Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche sociale et économique

NON-PARTICIPANTS ADMISSIBLES ET INDIVIDUS NON ADMISSIBLES EN TANT QUE DOUBLE GROUPE TÉMOIN DANS LES PLANS EXPÉRIMENTAUX DE DISCONTINUITÉ DE LA RÉGRESSION

Erich Battistin[1], Enrico Rettore[2]

RÉSUMÉ

L'attrait du plan expérimental de discontinuité de la régression (PEDR), dans sa formulation précise ou floue, tient à sa grande similarité avec un plan expérimental formel. Par contre, son applicabilité est limitée, puisqu'il n'est pas très fréquent que les individus soient affectés au groupe exposé au traitement d'après une mesure observable par l'analyste avant le programme. En outre, il permet de déterminer l'effet moyen du programme seulement sur une sous-population très spécifique d'individus. Dans le présent article, nous montrons que le PEDR se généralise facilement aux cas où on établit l'admissibilité au programme d'après une mesure observable avant le programme et où on permet l'autosélection des individus admissibles dans le groupe exposé au traitement selon un processus inconnu. Ces conditions s'avèrent aussi fort pratiques pour la définition d'un test de spécification applicable aux estimateurs non expérimentaux conventionnels de l'effet d'un programme. Nous décrivons explicitement les exigences concernant les données.

MOTS CLÉS : Évaluation de programme; deuxième groupe témoin; tests de spécification.
Classification JEL : C4, C8.

1. INTRODUCTION

Le problème central de l'évaluation des politiques publiques est de faire la distinction entre leur effet causal et l'effet confusionnel des autres facteurs qui influent sur le résultat étudié. La sélection aléatoire des unités faisant l'objet de l'intervention produit un groupe de traitement et un groupe témoin qui sont équivalents à tous les égards, sauf leur situation d'exposition. Donc, par construction, dans une expérience entièrement aléatoire, toute différence entre les deux groupes après l'intervention n'est pas le reflet de différences avant l'intervention. Par conséquent, les différences entre les unités exposées et les unités témoins sont dues entièrement à l'intervention elle-même.

Malheureusement, dans la plupart des cas, la randomisation est irréalisable, pour des raisons d'éthique ou simplement parce que l'affectation au traitement ne peut pas être contrôlée par l'analyste. En outre, même dans les cas où l'analyste est capable de randomiser l'affectation, les unités peuvent ne pas se conformer au résultat et se retirer du programme d'intervention ou en chercher un autre (voir Heckman et Smith, 1995). Un exemple bien connu et très largement utilisé d'affectation randomisée est le programme JTPA aux États-Unis qui, à l'heure actuelle, dessert chaque année presque un million de personnes économiquement défavorisées (voir Friedlander et coll., 1997). L'affectation aléatoire a lieu avant l'inscription effective au programme, mais une fraction constante des individus affectés aléatoirement au groupe de traitement n'y participe pas. Pour certaines composantes du JTPA, ce comportement d'inobservation semble être non négligeable (voir, par exemple, Heckman et coll., 1998b).

Dans cette situation, l'expérience idéale n'est pas réalisée entièrement, puisque la participation devient (du moins partiellement) volontaire : la formation est prodiguée uniquement aux individus qui satisfont certains critères de besoins et se conforment aux résultats de la randomisation. Il s'ensuit que la participation dépend de caractéristiques observables et inobservables des individus qui pourraient être corrélées aux résultats d'intérêt. Dans cette situation,

[1]Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK, erich_b@ifs.org.uk
[2]Département de statistique, Université de Padoue, Via Cesare Battisti 241, 35121 Padoue, Italie, rettore@stat.unipd.it

les différences entre le groupe de traitement et le groupe témoin par rapport aux résultats d'intérêt pourraient résulter de l'autosélection des unités dans le programme d'intervention.

Déterminer si les variations observées du résultat d'intérêt peuvent être attribuées à l'intervention elle-même et non à d'autres causes possibles s'avère encore plus compliqué dans des conditions non expérimentales. Dans cette situation, l'estimation des relations de cause à effet que l'on pourrait croire être vérifiée entre le programme et ses résultats dépend, en fait, d'hypothèses non vérifiables concernant le comportement des individus. Étant donné que la situation idéale pour les évaluateurs des politiques est de connaître entièrement les mécanismes menant les individus à participer au programme, et que, dans la plupart des cas, un tel mécanisme est inconnu (soit à cause de la non-conformité des individus dans des conditions expérimentales ou à cause de l'insuffisance des connaissances dégagées des études observationnelles), la question se pose alors de savoir comment tirer le maximum de chaque plan expérimental pour obtenir des estimations raisonnables des effets du programme.

Certains cas donnent lieu au plan connu sous le nom de plan expérimental de discontinuité de la régression (PEDR) [pour *Regression Discontinuity Design*] (voir Campbell, 1969, Rubin, 1977, Trochim, 1984). Selon ce plan expérimental, l'affectation des unités est fondée uniquement sur des variables de pré-intervention observables par l'analyste et la probabilité de participation varie de façon discontinue en fonction de ces variables. Pour fixer les idées, considérons le cas où un groupe d'unités disposées à participer est subdivisé en deux groupes en prenant comme critère le fait que la mesure avant l'intervention est supérieure ou inférieure à un seuil connu. Les individus dont le score est supérieur au seuil sont exposés à l'intervention, tandis que ceux dont le score est inférieur n'y sont pas exposés.

Ce plan expérimental présente à la fois des avantages et des inconvénients comparativement à ses concurrents. D'une part, dans le voisinage du seuil de sélection, un PEDR présente certaines caractéristiques d'une expérience pure. En ce sens, il est certainement plus intéressant qu'un plan non expérimental. Puisque les sujets qui appartiennent au groupe exposé au traitement et au groupe témoin diffèrent uniquement par rapport à la variable en fonction de laquelle l'affectation à l'intervention est établie (et par rapport à tout autre variable qui y est corrélée), on peut neutraliser les facteurs confusionnels en faisant la distinction entre les participants marginaux et les non-participants marginaux. Dans ce contexte, le terme *marginal* désigne les unités qui ne sont *pas trop éloignées* du seuil de sélection. La comparaison des résultats moyens obtenus pour les unités marginales traitées et témoins permet de déterminer l'effet moyen de l'intervention localement par rapport au seuil de sélection. Intuitivement, pour que la détermination au seuil d'inclusion soit vérifiée, il faut que toute discontinuité dans la relation entre les résultats étudiés et la variable déterminant la situation de traitement soit entièrement attribuable au traitement lui-même (ce qui nécessite certaines conditions de régularité au seuil de sélection discutées par Hahn et coll., 2001).

D'autre part, le PEDR présente deux limites importantes. Premièrement, par définition, sa faisabilité est limitée aux cas où la sélection se fait sur des variables observables avant l'intervention; par contre, il n'en est pas souvent ainsi. Deuxièmement, même quand un tel plan expérimental s'applique, en présence d'effets hétérogènes, il permet de déterminer l'effet moyen de l'intervention seulement au seuil de sélection. Dans la situation raisonnable de l'hétérogénéité de l'effet entre les unités, cet effet *local* pourrait être fort différent de l'effet pour les unités éloignées du seuil de sélection. Pour déterminer l'effet moyen sur une population plus large, on ne peut recourir qu'à un estimateur non expérimental dont la cohérence pour l'effet prévu dépend intrinsèquement d'hypothèses de comportement (non vérifiables).

Dans le présent article, nous montrons que le champ d'applicabilité du PEDR inclut tous les cas où la population pertinente est subdivisée en deux sous-populations — d'individus admissibles et non admissibles — à condition que i) la situation d'admissibilité soit établie par rapport à une variable continue et ii) l'information sur les unités non admissibles et sur les unités admissibles non participantes soit disponible. Alors, l'effet moyen sur les unités participantes situées dans le voisinage du seuil d'admissibilité est déterminé sous des conditions de régularité très faibles, quelle que soit la façon dont les unités admissibles s'auto-sélectionnent dans le programme. Nous établissons un lien explicite avec la littérature sur le PEDR (uniquement des références implicites jusqu'à présent; Angrist, 1998) et nous déterminons les conditions de régularité nécessaires pour que la détermination de l'effet soit vérifiée (van der Klaauw, 2002 et Heckman et coll., 1999 font remarquer qu'un PEDR découle des règles d'admissibilité, mais ils ne discutent pas des conditions de détermination de l'effet). Nous montrons aussi que la discontinuité dans la

règle d'admissibilité mène à des conditions de régularité pour l'identification plus faibles que celles découlant du PEDR dit *flou* (Hahn et coll., 2001) auquel appartiennent à première vue les conditions dont nous discutons.

Deuxièmement, un corollaire direct du résultat qui précède est que le biais de sélection au seuil d'admissibilité est déterminable. On peut alors vérifier formellement si l'une des longues séries d'estimateurs non expérimentaux existants nous permet de corriger ce biais. Si un estimateur non expérimental permet de corriger le biais de sélection, même si ce n'est que pour une sous-population particulière — plus précisément, les unités situées dans le voisinage du seuil d'admissibilité — on pourra l'utiliser avec plus de confiance pour estimer les effets causals sur une population plus large.

Nous établissons les liens avec les résultats déjà publiés. En particulier, nous montrons que notre premier résultat est étroitement lié à ceux de Bloom (1984) et d'Angrist et Imbens (1991). Nous soulignons aussi que notre résultat est étroitement lié à l'idée énoncée par Rosenbaum (1987) d'utiliser deux groupes de comparaison distincts pour l'identification des effets causals. Enfin, nous montrons les similarités entre notre test de spécification d'un estimateur non expérimental et les tests de spécification établis par Heckman et Hotz (1989), ainsi que le lien avec la caractérisation du biais de sélection donnée par Heckman et coll. (1998a).

La suite de l'article est présentée comme suit. À la section 2, nous discutons des similarités entre l'expérience entièrement randomisée et un PEDR. À la section 3, nous généralisons l'utilisation d'un PEDR lorsque la participation au groupe de traitement est déterminée par autosélection. À la section 4, nous montrons comment valider l'utilisation des estimateurs non expérimentaux pour estimer l'effet du traitement en utilisant un PEDR. Enfin, à la section 5, nous présentons certaines conclusions.

2. PLAN EXPÉRIMENTAL DE DISCONTINUITÉ DE LA RÉGRESSION ET EXPÉRIENCES RANDOMISÉES

Nous soulignons ici les similarités entre une expérience entièrement randomisée et un PEDR. Suivant la notation de la méthode d'inférence causale fondée sur les résultats possibles (voir Rubin, 1974), soit (Y^T,Y^{NT}) les résultats possibles de la participation et de la non-participation au programme, respectivement. Nous définissons alors l'effet causal du traitement comme étant la différence entre ces deux résultats, $\beta=Y^T-Y^{NT}$, différence qui n'est pas observable, puisque le fait d'être exposé (de se voir refuser l'accès) au programme révèle Y^T (Y^{NT}), mais masque l'autre résultat possible.

Soit D la variable binaire de situation de traitement, où D=1 pour les participants et D=0 pour les non-participants. Si l'affectation est déterminée aléatoirement, la situation de traitement ne dépend pas des caractéristiques de l'individu et la condition suivante est vérifiée

$$(Y^T, Y^{NT}) \perp D. \tag{1}$$

Habituellement, on n'applique la randomisation qu'aux personnes ayant fait antérieurement la demande d'inscription à un programme particulier, personnes qui ne sont généralement pas représentatives de l'ensemble de la population (comme pour le JTPA mentionné plus haut). Dans cette situation, la condition (1) est vérifiée par rapport au groupe d'unités réellement randomisées, mais non pas par rapport à l'ensemble de la population.

L'attrait de la randomisation est que la différence entre le résultat moyen pour les unités traitées et le résultat moyen pour les unités témoins représente l'effet moyen du programme.

$$E(Y^T - Y^{NT}) = E(Y^T \mid D = 1) - E(Y^{NT} \mid D = 0), \tag{2}$$

puisqu'imposer les valeurs de D comme conditions dans le deuxième membre de (2) est sans aucune importance par construction. Autrement dit, la randomisation permet d'utiliser l'information sur les non-participants pour déterminer le résultat contrefactuel moyen pour les participants, c'est-à-dire ce que les participants auraient vécu s'ils n'avaient pas participé au programme.

Bien que le PEDR ne comprenne pas l'affectation aléatoire des unités au groupe de traitement, il a certaines caractéristiques intéressantes en commun avec un plan expérimental. À la présente section, nous nous concentrons sur le PEDR dit *précis* (Trochim, 1984), c'est-à-dire un PEDR où la situation de participation est déterminée d'après la fonction déterministe de la caractéristique observable S qui suit

$$D = 1(S \geq \overline{s}), \tag{3}$$

où $\overline{s}$ est un seuil de sélection. Il s'agit en fait de la formulation originale de ce genre de plan expérimental dont a discuté Campbell (1969). Les unités sont affectées au groupe d'exposition au traitement si, et uniquement si, elles obtiennent un score égal ou supérieur à $\overline{s}$, ce qui sous-entend que la probabilité de participation sous la condition de S est discontinue au seuil $\overline{s}$, faisant un pas de 0 à 1 lorsque S passe le seuil $\overline{s}$. En tirant parti de la relation entre S et D dans (3), il s'ensuit que, pour un PEDR précis, la condition suivante est vérifiée

$$(Y^T, Y^{NT}) \perp D \mid S = \overline{s}\ .$$

À cause de la similarité avec la condition (1), on qualifie souvent un PEDR précis comme étant plan quasi expérimental (Cook et Campbell, 1979). Dans ce contexte, imposer S comme condition permet de déterminer l'effet moyen du programme sur les individus ayant un score $\overline{s}$, donc une version locale du paramètre en (2). En fait, dans le voisinage de $\overline{s}$, ce plan expérimental présente les mêmes caractéristiques qu'une expérience randomisée pure, puisque, pour toute valeur positive de ε, la condition qui suit est approximativement vérifiée

$$(Y^T, Y^{NT}) \mid S = \overline{s} - \varepsilon \quad \approx \quad (Y^T, Y^{NT}) \mid S = \overline{s} + \varepsilon.$$

Notons que, pour définir de façon valable les unités marginales (par rapport à $\overline{s}$), il faut que S soit continu. Dans le cas d'un échantillon fini, pour que la condition soit vérifiée, la valeur de ε doit tendre vers zéro à une vitesse appropriée à mesure que la taille de l'échantillon tend vers l'infini, ce qui sous-entend une théorie asymptotique non standard pour l'estimateur résultant de l'effet moyen (voir Hahn et coll., 2001 et Porter, 2002).

Dans certains cas, les unités ne se conforment pas à la situation qui leur est attribuée, et quittent le programme ou en recherchent d'autres. N'importe laquelle de ces violations de l'affectation originale pourrait mener à des conclusions biaisées sur les effets du programme, puisque les conditions (1) ou (4) ne sont plus valides. En fait, la présence d'unités qui n'observent pas l'affectation lorsque le mécanisme d'affectation est donné par (3) rend de nouveau la probabilité de participation discontinue au seuil $\overline{s}$, mais la situation de traitement n'est plus une fonction déterministe de la variable S (voir Hahn et coll., 2001, et Battistin et Rettore, 2002).

Deux inconvénients importants restreignent l'applicabilité des PEDR. Premièrement, lors d'une étude d'observation, il est plus fréquent que les unités soient affectées au programme de traitement par autosélection que par sélection exogène d'après une mesure faite avant le programme. Deuxièmement, même dans les cas où le PEDR s'applique, un tel plan ne renseigne pas sur l'effet qu'à le traitement sur les unités éloignées de $\overline{s}$. Ces deux problèmes sont ceux que nous examinons aux sections qui suivent.

3. UNE GÉNÉRALISATION DU PLAN EXPÉRIMENTAL PRÉCIS DE LA DISCONTINUITÉ DE LA RÉGRESSION

3.1 Résultats de la détermination de l'effet du programme

Nous discutons ici de l'utilisation des propriétés d'un PEDR dans des conditions fréquemment observées lors de l'évaluation des politiques sociales. Supposons que le programme vise un groupe particulier d'individus dont l'admissibilité dépend d'une caractéristique connue (comme l'âge, le revenu ou la durée du chômage) et que, sous la condition de l'admissibilité, les individus choisissent de participer au programme selon un processus inconnu de

l'analyste. Sans perte de généralité, supposons que les individus admissibles sont ceux pour lesquels les valeurs de la variable S sont supérieurs à un seuil donné $\bar{s}$. Conséquemment, la situation d'admissibilité est déterminée d'après une règle déterministe. Si toutes les unités admissibles participaient au programme, on obtiendrait un PEDR précis et on pourrait déterminer l'effet moyen sur les unités dans le voisinage de $\bar{s}$.

En fait, il est généralement prouvé que les unités admissibles ne participent pas toutes au programme auquel elles ont droit. Dans plusieurs cas, l'hétérogénéité de l'information disponible sur le programme, des préférences individuelles et des coûts d'option sont des facteurs susceptibles d'influer sur la participation. Dans ces conditions, la probabilité de participation varie encore de façon discontinue comme une fonction de S, mais elle n'est plus une fonction déterministe de cette variable. En fait, la probabilité de participation est nulle pour les individus qui ne sont pas admissibles au programme ($S< \bar{s}$) et elle est inférieure à l'unité pour ceux qui, au contraire, y sont admissibles ($S\geq \bar{s}$). Le plan expérimental résultant appartient au PEDR dit *flou* (Trochim, 1984).

À cause à la fois de la règle d'admissibilité et du processus menant à la participation, la population se retrouve divisée en trois sous-groupes : les *individus non-admissibles*, les *non-participants admissibles* et les *participants*. Pour étiqueter ces sous-groupes, nous introduisons une autre variable binaire permettant de distinguer, parmi les individus qui sont admissibles au traitement, ceux qui le reçoivent effectivement. Soit Z la situation d'admissibilité au programme, où Z=1 (Z=0) pour les individus qui sont admissibles (non admissibles) au programme, et soit D la variable binaire indiquant la situation de traitement, telle que définie plus haut. Le groupe de non-participants comprend un mélange d'individus qui ne satisfont pas les critères d'admissibilité (Z=0) et d'individus qui choisissent de ne pas adhérer au programme (Z=1, D=0). Il faut souligner que, contrairement au cas considéré à la section précédente, dans les conditions décrites ici, la participation des individus admissibles au programme n'a pas lieu en vertu du plan expérimental, mais résulte d'une autosélection.

Soit

$$E(Y^T \mid Z=1, D=1, S=s) \tag{4}$$

le résultat moyen pour les unités admissibles obtenant le score S=s et recevant effectivement le traitement, où $S\geq \bar{s}$. Cette quantité est déterminée en exploitant l'information sur les participants pour toute valeur donnée de S. Soit

$$E(Y^{NT} \mid Z=1, D=1, S=s) \tag{5}$$

le résultat contrefactuel moyen pour le même groupe d'unités, c'est-à-dire ce que leur réponse aurait été si elles n'avaient pas participé au programme. L'effet moyen du programme sur les unités traitées dont le score est S=s est alors déterminé comme étant la différence entre les résultats factuel et contrefactuel donnés par (4) et (5), soit

$$\tau(s) = E(Y^T \mid Z=1, D=1, S=s) - E(Y^{NT} \mid Z=1, D=1, S=s).$$

Par conséquent, on obtient l'effet moyen sur les participants, τ, sous forme de moyenne pondérée de ces quantités, les poids étant donnés par la proportion d'unités admissibles dont le score est S=s.

Ni $\tau(s)$ ni τ ne peuvent être déterminés directement, puisque, par construction, le résultat contrefactuel moyen donné par (5) n'est pas observé. Nous ne pouvons pas non plus le remplacer par le résultat factuel moyen observé pour les non-participants admissibles. En fait, à cause du processus d'autosélection déterminant le groupe de participants (c'est-à-dire ceux pour lesquels Z=1 et D=1) et le groupe de non-participants (c'est-à-dire ceux pour lesquels E=1 et D=0), les non-participants admissibles ne représentent pas un échantillon aléatoire provenant du groupe d'unités admissibles, ce qui implique en général que

$$E(Y^{NT} \mid Z=1, D=0, S=s) \tag{6}$$

diffère de (5). Notons que ce résultat est vérifié pour toute valeur donnée de S, en particulier quand $S= \bar{s}$.

Supposons que l'analyste dispose de l'information sur les résultats pour les unités non admissibles (Z=0). Puisque ce groupe d'unités est, par construction, caractérisé par des valeurs de S inférieures au seuil de sélection $\bar{s}$, on ne peut l'utiliser pour proximer les résultats contrefactuels des participants. Nous ne pouvons pas non plus utiliser les unités non admissibles dans le voisinage de $\bar{s}$ pour approximer le résultat contrefactuel moyen des unités participantes dans le voisinage de $\bar{s}$. La quantité

$$E(Y^{NT} \mid Z = 0, S = \bar{s}) \tag{7}$$

est, en fait, différente du résultat contrefactuel (5) évalué à $\bar{s}$, à cause de la sélection non aléatoire des unités dans le groupe exposé au traitement discuté plus haut. Les unités non admissibles ne permettent pas, à elles seules, de résoudre le problème.

Il faut, pour le résoudre (du moins pour une sous-population particulière de participants), utiliser conjointement l'information sur les unités non admissibles *et* sur les non-participants admissibles. La relation clé permettant d'obtenir ce résultat est

$$E(Y^{NT} \mid Z = 1, S = \bar{s}) = E(Y^{NT} \mid Z = 0, S = \bar{s}), \tag{8}$$

qui découle directement de la règle d'admissibilité. Dans le voisinage du seuil de sélection $\bar{s}$, les unités admissibles et non admissibles sont presque identiques en ce qui concerne S, si bien que, dans le scénario contrefactuel, les deux groupes marginaux auraient connu le même résultat moyen. Ce résultat repose sur le PEDR précis, tel que nous l'avons examiné à la section précédente. Le premier membre de l'équation (8) peut s'écrire sous forme d'une moyenne pondérée des résultats connus par les participants admissibles et les non-participants admissibles, respectivement, dans un voisinage de $\bar{s}$; soit

$$E(Y^{NT} \mid Z = 1, D = 1, S = \bar{s})\phi(\bar{s}) + E(Y^{NT} \mid Z = 1, D = 0, S = \bar{s})(1 - \phi(\bar{s})),$$

où $\phi(\bar{s})=\Pr(D=1|Z=1,S=\bar{s})$ est la probabilité d'autosélection dans le programme pour les unités marginalement admissibles. En introduisant la dernière expression par substitution dans (8), nous obtenons

$$E(Y^{NT} \mid Z = 1, D = 1, S = \bar{s}) = \frac{E(Y^{NT} \mid Z = 0, S = \bar{s})}{\phi(\bar{s})} - \frac{E(Y^{NT} \mid Z = 1, D = 0, S = \bar{s})(1 - \phi(\bar{s}))}{\phi(\bar{s})}. \tag{9}$$

Nommément, le résultat contrefactuel moyen pour les participants présentant $S=\bar{s}$ est une combinaison linéaire du résultat factuel moyen pour les unités non admissibles à $S=\bar{s}$ et du résultat factuel moyen pour les non-participants admissibles à $\bar{s}$. Les coefficients de la combinaison linéaire sont tels que leur somme est égale à un et qu'ils sont fonction de la probabilité $\phi(\bar{s})$, qui est déterminable. Donc, $\tau(\bar{s})$, c'est-à-dire l'effet moyen sur les participants au seuil $\bar{s}$, est déterminable et peut s'exprimer sous la forme

$$\frac{E(Y \mid Z = 1, S = \bar{s}) - E(Y \mid Z = 0, S = \bar{s})}{\phi(\bar{s})},$$

en soustrayant (9) de (4). Nous pouvons interpréter la dernière expression comme étant le ratio entre l'effet de l'intention d'être exposé au traitement, c'est-à-dire l'effet moyen que nous observerions si toutes les unités admissibles participaient effectivement au programme, et l'effet moyen de Z sur D au seuil $\bar{s}$. Les résultats obtenus par Hahn et coll. (2001) et par Porter (2002) pour l'estimation non paramétrique dans un PEDR s'appliquent sans difficulté.

Notons que la condition (8) est la pierre angulaire sur laquelle nous construisons le résultat. Autrement dit, il est essentiel que la règle d'admissibilité soit déterminée conformément à un PEDR *flou*. La conséquence principale est que, bien que le PEDR décrit à la présente section soit *flou*, les conditions de régularité pour la détermination de

$\tau(\bar{s})$ sont celles requises sous un PEDR *précis* (et sont par conséquent *plus faibles*). De surcroît, pour obtenir les résultats, nous n'avons pas besoin de préciser comment les unités admissibles s'autosélectionnent dans le programme de traitement. Donc, la capacité de déterminer $\tau(\bar{s})$ ne nécessite aucune hypothèse de comportement relative au processus de sélection. Cependant, pour procéder à la détermination, nous devons avoir accès à des renseignements pour les trois groupes distincts d'unités, c'est-à-dire les participants, les non-participants admissibles et les non-admissibles.

3.2 Résultats connexes

Dans une expérience entièrement randomisée, Bloom (1984) traite le cas où certaines unités affectées au programme n'y participent pas en réalité (désistements). En exploitant l'information sur les participants, les non-participants admissibles et les individus non admissibles, l'auteur prouve que l'on peut déterminer l'effet moyen sur les participants. Le résultat présenté à la section précédente peut être considéré comme un cas spécial du présent résultat, où la randomisation aurait lieu au seuil d'admissibilité $\bar{s}$. Dans notre cas, les non-participants admissibles au seuil $\bar{s}$ jouent le rôle des désistements de Bloom (1984).

Notre résultat (ainsi que celui de Bloom) peut aussi être obtenu à titre de cas spécial d'Angrist et Imbens (1991). Les auteurs prouvent que, même si la participation est le résultat d'une autosélection, l'effet moyen sur les participants est déterminable à condition i) qu'il existe une variable aléatoire A affectant la participation au programme et orthogonale aux résultats cibles (Y^T, Y^{NT}) et ii) que la probabilité de participation sous la condition de A est nulle pour au moins une valeur de A. La condition (i) fait de A une variable instrumentale pour la situation de traitement.

Dans le contexte de Bloom (1984), l'autosélection est une conséquence du comportement d'inobservation de certaines unités affectées aléatoirement au programme. Dans ces conditions, le choix naturel pour A est la situation prescrite, puisqu'elle résulte de la randomisation. La condition (i) est satisfaite, puisque $Pr(D=1|A=1)>Pr(D=1|A=0)$ et que A est orthogonale aux résultats possibles, et la condition (ii) est satisfaite puisque $Pr(D=1|A=0)=0$. Dans notre cas, puisque Z est orthogonale aux résultats possibles dans le voisinage de $\bar{s}$ et que $Pr(D=1|Z=0)=0$, Z satisfait les conditions énoncées par Angrist et Imbens (1991) dans le voisinage de $\bar{s}$. Donc, la détermination de l'effet moyen sur les participants au seuil $\bar{s}$ en découle.

4. VALIDATION DES ESTIMATEURS NON EXPÉRIMENTAUX DE L'EFFET MOYEN SUR LES PARTICIPANTS

4.1 Tests de spécification

À la section précédente, nous avons montré que l'existence d'une règle d'admissibilité permet de déterminer l'effet moyen d'une intervention sur des participants marginalement admissibles, même si les participants sont autosélectionnés à partir du groupe d'individus admissibles. Si l'effet du traitement est hétérogène par rapport à S, l'effet sur les participants marginaux ne fournit pas d'information sur l'effet de l'intervention sur les unités éloignées du seuil d'admissibilité. On peut utiliser les unités non admissibles et les non-participants admissibles comme groupes de comparaison valides, puisqu'ils diffèrent systématiquement des participants (les premiers par rapport à S et les seconds par rapport aux variables déterminant l'autosélection).

Afin de déterminer l'effet moyen sur l'ensemble de la population de participants, il faut recourir à l'une des longues séries d'estimateurs non expérimentaux décrits dans la littérature qui comportent une correction pour le biais de sélection sous diverses hypothèses (voir Heckman et coll., 1999 et Blundell et Costa Dias, 2000 pour une revue). L'inconvénient principal de cette approche est que les estimateurs de rechange pour le paramètre d'intérêt sont convergents sous des hypothèses qui, la plupart du temps, ne sont pas vérifiables.

Au fil des ans, selon la littérature, deux grandes approches se sont profilées pour traiter le problème. La première revient à déterminer si une théorie du comportement du phénomène étudié engendre une contrainte de suridentification du processus de génération des données et à exploiter éventuellement ce genre de contrainte pour

tester les hypothèses sur lesquelles s'appuient les estimateurs non expérimentaux (voir Rosenbaum, 1984 et Heckman et Hotz, 1989).

La deuxième approche n'est possible que si l'on a exécuté un plan expérimental, de sorte que l'on dispose d'une estimation expérimentale de l'effet. Alors, en plus d'estimer l'effet moyen, on peut exploiter les conditions expérimentales pour étudier le biais de sélection et pour déterminer si les estimateurs non expérimentaux peuvent reproduire l'estimation expérimentale (voir LaLonde, 1986 et Heckman et coll., 1998a). Lorsqu'on dispose des données d'une expérience randomisée, on peut vérifier valablement dans quelle mesure les estimations obtenues par les méthodes non expérimentales des groupes de comparaison s'approchent des estimations expérimentales de l'effet. Parallèlement, cela nous permet d'évaluer les propriétés d'autres estimateurs non expérimentaux de l'effet du traitement, donc de proposer la meilleure stratégie à adopter lorsqu'on ne dispose pas de données expérimentales.

À la présente section, nous montrons que si l'on dispose de renseignements sur les trois groupes d'unités résultant du plan établi à la section 3.1, on peut alors tester la validité de tout estimateur non expérimental sur une sous-population spécifique. À titre d'illustration, nous nous concentrons sur l'estimateur par appariement bien connu, mais le même raisonnement s'applique à d'autres estimateurs non expérimentaux. L'hypothèse fondamentale sur laquelle s'appuie l'estimateur par appariement est celle selon laquelle l'analyste peut observer toutes les variables qui sous-tendent le processus d'autosélection *et* sont corrélées au résultat. Formellement, l'affectation au groupe exposé au traitement est dite *fortement ignorable* étant donné un ensemble de caractéristiques x si, sous la condition de x, on peut considérer le traitement comme étant affecté aléatoirement aux unités, à condition que, pour chaque valeur de x il existe une probabilité positive de subir le traitement

$$(Y^T, Y^{NT}) \perp D \mid x, \quad 0 < Pr(D = 1 \mid x) < 1. \tag{10}$$

Si cette condition est vérifiée, la situation équivaut à ce que les unités soient affectées aléatoirement à l'exposition au traitement avec une probabilité qui dépend de x; on peut utiliser comme approximation du résultat contrefactuel pour les participants présentant les caractéristiques x le résultat réel pour les non-participants présentant les mêmes caractéristiques. Puisque les unités présentant les caractéristiques x ont la même probabilité d'entrer dans le programme, alors une règle opérationnelle en vue d'obtenir des données de genre expérimental *ex post* consiste à apparier les participants aux non-participants en se fondant sur une telle probabilité (le prétendu *score de propension*), dont la dimension est invariable par rapport à la dimension de x (voir Rosenbaum et Rubin, 1983).

Dans cette méthode, l'hypothèse critique est que l'ensemble x est suffisamment riche pour garantir la condition d'orthogonalité dans (10). En principe, ceci impose des exigences fortes sur la collecte des données. En outre, la violation de la deuxième condition dans (10) soulèverait le problème dit de soutien commun (voir, par exemple, Heckman et coll., 1998a, et Lechner, 2001).
Soit

$$sb(s) = E(Y^{NT} \mid E = 1, D = 1, S = s) - E(Y^{NT} \mid E = 1, D = 0, S = s) \tag{11}$$

le *biais de sélection* qui influe sur la comparaison grossière des participants admissibles et des non-participants admissibles. Le premier terme du deuxième membre est un résultat contrefactuel moyen, tandis que le second est un résultat factuel moyen. Cette quantité reflète les différences avant l'intervention entre les unités admissibles autosélectionnées dans le programme d'intervention et hors de celui-ci, respectivement, à chaque niveau de S, pour $S \geq \overline{s}$.

En utilisant les résultats de la section précédente, nous pouvons déterminer le résultat contrefactuel moyen pour les participants dans le voisinage du seuil $\overline{s}$ au moyen de (9). Cela sous-entend aussi que nous pouvons déterminer le biais de sélection pour les unités marginalement admissibles, sb($\overline{s}$), comme étant la différence entre (9) et (6) évalué au seuil $\overline{s}$. Notons que la détermination du terme contrefactuel du deuxième membre de (11) au seuil $\overline{s}$ s'appuie sur l'information sur le sous-groupe d'unités non admissibles les plus proches du groupe d'unités admissibles. Apparemment, à mesure que S s'écarte de $\overline{s}$, l'identification n'est plus possible.

Alors, soit

$$sb(s,x) = E(Y^{NT} \mid E=1, D=1, x, S=s) - E(Y^{NT} \mid E=1, D=0, x, S=s)$$

le biais de sélection sur la sous-population particulière marquée de l'indice x, où x représente les variables utilisées pour expliquer le biais de sélection dans une estimation par appariement de l'effet de l'intervention. Si la condition d'orthogonalité incluse dans (10) est vérifiée, alors sb(s,x)=0 uniformément par rapport à x et à S. En particulier, une condition nécessaire pour que l'estimateur par appariement donne de bons résultats est que sb($\bar{s}$,x)=0, ce qui est directement vérifiable.

Du point de vue opérationnel, dans le voisinage de $\bar{s}$, tout test de l'égalité des résultats moyens des unités non admissibles et des non-participants admissibles, respectivement, sous la condition de x, est un test de la forte ignorabilité de l'affectation à l'intervention, donc un test de la validité de l'estimateur par appariement. De toute évidence, le rejet de l'hypothèse nulle est suffisant pour conclure que la condition (10) n'est pas vérifiée.

Par ailleurs, en acceptant l'hypothèse nulle, on pourrait utiliser avec plus de confiance l'estimateur par appariement. Cependant, on ne peut pas dire que la validité de l'estimateur a été prouvée : en fait, le test est non informatif en ce qui concerne la question de savoir si la condition d'ignorabilité est vérifiée lorsque l'on s'écarte du seuil $\bar{s}$.

4.2 Résultats connexes

Puisque le PEDR peut être considéré comme l'expérience formelle au seuil $\bar{s}$, le test de spécification élaboré plus haut présente une similarité avec le test développé par Heckman et coll. (1998a) dans des conditions expérimentales. Dans les deux cas, on utilise une estimation de référence de l'effet moyen de l'intervention — l'estimation du PEDR dans le premier et l'estimation expérimentale dans le second — que l'analyste est disposé à considérer comme crédible. Alors, l'analyste teste les estimateurs non expérimentaux comparativement à l'estimation de référence pour déterminer si les hypothèses sur lesquelles ils s'appuient sont satisfaites.

La similarité entre les deux approches s'arrête là. L'existence de conditions expérimentales, comme dans Heckman et coll. (1998a), permet de caractériser complètement le biais de sélection et de tester les estimateurs non expérimentaux par rapport à la population de participants. Si un PEDR est disponible, ces exercices ne sont possibles que pour la population de participants au seuil $\bar{s}$.

Cependant, l'existence de données expérimentales est rare dans les évaluations des politiques sociales, particulièrement dans les pays européens. Par contre, il arrive souvent qu'une politique vise une population d'individus admissibles dont la participation au programme est volontaire. Dans cette situation, l'information sur les trois groupes d'individus requise pour appliquer les résultats décrits dans le présent article est, en principe, disponible. Il serait alors possible d'envisager l'application systématique du test de spécification fondé sur le PEDR comme outil de validation des estimateurs non expérimentaux de l'effet moyen sur les participants.

Dans sa discussion du rôle d'un deuxième groupe témoin dans une étude observationnelle, Rosenbaum (1987) donne un exemple qui ressemble, quoique vaguement, aux conditions auxquelles nous nous référons (exemple 2 à la page 294). L'Advanced Placement Program (programme AP) donne aux élèves du secondaire la possibilité d'accumuler des unités de cours collégiaux pour des travaux faits au secondaire. Les écoles secondaires n'offrent pas toutes le programme AP et, parmi celles qui le font, seule une faible majorité d'élèves participent. Deux groupes de comparaison se définissent naturellement dans ce contexte : les élèves inscrits dans une école secondaire n'offrant pas le programme et ceux inscrits dans une école offrant le programme mais qui n'y participent pas.

Puis, Rosenbaum (1987) discute de la façon dont on peut exploiter l'existence des deux groupes de comparaison pour tester la condition de forte ignorabilité nécessaire pour avoir confiance dans les résultats d'un estimateur par appariement. À première vue, son premier groupe de comparaison ressemble à notre groupe d'individus inadmissibles, tandis que le deuxième ressemble à notre groupe de non-participants admissibles. La différence essentielle entre l'exemple de Rosenbaum et notre plan est que, dans son exemple, l'analyste ne connaît pas la règle selon laquelle les écoles secondaires décident d'offrir ou non le programme AP, tandis que dans notre plan, la règle d'admissibilité est précisée complètement. C'est précisément cette caractéristique qui permet d'identifier l'effet

moyen sur les participants au seuil $\bar{s}$ et de tester la validité de tout autre estimateur non expérimental, même si seulement au seuil $\bar{s}$.

5. CONCLUSIONS

Le message principal qui se dégage du présent article est que, si une intervention vise une population d'unités admissibles, mais qu'elle est effectivement appliquée à un sous-ensemble d'unités admissibles autosélectionnées, cela vaut la peine de recueillir des renseignements séparément sur trois groupes d'unités : les individus non admissibles, les non-participants admissibles et les participants. En outre, il faut enregistrer les variables en fonction desquelles est établie l'admissibilité.

D'autres auteurs ont déjà insisté sur le fait qu'il est pertinent de faire la distinction entre les individus non admissibles et les non-participants admissibles pour améliorer la comparabilité du groupe exposé au traitement et du groupe témoin (voir, entre autres, Heckman et coll., 1998a). Nous avons montré que, si la règle d'admissibilité se fonde sur une variable continue et que l'observation est faite à la fois pour les individus non admissibles et pour les non-participants admissibles, on peut déterminer l'effet moyen sur les participants qui sont marginalement admissibles au programme, quelle que soit la façon dont se produit l'autosélection des participants. Nous avons aussi montré que le plan d'expérience résultant concorde avec un PEDR *flou*, mais que les discontinuités dans la règle d'admissibilité mènent à des conditions de régularité pour la détermination de l'effet qui sont caractéristiques d'un PEDR précis. Enfin, nous avons montré qu'une conséquence directe du résultat précédent est que le biais de sélection pour les unités à la limite entre l'admissibilité et la non-admissibilité est déterminable. Ces résultats suggèrent l'utilisation d'un test de spécification dans le voisinage du seuil d'admissibilité afin de pouvoir évaluer les propriétés des estimateurs non expérimentaux. Par conception, un test de ce genre nous renseigne sur les propriétés des estimateurs non expérimentaux que pour un sous-groupe particulier d'unités, donc les résultats ne peuvent être généralisés à l'ensemble de la population (à moins que nous soyons disposés à imposer d'autres contraintes d'identification). La valeur du test de spécification est que, s'il mène au rejet de l'estimateur localement, alors il est suffisant pour indiquer son rejet dans l'ensemble.

REMERCIEMENTS

La rédaction du présent article a bénéficié de discussions fructueuses avec David Card, Hide Ichimura et Andrea Ichino et de commentaires formulés par les participants à l'ESEM 2002, à l'atelier CEPR/IZA ayant pour thème l'amélioration du rendement du marché du travail qui a eu lieu à Bonn en octobre 2002, au Symposium 2002 de Statistique Canada et à la conférence LABORatorio ayant pour thème les nouvelles perspectives en évaluation des politiques publiques qui a eu lieu à Turin en novembre 2002. Les auteurs sont reconnaissants au MIUR d'avoir appuyé financièrement le projet sur la dynamique et l'inertie sur le marché italien du travail et l'évaluation des politiques (base de données, problèmes de mesure, analyses de fond). Les avis de non-responsabilité habituels s'appliquent.

RÉFÉRENCES

Angrist, J.D (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants", *Econometrica*, 66, 2, pp. 249-288.

Angrist, J.D, et G.W. Imbens, (1991), "Sources of Identifying Information in Evaluation Models", *NBER Technical Working Paper* 117.

Battistin, E., et E. Rettore (2002), "Testing for programme effects in a regression discontinuity design with imperfect compliance", *Journal of the Royal Statistical Society A*, 165, 1, pp. 1-19.

Bloom, H.S. (1984), "Accounting for No-Shows in Experimental Evaluation Designs", *Evaluation Review*, 8, pp. 225-246.

Blundell, R., et M. Costa Dias (2000), "Evaluation methods for non-experimental data", *Fiscal Studies*, 21, 4, pp. 427-468.

Campbell, D.T. (1969), "Reforms as experiment", *The American Psychologist*, 24**,** pp. 409-429.

Cook, T.D., et D.T. Campbell (1979), *Quasi-Experimentation. Design and Analisys Issues for Field Settigs*, Boston: Houghton Mifflin Company.

Friedlander, D., Greenberg, D.H., et P.K. Robins (1997), "Evaluating Government Training Programs for the Economically Disadvantaged", *Journal of Economic Literature*, 35, 4, pp. 1809-1855.

Hahn, J., Todd, P., et W. Van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design", *Econometrica*, 69, 3, pp. 201-209.

Heckman, J.J., et V.J. Hotz (1989), "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, 84, pp. 862-874.

Heckman, J.J., et J. Smith (1995), "Assessing the case for social experiments", *Journal of Economic Perspectives*, 9, 2, pp. 85-110.

Heckman, J.J., Ichimura, H., Smith, J., et P. Todd (1998a), "Characterizing Selection Bias Using Experimental Data", *Econometrica*, 66, pp. 1017-1098.

Heckman, J.J., Smith, J., et C. Taber (1998b), "Accounting for Dropouts in Evaluations of Social Experiments", *The Review of Economics and Statistics*, 80, 1, pp. 1-14.

Heckman, J.J., Lalonde, R., et J. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs", in Ashenfelter, A. and D. Card (eds.) *Handbook of Labor Economics, Volume 3*, Amsterdam: Elsevier Science.

van der Klaauw, W. (2002), "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach", *International Economic Review*, 43, 4, pp.

LaLonde, R. (1986), "Evaluating the econometric evaluations of training programs with experimental data", *American Economic Review*, 76, pp. 604-20.

Lechner, M. (2001), "A note on the common support problem in applied evaluation studies", *Discussion Paper 2001-01*, Department of Economics, University of St. Gallen.

Porter, J. (2002), "Asymptotic bias and optimal convergence rates for semiparametric kernel estimators in the regression discontinuity model", *Discussion Paper 1989*, Harvard Institute of Economic Research.

Rosenbaum, P.R. (1984), "From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment", *Journal of the American Statistical Association*, 79, 385, pp. 41-48.

Rosenbaum, P.R. (1987), "The Role of a Second Control Group in an Observational Study", *Statistical Science*, 2, 3, pp. 292-306.

Rosenbaum, P.R., et D.B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, 70, pp. 41-55.

Rubin, D.B. (1974), “Estimating causal effects of treatments in randomized and nonrandomized studies”, *Journal of Educational Psycology*, 66, pp. 688-701.

Rubin, D.B. (1977), “Assignment to Treatment Group on the Basis of a Covariate”, *Journal of Educational Statistics*, 2, pp. 4-58.

Trochim, W. (1984), *Research Design for Program Evaluation: the Regression-Discontinuity Approach*, Beverly Hills: Sage Publications.