Éléments sur les mécanismes de sélection dans les enquêtes et sur la non-réponse non-ignorable

Eric GAUTIER

INSEE, Unité Méthodes Statistiques

Introduction

Les mécanismes de sélection sont présents à plusieurs niveaux lors de la collecte d'une enquête. Tout d'abord la théorie des sondages (approche basée sur le plan de sondage) a pour objectif d'inférer des grandeurs sur la population totale finie au vu d'un échantillon tiré au hasard. Nous parlerons de sélection de phase 1. Dans un second temps vient la non-réponse totale. Celle-ci est multiforme et regroupe notamment les impossibles à joindre, les refus, les inaptes et les cas où trop de non-réponse partielle rend le questionnaire inexploitable. Nous qualifierons cette non-réponse de sélection de phase 2. Enfin, parmi les questionnaires collectés restant, une partie de l'information est manquante, il s'agit de la non-réponse partielle que nous appellerons aussi sélection de phase 3. Les données recueillies peuvent alors être utilisées, sous une approche de théorie des sondage « basée sur le plan de sondage » pour mener des estimations de quantités sur la population totale, mais aussi de paramètres de lois de comportement en utilisant des techniques économétriques.

Comme il est usuel lorsque l'on s'intéresse à des problèmes de données manquantes dans les enquêtes, on peut ou bien supposer que la sélection uniquement est aléatoire ou bien que la sélection et les variables que l'on mesure sont aléatoires. Ainsi quoiqu'il arrive la sélection, aux différentes phases, est aléatoire et caractérisée par une variable S qui vaut 1 si il y a sélection et 0 sinon. Pour des questions pratiques nous supposerons que la sélection est la traduction du fait que la propension à répondre S^*, variable aléatoire réelle latente, dépasse un certain seuil (fixé en pratique à 0, ce n'est pas une restriction). Remarquons aussi que dans le cas de la sélection de phase 3, il s'agit d'une sélection pour chaque « item », c'est à dire pour chaque variable de l'enquête[1]. Nous notons dans ce qui suit $\left(S^i\right)_{i=1,2,3}$ les variables aléatoires correspondant aux mécanismes de sélection de chacune des phases et $\left(S^{*i}\right)_{i=1,2,3}$ les variables latentes associées. Remarquons que nous observons des réalisations de la loi de S^2 conditionnelle à $S^1=1$ et de S^3 sachant $S^1=1$ et $S^2=1$.

Nous adoptons dans le présent document une approche plutôt modèle où la sélection et les variables sont supposées être aléatoires. En quelque sorte les données sont par nature aléatoire, l'observation en population totale de taille N correspondant à l'observation d'un échantillon

[1] On dit souvent que c'est pour cela que l'on préfère l'imputation à la repondération pour traiter la non-réponse partielle. Il y aurait sinon autant de systèmes de poids que de variables et il serait difficile d'étudier des corrélations. Pour une sélection NMAR, voir plus loin, il est impératif d'étudier simultanément sélection et comportement et nous avons aussi autant de modèles que de sélections. Ceux-ci devant être des modèles joints.

indépendant et identiquement distribué (iid) dans une certaine loi. On peut aussi supposer que le tirage de l'échantillon à ce stade (penser à un tirage fait par Dieu) est « uniforme » et que la sélection, sélection de phase 0, ne dépend pas des valeurs prises par les données. Selon l'approche modèle en théorie des sondages l'objectif est alors d'inférer des grandeurs sur les lois des variables aléatoires dont on dispose un échantillon[2]. Le nombre N étant un grand nombre on peut espérer que les grandeurs en population totale soient proches des grandeurs que l'on obtiendrait connaissant la vraie loi. Sous cette approche on utilise le même langage pour faire des inférences de sondage et des inférences économétriques.

Alors nous pouvons définir la typologie suivante des mécanismes de sélection. La terminologie initiale parle de mécanisme de données manquantes et nous parlerons ici plutôt de mécanismes de sélection afin de traiter de la même façon la phase d'échantillonnage. Un mécanisme de sélection est *Missing at Random* (MAR) si quitte à conditionner par suffisamment de variables on obtienne une loi de comportement en population générale égale à la loi de comportement conditionnelle à la sélection. D'après la formule de Bayes cela revient aussi à supposer que de la loi du mécanisme de sélection (ou de manière plus restrictive de la propension à répondre) conditionnelle à l'observation de co-variables est indépendante du comportement. Le cas où la loi de la sélection est indépendante de toutes les variables du questionnaire correspond au cas des mécanismes *Missing Completely At Random* (MCAR), on dit aussi que la sélection est uniforme. Cette deuxième hypothèse est souvent peu réaliste. Enfin parfois la sélection n'est pas MAR, on l'appelle alors NMAR (*Not Missing At Random*). Dans la terminologie initiale propre à D. Rubin, la notion de sélection ignorable est légèrement plus restrictive que celle de sélection MAR et vaut pour des modèles paramétrés de sélection et de comportement lorsque les paramètres du modèle de sélection et de comportement vivent dans un espace produit. Beaucoup d'auteurs ne font pas la distinction, nous ne la ferons pas non plus. Finalement une sélection est *ignorable* si on peut **l'ignorer**, **quitte à conditionner** par suffisamment de variables, lorsque l'on infère sur des quantités en population totale ou des comportements en population générale. Remarquons que dans les techniques de repondération utilisées par exemple pour le traitement de la non-réponse partielle on n'ignore pas le mécanisme de non-réponse, bien au contraire celui-ci est central dans la construction des estimateurs, par contre on ignore la loi de comportement que l'on ne désire pas modéliser. Mais ceci est parfaitement équivalent d'après la formule de Bayes. Finalement nous avons aussi que si la sélection est non-ignorable et que l'on omet de modéliser la non-réponse comme dépendant de la variable d'intérêt en partie inobservable, les probabilités de réponse individuelles et donc les estimateurs par repondération seront biaisés. Le cas d'une sélection non-ignorable est le plus général mais on s'attend à ce qu'il soit tout de même assez répandu. On peut en effet penser que la non-réponse à une question sensible portant sur le revenu, le patrimoine ou des pratiques sexuelles soit non-ignorable. Il peut s'agir d'une sélection de phase 2, l'enquêté ayant reçu une lettre avis connaît le sujet de l'enquête et n'ouvre pas ou est délibérément absent au moment de l'enquête. Il peut aussi s'agir d'une sélection de phase 3, l'enquêté est de bonne volonté mais décide malgré tout de ne pas répondre aux questions sensibles. Cette typologie n'est pas anecdotique, elle est même tout à fait fondamentale pour comprendre les éventuels biais de sélection des inférences obtenues à partir des éléments sélectionnés. En effet en présence de non-réponse et de sélection le problème de biais est le plus préoccupant.

Le papier présente quelques aspects et méthodes existantes et ne se veut pas exhaustif. Il est organisé de la façon suivante. Dans une première partie nous revenons sur la notion de sélection MAR et rappelons notamment que des biais peuvent subsister si on omet certaines variables qui permettraient pourtant par conditionnement d'obtenir de l'indépendance entre le comportement et la sélection. Dans une seconde partie nous présentons des approches paramétriques pour une sélection non-ignorable basées ou bien sur des modèles de mélange ou bien sur des modèles de sélection de type modèle Tobit généralisé. Enfin la dernière partie est une petite introduction aux options non-paramétriques qui s'offrent à nous.

[2] Le modèle sert à prédire les données non sélectionnées. Il existe deux variantes : les modèles de superpopulation et la modélisation bayésienne, c.f. par exemple la référence [9]. Lorsqu'il y a de la sélection et que nous ne disposons pas de variables pour l'appréhender, voir plus loin, les inférences sont basées sur une modélisation jointe des variables et de la sélection. Enfin, en pratique nous disposons de peu de co-variables pour prédire les données non sélectionnées.

1. La sélection MAR

1.1. Définition

Ecrivons plus précisément la définition d'un mécanisme de sélection MAR. Nous supposons ici et dans ce qui suit que les enregistrements sont indépendants d'une unité (ménage, individu) à l'autre. Nous disons que la sélection S qui opère sur une variable Y est MAR lorsque

$$\forall i = 1,...,n, P\left(S_i = 1 \middle| Y_i, X_i\right) = P\left(S_i = 1 \middle| X_i\right),$$

où X_i est le vecteur aléatoire correspondant aux variables disponibles renseignées pour l'unité i, ce sont des variables de l'enquête, de la base de sondage ou d'une base de suivi de collecte , et n est la taille de notre tableau « à trous », $\left(S_i, Y_i, X_i\right)_{i=1}^{n}$ sont supposés être iid.

Supposons désormais pour simplifier qu'il n'y a qu'une seule variable sujette à sélection, il n'est plus alors nécessaire de faire figurer les indices i. Nous obtenons que, de manière équivalente, l'hypothèse sur le mécanisme de sélection est aussi une hypothèse sur le comportement et nous obtenons

$$P\left(Y \in A \middle| S = 1, X = x\right) = P\left(Y \in A \middle| S = 0, X = x\right) = P\left(Y \middle| X = x\right),$$

où x correspond à une observation de la variable X, A est un évènement. Ainsi si on étudie un comportement conditionnel à toutes les variables de l'enquête on peut ignorer la sélection.
Mais alors ne conditionner que par un sous ensemble de X peut parfois faire que l'identité ci-dessus n'est plus vraie même si la sélection est ignorable.

Un cas particulier de sélection MAR est la sélection MCAR pour laquelle nous avons de manière équivalente

$$P\left(S = 1 \middle| Y, X\right) = cste \text{ et } P\left((Y, X) \in A \middle| S = 1\right) = P\left((Y, X) \in A \middle| S = 0, X = x\right) = P\left((Y, X) \in A\right),$$

Cette seconde hypothèse est très restrictive et en pratique on pose l'hypothèse de sélection MCAR par classe, les classes étant définies à partir des co-variables. Elle est testable par exemple via l'estimation d'un modèle logistique et la lecture du test de nullité des coefficients. R.J.A. Little propose un autre test de l'hypothèse MCAR dans [11]. Lorsque la sélection est MCAR travailler pour l'inférence sur Y sur les observations[3] sans procéder à de l'imputation n'introduit pas de biais de sélection. Si par contre la sélection n'est pas MCAR, imputer Y au vu des co-variables X peut permettre d'obtenir des estimations non biaisées tout en ignorant la loi du mécanisme de sélection. Mais aussi la loi de Y conditionnelle à X, qui peut être à estimer si on souhaite imputer, est non seulement celle des éléments sélectionnés mais aussi celle en population général, c'est-à-dire non conditionnelle à la sélection.

1.2. Un exemple où exclure une co-variable entraîne des biais pour une inférence sur une loi en population générale

Dans l'exemple qui suit l'objectif est de faire une estimation de paramètres de comportement à partir d'observations d'une variable sujette à sélection et de co-variables. Supposons que nous ne disposions que d'au plus trois variables Y, X_1 et X_2 et que les deux dernières variables soient dichotomiques (pour simplifier) et parfaitement observée alors que la première variable est sujette à sélection.

[3] En anglais la terminologie associée est « available case analysis ».

Nous supposons que la loi de Y sachant $X_1 = x_1$ et $X_2 = x_2$ est de moyenne $a + bx_1 + cx_2$. Dans ce cas nous pouvons voir que la loi de Y sachant $X_1 = x_1$ est quant à elle de moyenne

$$a + cE\left[X_2 \middle| X_1 = 0\right] + \left(b + cE\left[X_2 \middle| X_1 = 1\right] - cE\left[X_2 \middle| X_1 = 0\right]\right)x_1.$$

Supposons en outre que la sélection soit telle que

$$P\left(S = 1 \middle| X_2 = 1\right) = p\,,\ P\left(S = 0 \middle| X_2 = 1\right) = 1 - p$$
$$P\left(S = 1 \middle| X_2 = 0\right) = q \text{ et } P\left(S = 0 \middle| X_2 = 1\right) = 1 - q\,.$$

En quelque sorte la sélection sur-représente des modalités d'une variable X_2 si $p \neq q$.

Supposons que l'on souhaite inférer sur la loi de Y sachant $X_1 = x_1$ qui existe car la loi de Y sachant $X_1 = x_1$ et $X_2 = x_2$ existe. Une inférence sur la moyenne par exemple par MCO sur les données issues de la phase de sélection fournit un estimateur convergent de

$$E\left[Y \middle| X_1 = x_1, S = 1\right] = E\left[E\left[Y \middle| X_1 = x_1, X_2\right] S = 1\right]$$
$$= a + cE\left[X_2 \middle| X_1 = 0, S = 1\right] + \left(b + cE\left[X_2 \middle| X_1 = 1, S = 1\right] - cE\left[X_2 \middle| X_1 = 0, S = 1\right]\right)x_1.$$

Cette nouvelle moyenne est a priori différente de la moyenne de la loi en population générale si les lois de S sachant X_1 et de X_2 sachant X_1 ne sont pas indépendantes.

L'exemple qui précède n'est pas du tout anodin. Ce type de sur-représentation peut avoir lieu dès la phase de l'échantillonnage. Ainsi faire une inférence modèle ne dispense pas d'aller voir le plan de sondage au moment du tirage. Enfin, il peut aussi s'agir de sélection de phase 2 ou 3. Si elle est de phase 3 la variable dont dépend la sélection peut être une variable disponible dans l'enquête et en ne la prenant pas en compte on infère sur le comportement des gens qui sont sélectionnés et non sur le comportement en population générale. Mais aussi, cela peut être une variable inobservée, par exemple au stade de la non-réponse totale, étant corrélée avec la variable d'intérêt ; nous changeons alors de registre de difficulté et il s'agit cette fois-ci de sélection endogène ou de non-réponse non-ignorable. En tout cas, au stade de la sélection de phase 1, la sélection est connue, les variables sont à notre disposition et parfaitement observées et le problème beaucoup plus simple, il est donc facile dans ce cas de ne pas mener des estimations biaisées. Ce problème est central dans toutes les techniques d'importation d'une équation d'une enquête à une autre, pour détecter des points aberrants ou pour faire de l'imputation. Outre le problème de date, subsiste le problème de sélection qui sans forcément être non-ignorable peut être MAR. La phase 1 de sélection est connue et il convient au moins de vérifier que dans les deux enquêtes la même sur-représentation a opéré ou que l'équation prend en compte une variable qui permet d'ignorer cette sur-représentation. Enfin, comme en pratique subsiste de la sélection de phase 2 et 3 cet exercice devient encore plus délicat.

On pourrait se demander si dans ce cas utiliser des poids de sondage permettrait de corriger le biais issus d'une spécification inadaptée. La réponse dans le cas d'un modèle linéaire Gaussien est négative. Dans une approche modèle, prendre en compter des poids revient à faire figurer des poids dans la vraisemblance et finalement est analogue à poser un modèle avec hétéroscédasticité. Hors dans ce cas l'estimateur des MCOs reste « convergent ». Nous mettons « convergent » entre guillemets car l'estimateur du paramètre apparaissant dans la moyenne converge certes mais vers celui de la loi dont on dispose un échantillon c'est-à-dire de la loi conditionnelle à la sélection et non de la loi en population générale. Enfin subsiste une dernière difficulté que nous ne prenons pas en compte qui est que si l'on admet que les réalisations de (S, Y, X) sont indépendantes rien ne garantit que celles de (Y, X) conditionnelles à $S = 1$ le soient encore.

1.3. Le cas d'un estimateur du total en population finie

Soit Q_N le produit des lois de (Y_i, X_i, S_i^1). Alors l'espérance conditionnelle aux observations de Y et X de l'estimateur de Horvitz-Thompson du total est donnée par

$$E_{Q_N}\left[\sum_{i=1,...,N}\frac{Y_i S_i^1}{\pi_i}\Big|\forall i\in\{1,...,n\}, Y_i = y_i; \forall i\in\{1,...,m\}, X_i = x_i\right]$$
$$= E_{Q_N}\left[E_{Q_N}\left[\sum_{i=1,...,N}\frac{Y_i S_i^1}{\pi_i}\Big|Y_i, X_i, i=1,...,N\right]\Big|\forall i\in\{1,...,m\}, Y_i = y_i; \forall i\in\{1,...,m\}, X_i = x_i\right]$$
$$= E_{Q_N}\left[\sum_{i=1,...,N}\frac{Y_i P(S_i^1 = 1|Y_j, X_j, j=1...N)}{\pi_i}\Big|\forall i\in\{1,...,n\}, Y_i = y_i; \forall i\in\{1,...,m\}, X_i = x_i\right],$$

Si les variables X sont des variables de l'enquête alors $m = n$, si ce sont des variables de la base de sondage alors m est la taille de la base de sondage. On peut supposer par indépendance que

$$P(S_i^1 = 1|Y_j, Z_j, j = 1,..., N) = P(S_i^1 = 0|Y_i, Z_i).$$

L'estimateur de Horvitz-Thompson est donc sans biais lorsque

$$\forall i\in\{1,...,n\}, \pi_i = P(S_i^1 = 1|Y_i = y_i, X_i = x_i), \text{ et}$$
$$\forall i\notin\{n+1,...,m\}, \pi_i = E\left[P(S_i^1 = 1|Y_i, X_i = x_i)\right].$$

Le cas où la sélection de phase 1 suit un mécanisme MAR est celui où $P(S_i = 1|Y_i, X_i) = P(S_i = 1|X_i)$. Omettre certaines variables du vecteur X peut rendre la relation précédente fausse et rendre l'estimateur de Horvitz-Thompson biaisé, on rencontre ce type de problème lorsque l'on effectue de la repondération pour traiter la non-réponse totale et que l'on utilise un modèle Logit de non-réponse. Le cas où la sélection est non-ignorable et où la seule phase de sélection est la sélection de phase 1 est celui des plans de sondage informatifs où conditionnellement aux observations du vecteur X, il existe y_1 différent de y_2 et x tel que

$$P(S = 1|Y = y_1, X = x) \neq P(S = 1|Y = y_2, X = x).$$

2. La sélection non-ignorable, approche paramétrique

Il est souvent admis que pour beaucoup de questions la sélection de phase 3 puisse être ignorable. En effet, nous disposons dans l'enquête de nombreuses variables permettant par conditionnement d'obtenir de l'indépendance entre le comportement et la sélection. Par contre, lorsque l'enquêté est informé du sujet de l'enquête par lettre avis, il est tout à fait possible que la sélection de phase 2 soit « endogène » et que l'on ne dispose pas de variable (celles du plan de sondage ou de variables d'une base de suivi de collecte que l'on s'est constitué) pour pouvoir ignorer la sélection. Mais, même au niveau de la non-réponse partielle, il peut subsister de la sélection non-ignorable notamment pour des questions relatives aux revenus, au patrimoine ou aux pratiques sexuelles. Enfin, pour toutes les enquêtes usuelles, la sélection de phase 1 est ignorable. Le cas d'une sélection de phase 1 non-ignorable est très particulier et nous parlons alors de plans de sondage informatifs, cela peut être le cas dans les enquêtes par quotas.

2.1. Mélanges et identifiabilité

Une première approche au problème de sélection non-ignorable consiste à spécifier des comportements différents chez le groupe des sélectionnés et chez le groupe des non-sélectionnés.

Par exemple, sans mobiliser d'information sur des co-variables, nous pourrions poser un modèle de mélange de lois normales c'est-à-dire que la loi $L(Y|S=1)$ est une loi normale de moyenne μ_1 et de variance σ_1^2 et que la loi $L(Y|S=0)$ est une loi normale de moyenne μ_0 et de variance σ_0^2. L'hypothèse MCAR revient à supposer que $\mu_0 = \mu_1$ et que $\sigma_0^2 = \sigma_1^2$. Sans hypothèse sur la sélection et en supposant que la loi mélangeante est une loi de Bernoulli de paramètre p, nous obtenons que la loi mélangée, c'est-à-dire la loi en population générale ou déconditionnée de la sélection est de moyenne $p\mu_0 + (1-p)\mu_1$ et de variance $p\sigma_1^2 + (1-p)\sigma_2^2 + p(1-p)(\mu_1 - \mu_0)^2$ (en utilisant l'expression de la variance en fonction de la variance et de l'espérance conditionnelle et que la variance d'une variable aléatoire est invariante par translation). La loi mélangée n'est plus une loi normale, c'est un mélange de lois normales.

Nous pouvons constater ici qu'il n'est pas possible d'estimer les paramètres de la composante non sélectionnée du mélange et nous sommes confronté à un problème d'identification. Une façon de rendre le modèle identifiable est d'imposer des restrictions sur les paramètres. La restriction $\mu_0 = \mu_1$ et $\sigma_0^2 = \sigma_1^2$ revient par exemple à faire l'hypothèse que la sélection est MCAR. Nous allons voir qu'il existe des restrictions « moins restrictives ».

Supposons qu'à présent nous considérions une variable Y sujette à sélection et une variable X parfaitement observée et que nous considérions un mélange de deux lois normales bivariées de loi mélangeante une loi de Bernoulli de paramètre p. La loi de (Y, X) conditionnelle à la sélection est de moyenne $\begin{pmatrix} \mu_{1Y} \\ \mu_{2X} \end{pmatrix}$ et de matrice de covariance $\begin{pmatrix} \sigma_{1YY}^2 & \sigma_{1XY}^2 \\ \sigma_{1XY}^2 & \sigma_{1XX}^2 \end{pmatrix}$ et la loi conditionnelle à la non sélection de moyenne $\begin{pmatrix} \mu_{0Y} \\ \mu_{0X} \end{pmatrix}$ et de matrice de covariance $\begin{pmatrix} \sigma_{0YY}^2 & \sigma_{0XY}^2 \\ \sigma_{0XY}^2 & \sigma_{0XX}^2 \end{pmatrix}$. On peut identifier 8 des 11 paramètres du modèle, il s'agit des paramètres $p, \mu_{1X}, \mu_{0X}, \mu_{1Y}, \sigma_{1YY}, \sigma_{1XX}, \sigma_{1XY}$ et σ_{0XX}. Imposer une restriction afin de rendre le modèle identifiable peut être obtenu en faisant une hypothèse sur la loi de Y conditionnelle à $X = x$ et à la non-sélection. Si on suppose qu'elle est égale à la même loi conditionnelle à la sélection, ce qui correspond à un cas particulier de l'hypothèse MAR sur le mécanisme de sélection, le modèle de mélange est désormais identifiable. Little et Rubin proposent aussi dans [10] de postuler que la loi de X sachant Y est cette fois indépendante de la sélection. Ceci permet encore de rendre le modèle de mélange identifiable. Ils appellent cela une hypothèse protectrice et font remarquer qu'imputer dans la prédiction que fournit une estimation de la loi mélangée, sous la contrainte, revient à imputer par la régression inverse de X sur Y comme dans les problèmes de calibration. Enfin, ils proposent des solutions intermédiaires où ils supposent que la sélection ne dépend plus ou de X ou de Y mais plutôt d'une combinaison linéaire de X et Y de la forme $X + \lambda Y$. La loi du mélange est alors identifiable pour une valeur fixée du paramètre λ. Le paramètre λ n'est lui pas identifiable et ils proposent éventuellement d'effectuer plusieurs imputations pour plusieurs valeurs de λ afin de mesurer la sensibilité des inférences au paramètre et proposent également de faire des inférences bayésiennes en posant une loi a priori (par exemple non informative) sur le paramètre λ.

2.2. Un modèle de sélection, le modèle Tobit généralisé

Le modèle qui suit est aussi appelé modèle de sélection d'Heckman. Nous considérons le cas où une seule variable est partiellement manquante et faisons comme si la sélection au niveau de l'échantillonnage et celle de deuxième phase étaient parfaitement uniformes (i.e. MCAR).

2.2.1. Le modèle et une première approche de l'estimation

Posons, le modèle suivant pour la loi jointe du comportement et de la variable latente S^{*3} du mécanisme de sélection de phase 3. On observe des réalisations des vecteurs de co-variables X (qui contiennent éventuellement des variables du plan de sondage) pour tous les enregistrements et de la variable Y seulement lorsque S^{*3} est positive. On n'observe pas la variable latente mais on observe son « signe » S^3. Le système s'écrit :

$$\begin{cases} Y &= X\beta_1 + \varepsilon_1, \\ S^{*3} &= X\beta_2 + \varepsilon_2, \end{cases}$$

et Y est observé si $S^{*3} \geq 0$, Y est non-répondu sinon. Nous supposons, dans un premier temps, que la loi du vecteur des résidus est la loi d'un vecteur Gaussien, c'est à dire :

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \text{ suit la loi } N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right].$$

Le 1 dans la matrice de covariance figure afin que le modèle soit identifiable. L'emploi de la loi Gaussienne se justifie au moins pour des raisons pratiques. En effet, il est simple de conditionner dans un vecteur Gaussien car il existe des formules fermées connues.

Le mécanisme de sélection est bien non-ignorable
Par conditionnement dans un vecteur Gaussien nous pouvons écrire :

$$P\left(S^{*3} < 0 \middle| Y = y, X = x\right) = P\left(\varepsilon_2 < -x\beta_2 \middle| \varepsilon_1 = \frac{y - x\beta_1}{\sigma}, X = x\right)$$

$$= 1 - \Phi\left(\frac{x\beta_2 + \rho\sigma^{-1}\left(y - x\beta_1\right)}{\sqrt{1-\rho^2}}\right),$$

où Φ désigne la fonction de répartition de la loi normale centrée réduite. Si $\rho \neq 0$, nous sommes donc bien en présence d'un mécanisme de sélection non-ignorable. Remarquons aussi que cette probabilité est décroissante avec y lorsque $\rho > 0$ et croissante sinon.

La loi de comportement conditionnelle à la sélection
Si le modèle est le vrai modèle, les observations sont dans la loi $L(Y \middle| X = x, S^{*3} \geq 0)$ qui est une loi de moyenne $x\beta_1 + \rho\sigma\lambda\left(x\beta_2\right)$ et de variance

$$\sigma^2 + \rho^2\sigma^2\left(x\beta_2\lambda\left(x\beta_2\right) - \left(\lambda\left(x\beta_2\right)\right)^2\right),$$

où $\lambda = \frac{\varphi}{\Phi}$ est appelé l'inverse du ratio de Mills avec φ densité de la loi normale centrée réduite.
Cette loi correspond à celle de la loi marginale de ε_1 si et seulement si $\rho\sigma = 0$, c'est le cas s'il y indépendance des termes d'erreur (ou indépendance des lois de Y sachant $X = x$ et de S^*

sachant $X = x$) ce qui se traduit ici par une non-corrélation. Elle n'est plus nécessairement Gaussienne, alors que la loi $L(Y|X = x, S^{*3} = s)$ l'est.

La loi de comportement conditionnelle à la non-sélection

Les valeurs non observées sont alors dans la loi $L(Y|X = x, S^{*3} < 0)$ qui est de moyenne $x\beta_1 - \rho\sigma\lambda(-x\beta_2)$ et de variance

$$\sigma^2 + \rho^2\sigma^2\left(x\beta_2\lambda(-x\beta_2) - (\lambda(-x\beta_2))^2\right).$$

L'imputation aléatoire correspondrait alors à un tirage dans cette loi et l'imputation déterministe à la moyenne de cette loi.

Estimation en deux étapes

Il s'agit d'une méthode d'estimation facile à mettre en œuvre sous SAS. Elle consiste à estimer dans un premier temps le paramètre β_2 de la seconde équation par le modèle Probit (par exemple avec la PROC LOGISTIC) puis à estimer par moindres carrés ordinaires la régression de Y sur x et $\frac{\varphi}{\Phi}\left(x\hat{\beta}_2\right)$. Pourtant comme nous l'avons vu plus haut le résidu de la loi conditionnelle à la sélection est hétéroscédastique par nature. En outre les observations des nouveaux résidus obtenus en remplaçant dans l'inverse du ratio de Mills le paramètre β_2 par un estimateur sont corrélées. L'estimateur des moindres carrés ordinaires ne correspond donc plus à celui du maximum de vraisemblance, néanmoins on sait qu'il sera malgré tout convergent et asymptotiquement normal. Il existe une formule pour la matrice de covariance asymptotique de l'estimateur des paramètres par cette procédure, la formule usuelle issue des MCO ne s'applique plus car β_2 est connu de manière imparfaite et approché par $\hat{\beta}_2$. Un autre moyen de calculer la matrice de covariance est d'utiliser les techniques de bootstrap. Cela permet notamment d'effectuer des tests par exemple de nullité des coefficients de la moyenne de la loi conditionnelle à la sélection, ceux fournis par les MCOs en deuxième étape étant faux hormis pour le coefficient de l'inverse du ratio de Mills estimé. Enfin nous pouvons conclure que la procédure en deux étapes donne des estimateurs convergents même si l'hypothèse de binormalité est légèrement affaiblie, voir par exemple [1] et [13].

Un test de sélection MAR via l'estimation en deux étapes

Le test découle directement de l'estimation du modèle, en effet la statistique de Student du paramètre $\rho\sigma$ est correctement calculée sous l'hypothèse nulle $\rho = 0$, de sélection ignorable. Si le test est rejeté la sélection est non-ignorable.

L'imputation

Si la sélection est non-ignorable nous disposons d'estimateurs convergents des paramètres d'intérêt de la loi conditionnelle à la non-sélection hormis un estimateur du paramètre σ [4]. Nous pouvons donc mener une imputation déterministe en imputant par la moyenne conditionnelle de la loi où le paramètre est le paramètre estimé. Cette imputation est adaptée si, a posteriori, on souhaite estimer le coefficient d'un modèle de régression où la variable figure comme variable explicative, ou de la droite si cette variable est la variable expliquée ou si l'on cherche à estimer une moyenne ou plus généralement toute statistique linéaire en les observations. Par contre, les estimateurs sur données imputées de statistiques non linéaires en les observations seront biaisés. Dans ce cas il est préférable d'avoir recours à une imputation par simulation. Il faut pour cela une méthode d'estimation qui donne accès à tous les paramètres. Il est ensuite possible de mener une simulation utilisant un algorithme d'acceptation/rejet[5]. Il est par exemple possible de simuler

[4] On peut trouver dans la littérature économétrique une troisième étape qui permet d'obtenir un estimateur de σ.

[5] Mais il est à noter que l'acceptation/rejet est de moins en moins efficace lorsque la dimension augmente.

plusieurs fois le vecteur (Y, S^{3*}) et d'accepter la première valeur de Y pour laquelle $S^{3*} < 0$. Il semble difficile de simuler directement la loi conditionnelle à la non sélection car nous n'avons précisé que sa moyenne et sa variance et il est connu que la marginale d'une loi normale tronquée de dimension supérieure à 1 n'est pas en général une loi normale tronquée[6].

Estimation par maximisation de la vraisemblance
On peut estimer le modèle précédent par maximum de vraisemblance. L'estimateur sera convergent si les hypothèses constitutives du modèle (ici modèle linéaire Gaussien) sont satisfaites. La contribution à la vraisemblance d'un individu *i* sélectionné vaut:

$$L_i = P(S^{*3} \geq 0 | X = x_i, Y = y_i) \frac{1}{\sigma} \varphi\left(\frac{y_i - x_i \beta_1}{\sigma} \right)$$
$$= \Phi\left[\frac{1}{\sqrt{1-\rho^2}} \left(x_i \beta_2 + \frac{\rho}{\sigma}(y_i - x_i \beta_1) \right) \right] \frac{1}{\sigma} \varphi\left(\frac{y_i - x_i \beta_1}{\sigma} \right),$$

et

$$L_i = P(S^{*3} < 0 | X = x_i) = \int_R \Phi\left(- \frac{x_i \beta_2 + \rho u}{\sqrt{1-\rho^2}} \right) \varphi(u) du,$$

pour un non-sélectionné. Ce type de vraisemblance s'appelle aussi vraisemblance observée, il s'agit de l'intégrale de la vraisemblance si toutes les données étaient observées par rapport aux données manquantes. La maximisation de la vraisemblance peut être effectuée en SAS IML.

Pourquoi cela marche t'il ? Pourquoi cela ne marche-t-il pas ?
Le modèle est basé sur l'hypothèse non testable qu'en population générale le comportement est linéaire Gaussien. La loi conditionnelle à la sélection devient asymétrique du fait de la sélection à moins que ρ soit nul. A partir de l'échantillon des répondants ρ s'estime sur le décentrage et donc l'écart au modèle linéaire Gaussien. Une imputation menée comme au dessus complètera l'échantillon afin de le rendre linéaire Gaussien. Si l'hypothèse de départ que la loi en population générale est celle d'un modèle linéaire Gaussien n'est pas vérifiée, la correction du biais de sélection pourrait éventuellement ajouter du biais plutôt que d'en enlever !

Comment spécifier d'autres lois ?
Dans la référence [1], l'auteur présente une généralisation du modèle à d'autres lois, tout en permettant l'emploi d'une méthode en deux étapes. En particulier on pourrait choisir toute sorte de loi pour le deuxième résidu et imposer que la loi du premier conditionnelle au second soit linéaire en le second. Si on consent à utiliser des techniques de maximisation de la vraisemblance, on peut alors penser spécifier toute sorte de lois. Mais aussi on peut penser utiliser des fonctions copules si on souhaite associer des marginales que l'on connaît bien et mener des estimations paramétriques ou non-paramétriques de la fonction de dépendance.

Un test de spécification
Le fait de disposer de deux estimateurs des paramètres de la première équation (le maximum de vraisemblance et l'estimation en deux étapes) dont l'un est convergent même en relâchant l'hypothèse sur la loi du résidu alors que l'autre ne l'est pas, permet de mettre en œuvre un test de spécification « à la Haussman » portant sur la binormalité des termes d'erreurs. Pour plus de détail le lecteur peut consulter la référence [13].

Utiliser des co-variables différentes dans les modèles de comportement et de sélection
Il est aussi possible, nous ne l'avons pas fait afin de ne pas surcharger les notations, de prendre des co-variables différentes. Seulement lorsque l'on écrit la loi de la variable d'intérêt conditionnelle à la sélection ou à la non-sélection, nous conditionnons aussi par les observations

[6] Sans la troncature par contre il est vrai, et nous l'avons déjà utilisé, que conditionner dans un vecteur Gaussien donne bien toujours des vecteurs Gaussiens.

des co-variables communes aux deux équations. Il convient donc dans ce cas de prendre pour co-variables de la variable d'intérêt au moins toutes celles de la sélection, à moins qu'il y ait indépendance conditionnelle. Une estimation par MCO de la loi marginale conditionnelle aux co-variables de la sélection permet de se faire une première opinion sur l'opportunité ou non de faire figurer certaines variables de la sélection dans le modèle de la variable d'intérêt. Si on effectue une estimation par maximum de vraisemblance toutes les combinaisons sont possibles.

2.2.2. Estimation par des algorithmes EM

L'algorithme EM, développé initialement par A.P. Dempster, N.M. Laird et D. Rubin en 1977, le lecteur peut se référer aux références [10] et [14] pour plus de détails, est un algorithme itératif du calcul du maximum de vraisemblance. Il est particulièrement adapté aux problèmes à données manquantes car alors la vraisemblance, par rapport à la vraisemblance sans données manquantes, est intégrée. Il est basé sur le résultat central que le paramètre qui maximise la log-vraisemblance prédite (i.e. celle obtenue en intégrant la log-vraisemblance par rapport à la loi des données manquantes paramétrée par la valeur courante du paramètre de l'étape t) augmente du même coup de logarithme de la vraisemblance observée[7]. L'algorithme est donc une succession d'une étape (E), étape d'espérance, où on calcule l'espérance de la log-vraisemblance pour la valeur courante du paramètre puis d'une actualisation du paramètre en celui qui maximise cette nouvelle fonction du paramètre, étape (M). L'algorithme converge sous des hypothèses de régularité vers un point stationnaire, qui pourrait malheureusement être un point selle ou un extremum local. Mais c'est aussi le problème des algorithmes numériques déterministes de maximisation de la vraisemblance par exemple l'algorithme de Newton-Raphson. Des version stochastiques peuvent permettre d'outre passer ce problème comme par exemple l'algorithme SEM ou l'étape E est remplacée par une simulation dans la loi avec le paramètre estimé dans l'étape courante, mais pour ces méthodes nous n'avons pas nécessairement de résultats de convergence. En pratique, il convient de partir de plusieurs initialisations différentes des paramètres et de comparer les limites que l'on obtient. En outre, la convergence peut être lente, elle l'est d'autant plus que le taux de valeurs manquantes est élevé. Il prend par contre une forme très agréable lorsque les données sont issues d'un échantillon dans une loi de la famille exponentielle (le calcul de l'espérance revient à celui des statistiques exhaustives et le maximum se calcule par des formule fermées). Par rapport aux techniques d'optimisation usuelles couplées éventuellement à une intégration numérique, l'algorithme EM ne fournit pas de manière automatique les variances des estimateurs des paramètres. En effet, l'algorithme EM ne nécessite pas le calcul de la dérivée du logarithme de la vraisemblance observée et donc ne calcule pas non plus matrice de covariance asymptotique qui est l'inverse de le matrice d'information de Fisher calculée sur le logarithme de la vraisemblance observée en l'estimateur du paramètre. Par contre, pour un paramètre réel, la matrice de covariance asymptotique de données complètes, i.e l'inverse de la matrice d'information de Fisher pour la log-vraisemblance observée s'obtient à l'étape E. La matrice d'information de Fisher qui nous intéresse est alors obtenue par multiplication par un facteur issu de l'algorithme EM, voir par exemple la référence [14] p. 63 pour plus de détails.

Utilisation de l'algorithme EM avec des données tronquées
Dans l'exemple présenté ci-dessus nous sommes bien en présence de données tronquées. Si on reprend la preuve qu'à chaque itération de l'algorithme EM, la log-vraisemblance augmente on peut montrer qu'il est possible de remplacer l'étape E par une étape « EC » d'espérance conditionnelle à l'observation de tranches pour la variable d'intérêt ou la variable latente à la sélection et que l'on obtient toujours que la log-vraisemblance augmente à chaque itération des étapes EC et M. En effet nous pouvons écrire par exemple si z n'est jamais observé et y est observé dans une tranche T

$$f_\theta(y,z)\mathbb{1}_{y\in T} = \left(\iint_{T\times R} f_\theta(y,z)dydz\right) f_\theta\left(z|Y\in T\right),$$

[7] Le résultat résulte simplement de la convexité de la fonction $x \mapsto x\log(x)$ et de l'inégalité de Jensen.

puis en passant au logarithme, en multipliant les deux membres par $f_{\theta^t}(z|y \in T)$, en intégrant par rapport à z et étant donné que le membre de gauche est constant on obtient alors par l'inégalité de Jensen on augmente le dernier terme dès que $\theta^t \neq \theta$ et l'on souhaite à chaque étape maximiser le premier terme du second membre.

$$\log\left(\iint_{T\times R} f_\theta(y,z)dydz\right) = \int_{z\in R} \log\left(f_\theta(y,z)1_{y\in T}\right) f_{\theta^t}(z|Y\in T)dz - \int_{z\in R} \log\left(f_\theta(z|Y\in T)\right) f_{\theta^t}(z|Y\in T)dz;$$

Le calcul d'une espérance conditionnelle peut se faire de manière simple si l'on sait simuler la loi conditionnelle, ce qui est le cas grâce à des algorithme d'acceptation/rejet. En effet, dans ce cas l'étape EC peut être remplacée par une étape de calcul de Monte Carlo de l'intégrale qui consiste simplement à faire la moyenne d'un nombre élevé de simulations dans la loi conditionnelle (par application de la loi des grands nombres[8]). Ce type d'algorithme s'apparente aux méthodes MCEM (Monte Carlo EM). L'étape EC est remplacée par le calcul approché de l'espérance conditionnelle des statistiques exhaustives. Il ne s'agit pas ici, comme dans certains cas avec l'algorithme EM, d'une imputation déterministe des données manquantes à cause des termes d'ordre 2. En effet, dans la log-vraisemblance sur données « complétées », les données manquantes apparaissent parfois élevées au carré ou dans des doubles produits.

Une deuxième variante de l'algorithme EM est aussi nécessaire lorsque la loi est une loi jointe. Il s'agit de l'algorithme ECM où l'étape M cette fois est décomposée en une mise à jour successive des paramètres les uns après les autres. Le symbole CM signifie cette fois maximisation conditionnelle. Les étapes M successives augmentent elles aussi à chaque fois la vraisemblance.

Notons $\beta^{(t)}$ le vecteur des paramètres $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$ et $\Sigma^{(t)}$ la matrice de covariance à l'itération t, Ξ_i la matrice $\begin{pmatrix}X_i & 0\\ 0 & X_i\end{pmatrix}$, E_c^t l'espérance conditionnelle aux observations des tranches ou fourchettes pour la variable et la variable latente à la sélection pour la loi de paramètres ceux de l'étape t et Z le vecteur $\begin{pmatrix}Y\\ S\end{pmatrix}$. Nous avons donc les deux étapes suivantes

- ECM1 : $\beta^{(t+1)} = \left(\sum_{i=1}^n \Xi_i{}'\left(\Sigma^{(t)}\right)^{-1}\Xi_i\right)^{-1}\left(\sum_{i=1}^n \Xi_i{}'\left(\Sigma^{(t)}\right)^{-1}E_c^t[Z_i]\right)$, remarquons que cette statistique étant linéaire en les observations le calcul revient à faire des moindres carrés sur données imputées de manière déterministe,
- ECM2 : $\Sigma^{(t+1)} = \frac{1}{n}E_c^t\left[\sum_{i=1}^n \left(Z_i - \Xi_i\beta^{(t+1)}\right)\left(Z_i - \Xi_i\beta^{(t+1)}\right)'\right]$, cette deuxième quantité est quant à elle non linéaire en les observations.

Remarquons enfin que dans le cas du modèle Tobit généralisé, R.J.A. Little et D. Rubin donnent dans [10] des formules fermées et une façon plus simple de procéder à une séquence EM. Nous rappelons cela en annexe.

[8]La vitesse de convergence est celle du théorème central limite c'est-à-dire $1/\sqrt{n}$, ce qui est peu efficace en dimension 1 si la fonction à intégrer est régulière mais efficace en dimension élevée

2.2.3. Le cas de données en clair et en fourchettes ou de variables qualitatives

La procédure en deux étapes est spécifique au cas des variables quantitatives, seulement nous pouvons toujours procéder au calcul du maximum de vraisemblance et utiliser toute sorte de spécification de lois. Ceci permet alors de traiter le cas de variables qualitatives ou de comptage. Rappelons ci-après le cas de déclarations de montants en clair, en fourchettes ou en tranches et de non-réponses. La contribution à la vraisemblance observée de l'unité i s'écrit alors comme[9]

$$L_i = \int_{\frac{b_i - x_i\beta_1}{\sigma}}^{\frac{h_i - x_i\beta_1}{\sigma}} P(S^{*3} \geq 0 \big| X = x_i, Y = u\sigma + x_i\beta_1)\varphi(u)du$$

$$= \int_{\frac{b_i - x_i\beta_1}{\sigma}}^{\frac{h_i - x_i\beta_1}{\sigma}} \Phi\left[\frac{1}{\sqrt{1-\rho^2}}(x_i\beta_2 + \rho u)\right]\varphi(u)du,$$

lorsque la fourchette est observée et sinon

$$L_i = P(S^{*3} < 0 \big| X = x_i,) = \Phi(-x_i\beta_2).$$

Le cas des tranches est celui où h_i et b_i sont déterminés ex-ante et seul le choix de la tranche et non plus ses bornes est aléatoire. Le cas d'une variable dichotomique est celui où ou bien $h_i = +\infty$ et $b_i = 0$ ou bien $b_i = -\infty$ et $h_i = 0$.

La version de l'algorithme EM présentée plus haut pour les données tronquées peut s'appliquer pour estimer les différents paramètres et faire éventuellement une imputation aléatoire simple. Enfin les problèmes avec variables qualitatives se modélisent en général avec des variables latentes dont on n'observe que le signe ou l'appartenance à un intervalle. La version précédente de l'algorithme EM permet aussi de faire du maximum de vraisemblance dans ce cadre.

2.2.4. Ouverture

Une autre littérature économétrique par des méthodes semi paramétrique d'appariement (méthode est appelée Hot-Deck lorsque l'on fait de l'imputation) avec des scores (eux basés sur des modèles paramétriques) s'est aussi développée, avec entre autre des contributions de D. Rubin et de J.I. Heckman. Les applications usuelles sont l'estimation de l'effet moyen d'un traitement médical ou d'un programme de formation professionnelle. Nous ne traitons pas de cet aspect ce document.

2.3. Quelques premiers résultats

Nous présentons dans ce qui suit une expérimentation menée conjointement avec P. Biscourp sur les données de l'enquête emploi (EE) appariée avec l'enquête Revenu Fiscaux (ERF). Dans l'enquête emploi les individus fournissent ou bien une valeur de leur salaire mensuel au mois de l'enquête ou bien répondent en tranche ou bien ne répondent pas du tout. Le système de tranche permet de récupérer une information partielle chez des individus qui ne savaient pas ou ne désiraient pas répondre précisément à ce montant. Par ailleurs nous disposons grâce à l'appariement du salaire déclaré aux impôts pour l'année précédente ou l'année en cours. Plusieurs études ont déjà montré que le salaire enquête emploi est un salaire mesuré avec un bruit. Ce bruit correspond bien sûr à des erreurs de saisies ou des erreurs mensuel/annuel mais aussi à des erreurs d'arrondis. Mais ce n'est pas tout, dans certains cas il tend à y avoir des sous-

[9] Il s'agit comme auparavant de la contribution de l'unité si les données étaient observées intégrée « par rapport aux données manquantes », c'est à dire la probabilité élémentaire correspondant à l'observétion.

déclarations[10]. Nous avons travaillé dans un premier temps sur la variable ZTSAO de l'enquête ERF qui n'est pas retravaillé, il pourrait par contre contenir des montants de chômage, aussi nous nous sommes restreints aux individus ayant été en emploi toute l'année. En quelque sorte nous faisons, à une légère différence de concept près et à une petite différence de date près, comme si nous avions une mesure collectée sans erreur. Cela permet de décomposer le problème mais aussi à ne pas, dans un premier temps, avoir à faire du maximum de vraisemblance afin d'intégrer l'information en tranche. Enfin notons qu'un modèle de non-réponse en clair a déjà été estimé en faisant figuré le revenu ERF comme co-variable et qu'il s'agit bien d'une variable dont le test de nullité du coefficient est rejeté ceci nous faisant conclure que cette sélection est a priori non-ignorable. Nous avons voulu comparer avec ce que nous auraient donné d'autres tests.

En premier, voici un cas rare où nous avons chez les non-répondants et chez les répondants en tranche une valeur du revenu. Nous avons donc pu estimer les paramètres de lois dans les groupes des répondants en clair, des répondants en tranche et des non-répondants. Ceci correspond à un modèle de mélange. Nous avons construit un test d'égalité des coefficients dans les trois groupes et du vecteur des coefficients dans le groupe des répondants et dans le groupe des non-répondants et enfin des répondants et des répondants en tranche. Les tests correspondent à des tests du rapport de vraisemblance la statistique « équivalente » suit une loi de Fisher pour notre modèle linéaire Gaussien. Nous avons construit des variables interagies avec les indicatrices d'appartenance à chacun des trois groupes et en faisant figurer les indicatrices des trois groupes, le tout sans constante. En outre, nous avons mené l'estimation dans 4 groupes construits en croisant le sexe avec les CS commençant par 3 et 4 contre celles commençant par 5 et 6. En effet, il est couramment admis que le modèle n'est pas additif si on fait figurer ces variables comme des co-variables, le même test de Fisher permettrait de le vérifier. Les résultats sont assez mitigés et il semble que les comportements soient significativement différent surtout chez les hommes de CS 3 et 4. Sans faire de groupe le test fait conclure que les comportements sont à chaque fois différent. Enfin, on peut remarquer que, du fait du rattrapage par le système de tranche, la non-réponse est faible et encore plus faible lorsque nous considérons des modèles par groupe. Nous envisageons donc d'empiler plusieurs années et de mener des tests sur plus d'observations.

Dans un deuxième temps, nous avons cherché à mettre en pratique la méthode d'estimation en deux étapes du modèle Tobit généralisé. En effet, en pratique même s'il n'y avait pas d'erreur de mesure nous ne disposerions pas de valeurs pour les réponses en tranche ou les non-réponses. Nous avons mené plusieurs estimations, en définissant la sélection comme la non-réponse en clair ou comme la non-réponse en clair ou en tranche. Les tests de nullité des coefficients dans le modèle Probit sont presque tous acceptés. Les seules variables un temps soit peu significatives sont des modalités manquantes de variables explicatives ou que l'entretien se soit mal passé ou enfin que l'appariement avec le numéro SIRET a échoué. Ceci confirme qu'un mécanisme de non-réponse est quelque chose de difficile à appréhender[11]. Le test de sélection MAR construit sur cette spécification fait accepter la sélection MAR si la sélection correspond à la non-réponse en clair. Par contre sur données EE, le test rejette l'hypothèse MAR ce qui pourrait peut être faire conclure que c'est plutôt l'erreur de mesure que la variable elle-même qui est corrélée avec la sélection. Mais la conclusion n'est pas du tout robuste. En enlevant quelques points curieux en haut et en bas de la distribution le test est à nouveau accepté. Sur le revenu ERF, il est vraisemblable qu'il y ait moins de cas curieux, nous en avons pourtant détecté. Etant donné que la sélection définie plus haut est un agrégat de non déclaration dissimulation ou mauvaise connaissance nous avons décidé de changer de champ de sélection et de considérer la non-réponse en clair ou en tranche, certainement plus proche de la non-réponse dissimulation, le test est rejeté mais au seuil de 10%. Ainsi il semble pour l'instant que le test correspondant à l'estimation d'un modèle Tobit généralisé soit peu concluant et peu robuste. II s'agit donc a priori d'une méthode à manier avec précaution. Nous avons, entre autre, en projet de tester les méthodes d'imputation afférentes.

[10] Ce mécanisme complexe avec un bruit inconnu non nécessairement centré et de variance inconnue fait que le débruitage, aussi appelé déconvolution est un problème difficile.
[11] Tout d'abord on n'observe que le signe de la variable latente mais aussi de manière général le comportement de non-réponse est mal connu et moins étudié que le comportement lui même.

2.4. Le cas de l'imputation et des sélections de phase 1 et 2

L'imputation est utilisée généralement pour traiter la non-réponse partielle. Cela consiste à fabriquer des valeurs là où les enregistrements sont manquants. Le travail d'imputation est, comme toujours lorsque l'on traite de la non-réponse, un travail de modélisation. On peut décider d'imputer de façon déterministe et par exemple de produire la valeur que lui prédit le modèle. Dans ce cas, l'objectif n'étant pas de deviner les valeurs des non-répondants mais plutôt de faire des estimations le moins biaisées possible, on ne s'autorise sur données imputées qu'à ne calculer des statistiques linéaires en les « observations ». Si on s'intéresse par contre à des statistiques non linéaires comme par exemple des quantiles ou un indice de Gini, ces méthodes ne sont plus adaptées et on préfère imputer par simulation. Nous avons vu que si la non-réponse et non-ignorable nous devons pour produire une imputation aléatoire simple, dans la loi où les paramètres sont les paramètres[12] estimés, disposer non pas d'un estimateur de $\rho\sigma$ mais plutôt d'un estimateur de ρ et de σ. Mais ceci permet de traiter la sélection de phase 3. En pratique, il est courant de ne pas imputer pour traiter la non-réponse totale et de ne pas faire de l'inférence modèle de théorie des sondages. Si on fait une estimation de sondage basée sur le plan de sondage on applique couramment, dans un deuxième temps, sur données reconstruites un estimateur par repondération. Dans ce cas, il ne semble pas nécessaire de corriger à ce niveau de la sélection de phase 1 ou 2, sinon on effectue le travail deux fois de suite et l'on risque au contraire d'introduire un nouveau biais. Alors on souhaite compléter le tableau à trous de sorte de produire un échantillon dans la loi conditionnelle aux deux premières sélections et on peut bien faire comme s'il y avait une seule phase de sélection. Ainsi, il ne semble pas clair de faire intervenir des poids de sondage correspondant à la sélection de phase 1 lorsque l'on cherche à « corriger » de la sélection de phase 3. Il semble tout aussi peu clair de vouloir prendre en compte des poids après calage qui au mieux rendent comptent de la sélection 1+2.

2.5. Conclusion

Le problème principal des méthodes paramétriques est leur sensibilité à une mauvaise spécification. Or toutes les hypothèses du modèle ne sont pas testables du fait que les non-réponses sont justement des non-réponses. On peut, si l'hypothèse initiale est fausse, faire plus de mal que de bien. Certains auteurs suggèrent dans ce cas d'estimer plusieurs modèles et éventuellement de proposer une correction moyenne afin de rendre la procédure plus « robuste ». Cette solution moyenne est malgré tout assez décevante. Des méthodes semi paramétriques où on ne modélise plus le comportement ont été proposées et sont présentée dans la partie qui suit.

3. La sélection non-ignorable, approche semi paramétrique

Dans les deux méthodes qui suivent on ne désire pas modéliser les variables d'intérêt sujettes à sélection. Ceci est d'ailleurs le point de vue usuel du sondeur. En effet, les spécifications sont toutes critiquables, que ce soit la modélisation linéaire de la moyenne, l'additivité ou la loi. Une bonne démarche est d'effectuer des tests d'hypothèse. La loi normale, par exemple, apparaît souvent comme loi limite mais dire que chaque item est le tirage d'une loi normale peut être un peu restrictif. Par contre, et c'est pour cela que l'on a fait figuré l'approche paramétrique au début, cette approche est certainement la plus simple pour effectuer des tests et pour faire de l'imputation par simulation. Enfin des transformations des variables, comme la transformation en logarithme pour un salaire[13], peuvent permettre d'obtenir des distributions que l'on peut raisonnablement modéliser par exemple par des lois normales.

Dans les deux approches qui suivent on modélise de façon complètement paramétrique le mécanisme de sélection. A nouveau précisons que la sélection est peut être plus difficile à

[12] Une variance de sondage sur données imputées sous estime la variance même si le modèle était le vrai modèle. En outre les paramètres estimés sont eux-mêmes des quantités aléatoires.
[13] La loi du logarithme d'une variable positive a des queues moins épaisses que la loi d'une variable positive. La transformation en logarithme permet aussi d'obtenir une loi dont le support est toute la droite réelle.

modéliser que le comportement. Ce d'autant plus que la théorie par exemple économique nous renseigne sur la forme de certains comportements. Néanmoins il est souvent admis que des spécifications différentes des modèles de non-réponse (Probit, Logit…) donnent des résultats assez similaires.

3.1. Une méthode par moindres carrés pondérés

Cette méthode a été proposée par J.F. Beaumont dans la référence [3]. La variable que l'on modélise est supposée être quantitative. Dans le papier l'auteur propose dans un premier temps une estimation par maximisation de la vraisemblance observée pour un modèle paramétrique du type de celui de la partie 2. Dans l'article, le résidu de la deuxième équation, celle pour la variable latente à la sélection, est logistique. La deuxième méthode proposée est assez similaire à l'algorithme EM car il est itératif et procède en une succession de deux étapes. La preuve de la convergence de l'algorithme vers les paramètres d'intérêt n'est pas fournit mais l'auteur indique que l'algorithme a toujours convergé.

Etape 1 : Initialisation des paramètres du modèle de non-réponse.

Etape 2 : Les paramètres courants du modèle de non-réponse fournissent des probabilités de réponse estimées pour les unités répondantes. On estime alors les paramètres du modèle de la variable d'intérêt, modèle semi paramétrique[14] par moindres carrés pondérés. Les poids correspondent à l'inverse des probabilités de réponse. Ce type d'estimation ne nécessite pas de spécifier la loi du résidu.

Etape 3 : Au vu des paramètres estimés de la loi de la variable d'intérêt, on prédit les données manquantes (i.e. on effectue une imputation déterministe)[15].

Etape 4 : On estime les paramètres du modèle de non-réponse (ici logistique) faisant intervenir la variable d'intérêt comme variable explicative. On revient à l'étape 2.

Cette méthode utilise par ailleurs une approximation des espérances conditionnelles apparaissant pour la partie où la variable d'intérêt est non observée. Il est également proposé un test de normalité des résidus pour le modèle de la variable d'intérêt en population générale, c'est à dire ici que la variable soit observée ou non mais de phase 2. Il est basé sur une estimation de la fonction de répartition à partir des probabilités de réponse estimées[16]. Enfin l'inférence d'une moyenne peut se faire soit par repondération soit sur données imputées à partir des estimations obtenues in fine.

3.2. Les méthodes de vraisemblance empirique

Les méthodes de vraisemblance empirique sont apparues il y a une vingtaine d'années et ont été appliquées à des problèmes de statistique d'enquête. Le lecteur peut par exemple se référer à [6,8,15,16]. Nous présentons ci-après l'adaptation de ces méthodes à la sélection non-ignorable. Le papier de référence est celui de Qin, Leung et Shao, le lecteur peut consulter la référence [8] pour plus de détails.

La vraisemblance empirique est construite sur les observations jointes d'une variable Y partiellement observée et d'une variable X qui est elle toujours observée. Dans le cas de la non-réponse partielle, la variable X est une variable du plan de sondage ou de suivi de collecte si la sélection correspond à l'agrégation de la sélection de phase 2 et 3 ou une variable complètement renseignée de l'enquête s'il n'y a pas pour celle-ci de non-réponse totale.

[14] On ne précise pas la loi du résidu mais seulement une hypothèse de moment.
[15] Comme nous l'avons vu dans le cadre d'un modèle linéaire Gaussien l'étape E de l'algorithme EM ne revient pas tout à fait à prédire les données manquantes. C'est le cas par contre pour d'autres types de modèles.
[16] Notons que comme en régression linéaire usuelle si le modèle est vrai les résidus estimés sont pourtant corrélés et de variances distinctes. Il s'agit donc d'une approximation.

Afin de pouvoir s'affranchir d'une vraisemblance paramétrique qui serait la vraisemblance observée, i.e. calculée sur toutes les observations, même lorsque Y manque, les auteurs ne construisent la vraisemblance que sur les observations jointes des variables X et de Y. Alors, seule la loi de la sélection conditionnelle à Y et X est paramétrée, elle est notée $w(x,y,\theta) = P_\theta\left(S=1 \middle| X=x, Y=y\right)$, $f(x,y)$ est la densité de la loi jointe de (X,Y) dont nous ne donnons pas d'expression, $(x_i, y_i)_{i=1}^n$ sont les réalisations correspondant à l'observation jointe de X et de Y, les réalisations pour lesquelles Y n'est pas observé sont indicées de $n+1$ à N. Nous supposons dans un premier temps également que le mécanisme de non-réponse totale est MCAR et que le tirage de l'enquête est celui d'un échantillon aléatoire simple. Les auteurs précisent que la méthode peut être adaptée, en considérant une pseudo vraisemblance empirique où la vraisemblance intègre les spécificités du plan de sondage, c.f. par exemple la référence [6]. La vraisemblance « semi paramétrique » s'écrit alors

$$\left\{\prod_{i=1}^{n} w(x_i,y_i,\theta) f(x_i,y_i)\right\} \prod_{i=n+1}^{N} \iint \left(1-w(x,y,\theta)\right) f(x,y)dxdy = W^n (1-W)^{N-n} \prod_{i=1}^{n} \frac{w(x_i,y_i,\theta) f(x_i,y_i)}{W}$$

,

où $W = p(M=0) = \iint w(x,y,\theta) f(x,y)dxdy$.

La vraisemblance observée sur tout l'échantillon s'écrirait comme dans la partie 2 plutôt

$$\left\{\prod_{i=1}^{n} w(x_i,y_i,\theta) f(x_i,y_i)\right\} \prod_{i=n+1}^{N} \int \left(1-w(x,y,\theta)\right) f(x_i,y)dy,$$

mais elle nécessite plutôt l'emploi de techniques paramétriques. La vraisemblance semi paramétrique ne mobilise donc pas toute l'information recueillie.

La fonction de répartition $F(x,y)$, de « dérivée » partielle par rapport à x et y (si la loi est diffuse) $f(x,y)$, est approchée par la fonction en escalier de saut p_i en (x_i,y_i), p_i correspond à une masse, il s'agit de la probabilité empirique que (X,Y) soit dans l'élément de surface définit par les observations « voisines ». Cette fonction en escalier est bien la fonction de répartition empirique. Il y a alors une multitude de paramètres : les masses p_i, la probabilité non conditionnelle W et le paramètre θ du modèle de sélection. Le problème de maximisation se résout en ajoutant des contraintes qui sont aussi parfois appelées contraintes de calage et en effectuant la maximisation sous contraintes. Les auteurs proposent les contraintes suivantes

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \left(w(x_i,y_i,\theta) - W\right) = 0, \quad \sum_{i=1}^{n} p_i (x_i - \mu_x) = 0,$$

où μ_x est la moyenne de X connue par ailleurs ou la moyenne observée sinon. Il reste alors dans ce cas à résoudre un système non linéaire de quatre équations à quatre inconnues : les estimateurs des deux multiplicateurs de Lagrange des « contraintes de moment », les estimateurs de W et de θ. Remarquons que la contrainte sur la moyenne de X permet d'utiliser malgré tout les réalisations de X pour lesquelles Y n'est pas observé et qui n'interviennent pas dans la maximisation de la vraisemblance. Un estimateur sous l'approche modèle de théorie des sondage de la moyenne de la population est alors donné par $\sum_{i=1}^{n} \hat{p}_i y_i$.

Sous des hypothèses de régularité il y a consistance, c'est à dire que les estimateurs des paramètres
$(\hat{\theta}, \hat{W})$ convergent vers les vrais paramètres, et l'estimateur de la moyenne est convergent et asymptotiquement. L'auteur envisage, en collaboration avec H. Harari-Kermadec, voir la

référence [4], de mettre en pratique les méthodes de vraisemblance empirique et d'étudier quelques approfondissements. Il est en effet possible, voir [4], de généraliser les méthodes de vraisemblance empirique en considérant des familles plus générales de contrastes. Un contraste basé sur la mesure du chi2 permet notamment d'obtenir une solution plus directe, avec des formules fermées ne nécessitant pas le calcul numérique des zéros d'une fonction de plusieurs variables, du problème d'optimisation. Enfin il est a priori possible d'effectuer des tests d'hypothèses par cette approche et de tester la nullité d'un coefficient devant la variable Y dans un modèle de sélection Logit ou Probit. Enfin on peut aussi envisager l'étude d'autres quantités sur la loi de Y.

Concluons finalement que la sélection peut engendrer des biais dans les estimations. Ceci est tout à fait rédhibitoire si le biais ne tend pas vers 0. Un travail préliminaire et des tests d'hypothèses permettent de faire des estimations non biaisées si la sélection est MAR. De la littérature existe pour des modèles de sélection non-ignorable. Par contre l'auteur n'est pas en mesure de proposer une solution universelle. Il s'agit d'un domaine de recherche toujours actif et d'un problème complexe.

Bibliographie

[1] Agarwal S., « Learning from incomplete data », *Techical report*, University of California San Diego, 2001.

[2] Ardilly P., « Les techniques de sondage », *Technip,* 1994.

[3] Beaumont J.F., « Une méthode d'estimation en présence de non-réponse non-ignorable », *Techniques d'enquêtes*, vol 26, pp 145-151, 2000.

[4] Bertail P., Harari-Kermadec H., Ravaille D., « γ -divergence empirique et vraisemblance empirique généralisée », *Document de Travail du CREST*, n°2004-29, 2004.

[5] Caron N., « Les principales techniques de correction de la non-réponse, et les modèles associés », *Méthodologie statistique,* vol 9604, INSEE, 1996.

[6] Chen J., Sitter R.R., « A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys », *Statistica Sinica*, vol 9, pp 385-406, 1999.

[7] Nathan G., « L'analyse des données issues des enquêtes longitudinales sous un plan de sondage informatif », *Echantillonnage et méthode d'enquêtes* dir. Pascal Ardilly, Dunod, pp 227-240, 2004.

[8] Qin L., Leung D., Shao J., « Estimation with survey data under nonignorable or informative sampling », *Journal of the American Statistical Association*, vol 97, n°457, pp 193-200, 2002.

[9] Little R.J.A., « To model or not to model? Competing Modes of Inference for Finite Population Sampling », *Journal of the American Statistical Association*, vol 99, n°466, pp 546-556, 2004.

[10] Little R.J.A., Rubin D.B., « Statistical analysis with missing data (2nd editition) », *Wiley series in probability and statistics*, chapitre 15, 2002.

[11] Little R.J.A., « A test of missing completely at random for multivariate data with missing values », *Journal of the American Statistical Association*, vol 83, n°404, pp 1198-1202, 1988.

[12] Little R.J.A., « A note about models for selectivity bias », *Econometrica*, vol 53, pp 1469-1474, 1985.

[13] Lollivier S., « Endogénéité dans un système d'équations nomal bivarié avec variables qualitatives », *Journées de méthodologie statistique*, 2002.

[14] Schafer J., « Analysis of incomplete multivariate data », *Chapman & Hall, Monographs on Statistics and Applied Probability,* vol 72, 1997.

[15] Wang Q., Rao J.N.K., « Empirical likelihood-based inference under imputation for missing response data », *The Annals of Statistics,* vol 30, n°3, pp 896-924, 2002.

[16] Wu C., « Some algorithmic aspects of the empirical likelihood method in survey sampling », *Statistica Sinica,* vol 14, pp 1057-1067, 2004.

Annexe 1 : quelques résultats des estimations et tests menés sur l'appariement enquête emploi et enquête revenus fiscaux, année 1999.

Il s'agit de premiers résultats d'un travail en commun avec P. Biscourp, les résultats plus détaillés et les approfondissements seront publiés autre part.

Régression du logarithme du salaire ERF interagit avec 3 types de déclaration (en clair, en tranche, pas de déclaration):

Le groupe des hommes de CS commençant par 3 et 4 :

var	coef	Student	p-value
scont1	11,412	0,095	<,0001
stran1	10,952	0,264	<,0001
snrep1	11,326	0,419	<,0001
lnheur0	0,154	0,022	<,0001
lnheur1	0,317	0,063	<,0001
lnheur2	0,122	0,107	0,256
tpart0	-0,269	0,033	<,0001
tpart1	-0,086	0,094	0,364
tpart2	-0,448	0,139	0,001
a10	-0,181	0,018	<,0001
a11	-0,205	0,054	0,000
a12	-0,219	0,091	0,017
a30	0,086	0,014	<,0001
a31	0,068	0,039	0,081
a32	0,060	0,065	0,355
a40	0,140	0,012	<,0001
a41	0,157	0,034	<,0001
a42	0,078	0,056	0,170
anc10	-0,022	0,018	0,231
anc11	-0,059	0,055	0,285
anc12	0,121	0,092	0,190
var	coef	Student	p-value
cs420	-0,321	0,032	<,0001
cs421	-0,413	0,103	<,0001
cs422	-0,288	0,192	0,133
cs430	-0,380	0,032	<,0001
cs431	-0,367	0,133	0,006
cs432	-0,431	0,183	0,018
cs450	-0,284	0,030	<,0001
cs451	-0,397	0,091	<,0001
cs452	-0,188	0,127	0,139
cs460	-0,281	0,020	<,0001
cs461	-0,408	0,057	<,0001
cs462	-0,311	0,081	0,000
cs470	-0,346	0,022	<,0001
cs471	-0,383	0,063	<,0001
cs472	-0,359	0,091	<,0001
cs480	-0,309	0,024	<,0001
cs481	-0,375	0,068	<,0001
cs482	-0,264	0,095	0,005
fonc00	0,022	0,022	0,318
fonc01	-0,038	0,066	0,562
fonc02	0,005	0,098	0,959
fonc20	-0,036	0,019	0,062
fonc21	-0,131	0,060	0,029

var	coef	Student	p-value
anc30	0,042	0,017	0,013
anc31	0,087	0,052	0,093
anc32	0,089	0,077	0,252
anc40	0,115	0,016	<,0001
anc41	0,121	0,047	0,010
anc42	0,264	0,072	0,000
ancm0	-0,712	0,210	0,001
ancm1	0,086	0,228	0,706
ancm2	0,620	0,338	0,067
etcl0	0,053	0,018	0,003
etcl1	0,033	0,055	0,549
etcl2	0,120	0,091	0,185
dip10	0,305	0,022	<,0001
dip11	0,334	0,065	<,0001
dip12	0,719	0,097	<,0001
dip30	0,142	0,021	<,0001
dip31	0,083	0,063	0,186
dip32	0,405	0,091	<,0001
dip40	0,075	0,021	0,000
dip41	0,009	0,062	0,886
dip42	0,199	0,086	0,021
var	coef	Student	p-value
fonc22	-0,133	0,075	0,075
fonc40	-0,064	0,040	0,109
fonc41	-0,050	0,089	0,576
fonc42	0,230	0,174	0,186
fonc50	-0,060	0,058	0,300
fonc51	0,100	0,150	0,505
fonc52	0,305	0,331	0,356
fonc60	-0,015	0,022	0,474
fonc61	-0,050	0,062	0,419
fonc62	0,108	0,096	0,261
fonc70	0,006	0,022	0,790
fonc71	0,031	0,067	0,646
fonc72	0,025	0,084	0,763
fonc80	0,010	0,018	0,585
fonc81	-0,109	0,054	0,042
fonc82	-0,066	0,078	0,395
fonc90	0,197	0,027	<,0001
fonc91	0,115	0,075	0,125
fonc92	0,366	0,114	0,001
foncm0	-0,173	0,107	0,105
foncm1	-0,105	0,313	0,738
foncm2	0,736	0,272	0,007
EB0	-0,005	0,036	0,890

var	coef	Student	p-value
dip50	-0,001	0,020	0,951
dip51	-0,011	0,061	0,863
dip52	0,197	0,085	0,021
dip70	-0,083	0,024	0,001
dip71	-0,059	0,068	0,385
dip72	-0,005	0,098	0,961
cs310	-0,115	0,085	0,174
cs311	0,278	0,231	0,228
cs312	0,665	0,268	0,013
cs330	-0,090	0,028	0,001
cs331	-0,241	0,079	0,002
cs332	-0,317	0,123	0,010
cs340	-0,128	0,031	<,0001
cs341	-0,218	0,101	0,031
cs342	-0,234	0,145	0,107
cs350	-0,201	0,051	<,0001
cs351	-0,341	0,114	0,003
cs352	-0,126	0,220	0,568
cs380	-0,031	0,021	0,136
cs381	-0,063	0,055	0,252
cs382	-0,284	0,081	0,001
var	coef	Student	p-value
EB1	0,051	0,114	0,652
EB2	0,062	0,152	0,682
EC0	0,020	0,028	0,470
EC1	0,023	0,086	0,786
EC2	-0,140	0,128	0,273
ED0	0,020	0,032	0,539
ED1	-0,025	0,104	0,811
ED2	0,055	0,107	0,607
EE0	-0,053	0,021	0,013
EE1	0,019	0,061	0,750
EE2	-0,122	0,078	0,120
EG0	0,023	0,030	0,440
EG1	-0,002	0,081	0,984
EG2	-0,071	0,123	0,566
EH0	-0,048	0,026	0,060
EH1	-0,164	0,070	0,020
EH2	0,033	0,110	0,764
EJ0	-0,054	0,022	0,014
EJ1	-0,132	0,060	0,027
EJ2	-0,184	0,091	0,043
EK0	-0,044	0,029	0,124
EK1	-0,067	0,071	0,345
EK2	-0,286	0,119	0,017

var	coef	Student	p-value
EL0	0,001	0,027	0,976
EL1	-0,109	0,071	0,125
EL2	-0,232	0,112	0,039
EM0	-0,042	0,046	0,358
EM1	-0,001	0,161	0,994
EM2	-0,680	0,210	0,001
EN0	-0,071	0,020	0,000
EN1	-0,009	0,053	0,872
EN2	-0,317	0,084	0,000
EP0	-0,224	0,036	<,0001
EP1	-0,134	0,113	0,236
EP2	-0,151	0,137	0,271
EQ0	-0,109	0,029	0,000
EQ1	-0,127	0,096	0,187
EQ2	-0,359	0,128	0,005
ER0	-0,110	0,025	<,0001
ER1	-0,105	0,080	0,188
ER2	-0,295	0,108	0,006

var	coef	Student	p-value
tent10	-0,082	0,014	<,0001
tent11	0,002	0,042	0,959
tent12	-0,044	0,066	0,510
tent20	-0,045	0,015	0,003
tent21	0,031	0,045	0,498
tent30	-0,018	0,014	0,211
tent22	0,032	0,065	0,624
tent31	0,021	0,041	0,603
tent32	0,085	0,063	0,175
tentm0	-0,023	0,019	0,225
tentm1	0,104	0,055	0,060
tentm2	-0,229	0,087	0,009
tsoi0	0,043	0,012	0,000
tsoi1	0,093	0,035	0,008
tsoi2	0,133	0,048	0,005
tnui0	0,039	0,015	0,009
tnui1	-0,026	0,043	0,543
tnui2	-0,017	0,065	0,792

var	coef	Student	p-value
tsam0	-0,014	0,012	0,220
tsam1	-0,086	0,035	0,013
tsam2	-0,101	0,050	0,044
tdim0	0,043	0,014	0,002
tdim1	0,129	0,039	0,001
tdim2	0,093	0,057	0,101
idf0	0,110	0,012	<,0001
idf1	0,058	0,032	0,071
idf2	0,135	0,044	0,002
proxy0	0,048	0,009	<,0001
proxy1	0,058	0,027	0,029
proxy2	0,225	0,040	<,0001
quali10	0,051	0,032	0,119
quali11	0,032	0,053	0,551
quali12	0,051	0,051	0,318
sirmiss0	0,022	0,021	0,284
sirmiss1	0,053	0,056	0,343
sirmiss2	0,382	0,093	<,0001

Test d'égalité des coeffcients dans les groupes : réponse en clair/en tranche

	DF	carré	Fisher	p-value
Numérateur	61	0.14004	1.62	0.0017
Denominateur	5118	0.08646		

Test d'égalité des coeffcients dans les groupes : réponse en clair/aucune réponse

	DF	carré	Fisher	p-value
Numérateur	61	0.24423	2.82	<.0001
Denominateur	5118	0.08646		

Test d'égalité des coeffcients dans les 3 groupes

	DF	carré	Fisher	p-value
Numérateur	122	0.18161	2.10	<.0001
Denominateur	5118	0.08646		

Pour les autres groupes nous ne précisons que le résultat des tests du rapport de vraisemblance,

Le groupe des femmes de CS commençant par 3 et 4 :

Test d'égalité des coeffcients dans les groupes : réponse en clair/en tranche

	DF	carré	Fisher	p-value
Numérateur	57	0.08527	1.26	0.0903
Denominateur	3381	0.06757		

Test d'égalité des coeffcients dans les groupes : réponse en clair/aucune réponse

	DF	carré	Fisher	p-value
Numérateur	57	0.07439	1.10	0.2819
Denominateur	3381	0.06757		

Test d'égalité des coeffcients dans les 3 groupes

	DF	carré	Fisher	p-value
Numérateur	114	0.07959	1.18	0.0987
Denominateur	3381	0.06757		

Le groupe des hommes de CS commençant par 5 et 6 :

Test d'égalité des coeffcients dans les groupes : réponse en clair/en tranche

	DF	carré	Fisher	p-value
Numérateur	61	0.06211	1.18	0.1577
Denominateur	7225	0.05255		

Test d'égalité des coeffcients dans les groupes : réponse en clair/aucune réponse

	DF	carré	Fisher	p-value
Numérateur	61	0.089	1.69	0.0006
Denominateur	7225	0.05255		

Test d'égalité des coeffcients dans les 3 groupes

	DF	carré	Fisher	p-value
Numérateur	122	0.07516	1.43	0.0014
Denominateur	7225	0.05255		

Le groupe des femmes de CS commençant par 5 et 6 :

Test d'égalité des coeffcients dans les groupes : réponse en clair/en tranche

	DF	carré	Fisher	p-value
Numérateur	60	0.07273	1.07	0.3277
Denominateur	6340	0.06780		

Test d'égalité des coeffcients dans les groupes : réponse en clair/aucune réponse

	DF	carré	Fisher	p-value
Numérateur	60	0.08243	1.22	0.1233
Denominateur	6340	0.06780		

Test d'égalité des coeffcients dans les 3 groupes

	DF	carré	Fisher	p-value
Numérateur	120	0.07516	1.13	0.1547
Denominateur	6340	0.06780		

Régression du logarithme du salaire ERF avec comme variable supplémentaire l'inverse du ratio de Mills pour le paramètre du modèle de sélection estimé par un modèle Probit (après nettoyage succinct des points « aberrants »), cas de sélection correspondant à la non-réponse en clair ou en tranche:

var	coef	Student	p-value
Intercept	11,647	1,039	<,0001
mills	-1,032	0,580	0,075
lnheur	0,426	0,020	<,0001
tpart	-0,286	0,032	<,0001
femme	-0,069	0,019	0,000
a1	-0,111	0,019	<,0001
a3	-0,094	0,074	0,201
a4	-0,123	0,113	0,278
anc1	-0,096	0,042	0,022
anc3	0,081	0,014	<,0001
anc4	0,097	0,026	0,000
etcl	-0,002	0,030	0,944
dip1	0,269	0,034	<,0001
dip3	0,160	0,025	<,0001
dip4	0,046	0,025	0,065
dip5	0,027	0,012	0,020
dip7	-0,061	0,010	<,0001
dipm	5,172	2,946	0,079
cs31	-0,131	0,274	0,633
cs33	0,507	0,049	<,0001
cs34	0,539	0,049	<,0001
cs35	0,149	0,135	0,269
cs37	0,304	0,109	0,005
cs38	0,230	0,135	0,088
cs42	0,271	0,022	<,0001
cs43	0,326	0,065	<,0001

var	coef	Student	p-value
cs45	0,222	0,023	<,0001
cs46	-0,021	0,123	0,864
cs47	0,075	0,052	0,146
cs48	0,026	0,096	0,786
cs52	0,039	0,016	0,012
cs53	0,091	0,038	0,017
cs54	-0,171	0,131	0,194
cs55	-0,024	0,047	0,605
cs56	-0,176	0,019	<,0001
cs63	-0,085	0,039	0,027
cs64	-0,141	0,058	0,015
cs65	0,165	0,094	0,080
cs67	-0,342	0,131	0,009
cs68	-0,284	0,084	0,001
cs69	-0,144	0,036	<,0001
fonc0	0,035	0,016	0,029
fonc2	0,019	0,015	0,198
fonc3	0,071	0,099	0,474
fonc4	0,010	0,016	0,536
fonc5	0,022	0,018	0,235
fonc6	0,100	0,035	0,004
fonc7	0,032	0,014	0,024
fonc8	0,169	0,073	0,020
fonc9	0,241	0,050	<,0001
foncm	-0,691	0,398	0,083
EA	-0,154	0,043	0,000

var	coef	Student	p-value
EB	-0,061	0,019	0,002
EC	-0,073	0,021	0,001
ED	-0,209	0,092	0,024
EE	-0,126	0,055	0,021
EF	-0,172	0,072	0,017
EG	0,003	0,025	0,901
EH	-0,066	0,017	0,000
EJ	-0,212	0,060	0,000
EK	-0,120	0,048	0,012
EM	-0,185	0,038	<,0001
EN	-0,121	0,021	<,0001
EP	-0,343	0,109	0,002
EQ	-0,246	0,070	0,001
ER	-0,120	0,018	<,0001
tent1	-0,022	0,034	0,505
tent2	0,042	0,055	0,440
tent3	0,020	0,032	0,534
tentm	-0,044	0,012	0,000
tsoi	-0,009	0,023	0,708
tnui	0,084	0,019	<,0001
tsam	0,050	0,030	0,101
tdim	0,094	0,031	0,002
idf	-0,054	0,088	0,544
proxy	-0,130	0,073	0,075
quali1	0,769	0,441	0,081
sirmiss	-0,024	0,030	0,410

Si la sélection correspond à la non-réponse en clair, la p-value de la statistique de Student correspondant au test de nullité du coefficient devant l'inverse du ratio de Mills (la seule licite sous l'hypothèse nulle de sélection MAR) est de 0,154.

Annexe 2 : version de l'algorithme EM présentée dans le livre [10] de R.J.A. Little et D. Rubin

Les auteurs rappellent les formules fermées qui suivent pour les termes apparaissant lors du calcul de l'espérance conditionnelle de la log-vraisemblance totale. En effet les termes y_i , s_i^{*3}, $(y_i^*)^2$, $(s_i^{*3})^2$, $y^* s_i^{*3}$ apparaissent en développant le terme

$$-\frac{1}{\sigma^2(1-\rho^2)}\left[(y_i - x_i\beta_1)^2 - 2\rho\sigma(s_i - x_i\beta_2)(y_i - x_i\beta_1) + \sigma^2(s_i - x_i\beta_2)^2\right]$$

alors l'étape (EC) consiste à remplacer dans la log-vraisemblance totale les observations manquantes de y_i^*, s_i^{*3}, $(y_i^*)^2$, $(s_i^{*3})^2$, $y^* s_i^{*3}$ par leurs prédictions pour les valeurs courantes des paramètres. D'après les formules de conditionnement dans un vecteur Gaussien, nous obtenons

$$E\left[S^{*3}\middle|S^{*3} < 0, X = x_i\right] = x_i\beta_2^{(t)} - \lambda\left(-x_i\beta_2^{(t)}\right),$$

$$E\left[Y\middle|S^{*3} < 0, X = x_i\right] = x_i\beta_1^{(t)} - \rho^{(t)}\sigma^{(t)}\lambda\left(-x_i\beta_2^{(t)}\right),$$

$$E\left[Y^2\middle|S^{*3} < 0, X = x_i\right] = (x_i\beta_1^{(t)})^2 + (\sigma^{(t)})^2 - \rho^{(t)}\sigma^{(t)}\lambda\left(-x_i\beta_2^{(t)}\right)\left(2x_i\beta_1^{(t)} - \rho^{(t)}\sigma^{(t)}x_i\beta_2^{(t)}\right),$$

$$E\left[YS^{*3}\middle|S^{*3} < 0, X = x_i\right] = x_i\beta_1^{(t)}\left[x_i\beta_2^{(t)} - \lambda\left(-x_i\beta_2^{(t)}\right)\right] + \rho^{(t)}\sigma^{(t)},$$

et si y_i^* est observé

$$E\left[S^{*3}\middle|Y = y_i, S^{*3} \geq 0, X = x_i\right] = s_i^{(t)} + \sqrt{1-(\rho^{(t)})^2}\,\lambda\left(\frac{s_i^{(t)}}{\sqrt{1-(\rho^{(t)})^2}}\right),$$

$$E\left[(S^{*3})^2\middle|Y = y_i, S^{*3} \geq 0, X = x_i\right] = 1-(\rho^{(t)})^2 + (s_i^{(t)})^2 + s_i^{(t)}\sqrt{1-(\rho^{(t)})^2}\,\lambda\left(\frac{s_i^{(t)}}{\sqrt{1-(\rho^{(t)})^2}}\right)$$

$$\text{où } s_i^{(t)} = x_i\beta_2^{(t)} + \frac{\rho^{(t)}}{\sigma^{(t)}}\left(y_i - x_i\beta_1^{(t)}\right).$$

Les auteurs précisent que dans ce cas particulier l'étape (M) peut être obtenue de manière équivalente en :

- régressant S^{3*} sur X et changeant $\beta_2^{(t)}$ en $\beta_2^{(t+1)}$ le coefficient de la régression,
- régressant Y sur S^{3*} et X , ce qui donne les coefficients $\delta^{(t)}$ pour S^{3*} et $\gamma^{(t)}$ pour X et la variance $(\tilde{\sigma}^{(t)})^2$,
- posant $\beta_1^{(t+1)} = \gamma^{(t)} + \delta^{(t)}\beta_2^{(t)}$, $\sigma^{(t+1)} = (\tilde{\sigma}^{(t)})^2 + (\delta^{(t)})^2$ et $\rho^{(t+1)} = \frac{\delta^{(t)}}{\sigma^{(t+1)}}$.