_i^* est une variable latente qu'on n'observe pas. On observe y_i lorsque $r_i = 1$. $(\mathbf{z}_i, \mathbf{w}_i)$ est observé pour tout i. Ici, la variable d'intérêt y est continue. Il est également possible de traiter le cas des variables binaires en remplaçant l'équation 5-(i) par une variable latente (Castell et Sillard, 2021).

Dans ce modèle, l'équation de participation (5-(iii)) repose sur une variable latente r_i^* (relation (5-(ii)) qui fait intervenir les explicatives de y_i ainsi que des instruments $\mathbf{w}_i$, c'est-à-dire des variables qui expliquent la participation mais qui ne sont pas explicatives des y_i (on parle de conditions d'exclusion).

Formellement, les conditions d'identification du modèle précédent sont :

$$\left\{\begin{array}{l} E\left(\begin{pmatrix}\epsilon_i^0 \\ \epsilon_i^1\end{pmatrix}\middle|\, \mathbf{z}_i, \mathbf{w}_i\right) = 0 \\ \begin{pmatrix}\epsilon_i^0 \\ \epsilon_i^1\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \Sigma\right) \\ \Sigma = \begin{pmatrix}1 & \varrho\sigma \\ \varrho\sigma & \sigma^2\end{pmatrix} \end{array}\right. \tag{6}$$

Σ est une matrice de variance-covariance. C'est donc à travers cette matrice qu'est modélisée la formation simultanée de la variable d'intérêt y_i et de la participation r_i. Pour la suite, supposons une variance σ unitaire. Les coefficients de l'équation (5-(ii)) étant identifiables à un facteur multiplicatif près, cela n'a pas d'incidence sur les résultats.

La sélection endogène survient donc lorsque la corrélation ϱ entre les aléas des deux équations (5-(i) et 5-(ii)) est non nulle. Si $\varrho > 0$, alors le biais est positif, c'est-à-dire que les y_i observés sont, toutes choses égales par ailleurs, plus grands que les y_i non observés. A l'inverse, si $\varrho < 0$, alors le biais est négatif.

Le modèle de Heckman permet de calculer $E(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i)$. A partir de $E(y_i|\mathbf{z}_i, \mathbf{w}_i, r_i = 0)$, il est alors possible de construire un nouvel estimateur de μ par imputation. Celui-ci consiste à imputer les y_i pour les non-répondants, auxquels sont sommés les y_i observés des répondants, avec application des poids de tirage $1/\pi_i$. Cette méthode est la plus usuelle dans la littérature lorsqu'un modèle de Heckman est appliqué.

Avec les mêmes coefficients, le modèle de Heckman permet également de calculer $E(r_i|y_i, \mathbf{z}_i, \mathbf{w}_i)$, qu'on peut également écrire :

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

avec Φ la fonction de répartition de la loi normale centrée réduite. Tous les paramètre sont estimés par le modèle de Heckman.

Il est alors possible d'utiliser cette probabilité conditionnelle de participer, issue du modèle de Heckman, comme modélisation de la participation $\hat{\rho}_i$ pour disposer d'un estimateur par repondération de la forme de μ^1 (équation 1). Cette méthode originale correspond davantage à l'usage fait par les instituts de statistiques dans le cadre de la diffusion des données redressées des enquêtes. De fait, rares sont les enquêtes dont on diffuse les observations des non-répondants pour pouvoir réaliser une estimation corrigée par imputation.

2.2 Hypothèses et limites

Comme on vient de le voir, la repondération par la probabilité de participation modélisée dans le modèle de Heckman est conditionnelle à la variable d'intérêt modélisée. Ainsi, un poids corrigé de la sélection endogène est associé à une seule variable d'intérêt, contrairement au poids issus d'une correction de la non-réponse classique. En pratique, pour éviter la multiplication des poids et des difficultés d'utilisation, on peut essayer de trouver une pondération qui permet de corriger une grande partie du biais de sélection endogène pour plusieurs variables. Par exemple, dans le cadre de l'enquête EpiCov, le poids associé à l'indicatrice d'au moins un symptôme déclaré parmi une liste de symptômes permet de corriger une grande partie du biais de sélection observé pour chaque symptôme séparément ; c'est ce poids qu'il est conseillé d'utiliser pour toute analyse sur les symptômes (Castell *et alii*, à paraître).

La clé de l'identification de la sélection endogène repose sur l'identification de la relation qui existe entre r_i - plus exactement r_i^* - et y_i. L'utilisation du modèle d'Heckman nécessite de faire des hypothèses sur cette relation : du fait de la normalité des aléas, cette relation entre r_i^* et y_i est supposée linéaire. Pour la suite, considérons un instrument binaire. On dispose donc d'un sous-échantillon avec un effort de collecte normal ($w_i = 0$) et un sous-échantillon avec un effort de collecte renforcé conduisant à une participation plus élevée ($w_i = 1$). Le modèle de Heckman permet d'estimer le coefficient directeur de la tangente à la courbe qui relie y à $r^{*\,1}$, au point moyen entre les deux efforts de collecte.

Ainsi, la sélection endogène estimée au voisinage des taux de participation observée sera extrapolée de manière linéaire à l'ensemble des individus échantillonnés, avec un risque de biais par rapport à l'hypothèse de linéarité d'autant plus important que la différence de participation est faible. Par exemple, si on a un taux de réponse de 30 % pour le protocole normal et de 40 % pour le protocole renforcé, on extrapole à l'ensemble des individus échantillonnés la relation entre la variable d'intérêt et la participation par rapport à la relation observée au voisinage de 30-40 %. Le risque de biais lié à l'extrapolation linéaire est plus élevé que si on observait des taux de participation de 30 et 60 % par exemple. Cependant, même avec des taux de participation très différenciés, l'hypothèse de linéarité reste forte puisque rien n'assure que la relation qui relie y et r^* se caractérise par une droite entre les deux taux de participation observés.

Le modèle de Heckman nécessite la présence d'instruments $\mathbf{w}_i$ pour assurer sa bonne convergence. Ces instruments doivent expliquer la participation (équation 5-(ii)) mais pas la variable

1. Plus exactement, il s'agit de la courbe $y(\bar{r}_i^*)$, avec $\bar{r}_i^*$ la propension individuelle à participer à l'enquête, sous l'effort moyen de collecte, conditionnement aux caractéristiques individuelles $\mathbf{z}_i$.

d'intérêt (équation 5-(i)). Ces instruments impliquent que le principe de monotonie soit respecté. De fait, avec un instrument binaire, on a $r_i^*(\mathbf{w}_i = 1) - r_i^*(\mathbf{w}_i = 0) = \psi$ pour tout i. Ainsi, si $\psi \geqslant 0$, alors tout individu participant à l'enquête en l'absence d'instrument ($w_i = 0$) aurait nécessairement participé en présence de l'instrument ($w_i = 1$). Cette propriété de monotonie est valable pour tout individu, et pas seulement en moyenne.

Ce principe de monotonie a des implications sur le choix des instruments. En pratique, le protocole doit permettre de justifier que les individus du groupe dont le taux de réponse est le plus faible et qui ont participé à l'enquête auraient forcément participé s'ils avaient appartenu au groupe dont le taux de réponse est le plus élevé.

Pour répondre à ces différentes conditions, l'idéal est de disposer de sous-échantillons tirés aléatoirement impliquant des niveaux de participation nettement différents. Ainsi, l'indicatrice d'appartenance aux sous-échantillons est corrélée à la participation mais pas aux variables d'intérêt du fait du caractère aléatoire. Par ailleurs, ces sous-échantillons doivent répondre à la condition de monotonicité. On peut ainsi imaginer des protocoles avec un sous-échantillon dont la participation est renforcée par des incitations, qu'il s'agisse d'incitations financières ou d'efforts de relance plus importants.

On peut également envisager d'utiliser la combinaison de différents modes de collecte pour disposer de sous-échantillons aléatoires permettant d'obtenir un taux de participation plus élevé. Pour répondre à la condition de monotonicité, ces sous-échantillons doivent être "emboités", c'est-à-dire que le mode proposé dans le sous-échantillon avec un taux de participation normal doit également être proposé dans le sous-échantillon avec un taux de participation renforcé. De fait, si on dispose par exemple de deux sous-échantillons réalisés sur deux modes de collecte différents (par exemple Internet et téléphone), rien n'assure que les répondants Internet auraient participé par téléphone si on ne leur avait proposé que ce mode de collecte pour participer.

3 La combinaison de protocoles multimodes pour générer des instruments

Les protocoles multimodes apparaissent comme une solution efficace et relativement facile à mettre en œuvre pour corriger la sélection endogène. On a vu par exemple que les enquêtes par Internet peuvent être plus facilement sujettes à ce problème de sélection endogène, du fait d'un risque d'auto-sélection plus important et de taux de réponse relativement faibles. Pour corriger ce biais éventuel, il peut alors s'avérer utile de réaliser un sous-échantillon tiré aléatoirement en utilisant le mode Internet mais également un mode supplémentaire, intermédié par téléphone ou face à face par exemple. Cette solution permet d'envisager d'utiliser de gros échantillons Internet pour réaliser des extensions par exemple, tout en s'assurant de pouvoir identifier et corriger une éventuelle sélection endogène. Cependant, cette application ne peut être mise en œuvre que dans des situations précises, détaillées dans cette partie.

3.1 Une hypothèse forte : l'absence d'erreurs de mesure

L'utilisation comme instrument de protocoles qui diffèrent par la combinaison de plusieurs modes de collecte nécessite de faire une hypothèse forte : l'absence d'erreurs de mesure sur la variable d'intérêt concernée. De fait, s'il existe une erreur de mesure, c'est-à-dire que les individus répondent différemment selon le mode de collecte toutes choses égales par ailleurs, cela signifie que le mode est explicatif de la variable d'intérêt y. On se trouve dans la situation suivante par rapport au modèle présenté ci-dessus :

$$\begin{cases} \text{(i)} & y_i & = & c^1 + \mathbf{z}_i\chi + \sum_{j=1}^{J} \alpha_j \mathbb{1}(m_i = j) + \epsilon_i^1 \\ \text{(ii)} & r_i^* & = & c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \epsilon_i^0 \\ \text{(iii)} & r_i & = & \mathbb{1}(r_i^* \geqslant 0) \end{cases} \tag{7}$$

où m_i désigne le mode avec lequel l'individu i répond à l'enquête.

Ce modèle est identifiable si la variable m_i n'est pas colinéaire aux instruments $\mathbf{w}_i$. L'hypothèse identifiante est donc que le mode de collecte n'explique pas le choix de participer ou non à l'enquête. Or, dans le cas d'une enquête avec des sous-échantillons combinant différents modes de collecte, le fait de participer ou non à l'enquête a de grandes chances d'être lié aux modes de collecte proposés pour participer. La non inclusion du mode m_i dans l'équation de participation rejette le mode dans l'aléa de l'équation de participation, lequel est corrélé avec celui de l'équation de la variable d'intérêt dans le modèle de Heckman. Ainsi, le mode de collecte est corrélé à l'aléa dans l'équation 7-(i), de sorte que les coefficients α_j sur l'effet de mesure sont biaisés. En revanche, si les coefficients de l'équation de participation sont également biaisés du fait de la non inclusion du mode de collecte, l'estimation de la participation, et donc l'identification de la sélection endogène, n'est pas biaisée. A l'inverse, si on n'introduit pas le mode de collecte dans l'équation de la variable d'intérêt alors qu'il existe bien un effet de mesure, alors l'erreur de mesure est renvoyée dans l'aléa, et donc dans l'aléa de l'équation de participation du fait de la corrélation de ces aléas. Or les instruments $\mathbf{w}_i$ étant également liés au mode de collecte, les conditions d'exclusion du modèle de Heckman ne sont plus vérifiées.

Avec ce type de protocole utilisant un sous-échantillon monomode et un sous-échantillon multimode, il n'est donc pas possible d'identifier à la fois une erreur de mesure et une sélection endogène.

3.2 Quels protocoles pour identifier sélection endogène et erreur de mesure ?

D'autres méthodes permettent d'identifier à la fois une erreur de mesure et une sélection endogène. Ainsi, Lee (2009) propose une méthode d'estimation de l'effet d'un traitement en cas de sélection endogène. Contrairement au modèle de Heckman, cette variable de traitement peut expliquer la participation et la variable d'intérêt. Cette méthode, par ailleurs non paramétrique, permet d'estimer une borne de l'effet du traitement en situation de sélection endogène. Cependant, en pratique, cette borne est très large et permet difficilement de conclure (Castell *et alii*, à paraître).

D'autres pistes peuvent être envisagées pour identifier à la fois une erreur de mesure et un biais de sélection endogène. Suivant le modèle de Heckman développé ci-dessus, on peut imaginer un protocole permettant de disposer de plusieurs instruments, l'un permettant d'évaluer l'effet de mesure et l'autre permettant d'évaluer la sélection endogène. Supposons qu'on dispose de trois sous-échantillons tirés aléatoirement :

- un sous-échantillon monomode réalisé sur le mode de collecte A ($Mono_A$) ;
- un sous-échantillon monomode réalisé sur le mode de collecte B ($Mono_B$) ;
- un sous-échantillon multimode réalisé sur les modes de collecte A et B ($Multi$), avec un taux de participation renforcé par rapport aux deux autres sous-échantillons.

Avec un tel protocole, on pourrait modéliser l'effet de mesure et l'effet de sélection endogène en deux temps. Dans un premier temps, il s'agirait d'évaluer l'effet de mesure du mode B par rapport au mode A. Cela est possible avec des méthodes d'inférence causale usuelles dans l'évaluation des effets de mesure, en comparant les répondants aux deux lots monomodes. On peut également envisager d'évaluer cet effet à partir d'un modèle de Heckman, pour prendre en compte la sélection

endogène éventuelle :

$$\begin{cases} \text{(i)} & y_i & = & c^1 + \mathbf{z}_i\chi + \alpha\mathbb{1}(Mono_B) + \epsilon_i^1 \\ \text{(ii)} & r_i^* & = & c^0 + \mathbf{z}_i\beta + \gamma\mathbb{1}(Mono_B) + \psi\mathbb{1}(Multi) + \epsilon_i^0 \\ \text{(iii)} & r_i & = & \mathbb{1}(r_i^* \geqslant 0) \end{cases} \tag{8}$$

Contrairement à la situation précédente, le mode de collecte est introduit dans l'équation de participation (8-(ii)). Ainsi, les coefficients de l'équation de la variable d'intérêt (8-(i)), notamment l'effet de mesure α, sont sans biais. A cette étape, on ne s'intéresse qu'au coefficient mesurant l'erreur de mesure α. A partir de l'estimation de cette erreur de mesure, il est possible de la corriger dans les données, pour l'ensemble des répondants au mode B, qu'ils appartiennent au sous-échantillon $Mono_B$ ou au sous-échantillon $Multi$.

Après correction de l'erreur de mesure, on peut alors revenir au modèle de Heckman précédent (modèle 5), en utilisant comme instrument l'indicatrice d'appartenance au sous-échantillon multimode. La sélection endogène peut alors être identifiée et corrigée sans biais, sous les mêmes hypothèses que celles évoquées dans la partie précédente.

Cette méthode doit être plus précisément étudiée pour en évaluer les conditions d'identification et l'efficacité. Mais elle constitue une piste intéressante pour identifier et corriger à la fois un biais de mesure et un biais de sélection endogène dans les enquêtes multimodes.

4 Conclusion

La problématique de la sélection endogène dans les enquêtes auprès des ménages n'est pas nouvelle mais elle devient plus prégnante avec l'utilisation de nouveaux modes de collecte comme Internet, davantage sujets à cette problématique du fait d'un risque d'auto-sélection plus fort et de taux de réponse qui restent relativement faibles. La mise en place d'un protocole adapté avec plusieurs sous-échantillons aléatoires combinant plusieurs modes de collecte apparaît comme une solution pour identifier et corriger un biais de sélection endogène. Cependant, cette mise en œuvre est soumise à des hypothèses et doit être anticipée en amont de la collecte pour adapter le protocole en conséquence.

Cet article se fonde sur l'utilisation du modèle de Heckman mais on peut très bien envisager d'utiliser d'autres modèles moins paramétriques ou faisant d'autres hypothèses sur la distribution des aléas. Par ailleurs, pour être mis en œuvre en pratique, ce type de protocole nécessite d'être calibré plus précisément. Pour cela, des travaux de simultations sont en cours pour évaluer les conditions, et notamment les tailles de sous-échantillons, dans lesquelles une erreur de mesure et un biais de sélection endogène peuvent être identifiés et corrigés de façon concomitante.

Bibliographie

[1] Beck F., Castell L., Legleye S., Schreiber A., Le multimode dans les enquêtes auprès des ménages : une collecte modernisée, un processus complexifié, Courrier des statistiques, n° 7, Insee, 2022.

[2] Castell L., Favre Martinoz C., Paliod N., Redressements de la première vague de l'enquête EpiCov : un exemple de correction des effets de sélection dans les enquêtes multimodes, Documents de travail, Insee, à paraître.

[3] Castell L., Sillard P., Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman, Documents de travail, n° M2021-02, Insee, 2021.

[4] Groves R., Singer E., Corning A., Leverage-saliency theory of survey participation : Description and an illustration, Public Opinion Quaterly, vol 64, n° 3, pp 299-308, 2000.

[5] Heckman J.J., Sample selection bias as a specification error, Econometrica, vol 45, n° 1, pp 153-161, 1979.

[6] Lee D.S., Training, wages, and sample selection : Estimating sharp bounds on treatment effects, The Review of Economic Studies, vol 76, n° 3, pp 1071-1102, 2009.

[7] Tchetgen Tchetgen E.J., Wirth K.E., A general instrumental variable framework for regression analysis with outcome missing not at random, Biometrics, vol 73, n° 4, pp 1123-1131.