The Evaluation of Public Program Effect Using Regression Discontinuity Method : An introduction

Santarossa, Gino

10 October 2008

Note d'introduction sur l'évaluation d'impact d'un programme public par la méthode de régression par discontinuité

Gino Santarossa[†]

Octobre 2008

Résumé

Cette note décrit brièvement l'approche de régression par discontinuité qui vise à estimer l'impact d'un programme lorsque la participation dépend d'une valeur prédéterminée d'un facteur de sélection (seuil de participation au programme). L'évaluation peut s'appuyer sur une stratégie en deux étapes. La première consiste à visualiser graphiquement la relation entre la variable dépendante et le facteur de sélection afin de tirer des indications sur la nature de cette relation et le type de discontinuité au seuil de participation. La seconde étape s'appuie sur l'analyse économétrique et une analyse de sensibilité des résultats réalisée selon différentes formes fonctionnelles des effets du facteur de sélection et du programme sur la variable dépendante.

[†] Économiste-chercheur non affilié. Tout commentaire sera bien reçu et apprécié.

1 Introduction

Cette note décrit brièvement l'approche de régression par discontinuité qui vise à estimer l'impact d'un programme lorsque la participation dépend d'une valeur prédéterminée d'un facteur de sélection[1] soit le seuil de participation au programme. Au Québec, plusieurs programmes ou mesures publiques sont caractérisés par ce type de participation : le programme de péréquation municipale accorde une aide financière aux municipalités dont la richesse foncière uniformisée par habitant est inférieure à la médiane des richesses foncières de l'ensemble des municipalités québécoises ; le programme des prêts et bourses aux étudiants accorde une aide financière aux études postsecondaires en fonction de cinq paliers du revenu disponible parental[2]; enfin, le régime d'impôt sur le revenu des particuliers prévoit des calculs du taux effectif d'imposition en fonction notamment de trois catégories de revenus des ménages.

La méthode de régression par discontinuité consiste à « comparer » les résultats[3] des participants et des non-participants dont les valeurs du facteur de sélection sont localisées au « voisinage » du seuil de participation. Tout écart, à ce seuil, entre les résultats des participants et des non-participants, ou toute discontinuité de la relation entre la variable dépendante et le facteur de sélection, est alors attribué à l'impact du programme. Cette méthode est largement utilisée dans l'évaluation des programmes régis par des critères de participation avec seuil de participation (Trochim (1982), Hoxby (1998), Angrist et Lavy (1999)).

Cette note est organisée comme suit. La section suivante rappelle le problème général d'évaluation de l'impact d'un programme. La section 3 décrit la méthode de régression par discontinuité à l'aide d'analyses graphiques, d'une présentation mathématique et d'une description sommaire des méthodes d'estimation. La section 4 conclut.

2 Problème d'évaluation de l'impact d'un programme

L'évaluation d'impact d'un programme public peut être représentée à l'aide de l'approche économétrique des changements de régime (Quandt (1958,1972)) ou l'approche statistique des résultats potentiels (Rubin (1974)). Cette approche propose de « comparer » à un instant donné le résultat d'une unité *i* si elle participe au programme (Y_i^1) et son résultat sans une participation (Y_i^0). L'écart entre ces deux résultats est alors attribuable à l'effet du programme. Il s'écrit :

$$\delta_i = Y_i^1 - Y_i^0 \qquad (1)$$

[1] Un facteur de sélection détermine conjointement la participation au programme et la variable dépendante.
[2] Les modalités du programme sont décrites sur le site suivant sur le site du ministère de l'Éducation, du Loisir et du Sport à l'adresse suivante : www.afe.gouv.qc.ca/fr/pretsBourses/index.asp.
[3] Le terme « résultats » désigne les valeurs de la variable dépendante.

où δ_i symbolise cet effet. Le problème d'évaluation se définit par l'impossibilité d'observer les deux résultats Y_i^1 et Y_i^0 simultanément. L'impact du programme est donc indéterminé pour chaque unité d'une population visée par le programme. Il est toutefois possible de substituer à Y_i^1 et Y_i^0 des valeurs *attendues*[4] et *estimables* afin d'identifier l'effet du programme non plus pour une unité en particulier, mais pour différents regroupements d'unités. Par exemple, il est particulièrement avisé de s'intéresser au groupe des unités participantes à un programme afin d'en évaluer son efficacité. À cet égard, il est utile de définir une variable dichotomique *D* qui a pour rôle de nous renseigner sur la participation d'une unité au programme. Elle se définit comme :

$$D_i = \begin{Bmatrix} 1 \text{ si l'unité } i \text{ participe au programme} \\ 0, \text{ autrement} \end{Bmatrix}. \qquad (2)$$

La valeur attendue de Y_i^1 pour l'ensemble des unités participantes peut être ainsi représentée par le terme $E(Y_i^1 \mid D_i = 1)$[5] tandis que $E(Y_i^0 \mid D_i = 1)$ désigne leur valeur attendue sans une participation. L'impact du programme sur les unités participantes est alors le résultat de la différence entre ces deux termes et s'écrit :

$$E(\delta_i \mid D_i = 1) = E(Y_i^1 \mid D_i = 1) - E(Y_i^0 \mid D_i = 1) \qquad (3)$$

où $E(\delta_i \mid D_i = 1)$ symbolise l'effet du programme sur les participants. Cet effet reste cependant incalculable puisque le résultat attendu des participants s'ils n'avaient pas participé au programme ($E(Y_i^0 \mid D_i = 1)$) s'avère tout simplement inobservable[6]. Pour pallier ce problème de renseignement manquant, il est possible de remplacer $E(Y_i^0 \mid D_i = 1)$ par le résultat attendu de Y_i^0 pour les non-participants ($E(Y_i^0 \mid D_i = 0)$). Cette opération doit toutefois respecter une condition importante : les valeurs attendues de Y_i pour les unités participantes et non participantes doivent être identiques en l'absence du programme. À partir de cette condition, l'équation (3) peut être reformulée comme suit :

$$\begin{aligned} &E(\delta_i \mid D_i = 1) = E(Y_i^1 \mid D_i = 1) - E(Y_i^0 \mid D_i = 0) \\ &\text{à la condition que :} \\ &E(Y_i^0 \mid D_i = 1) = E(Y_i^0 \mid D_i = 0). \end{aligned} \qquad (4)$$

[4] La valeur attendue d'une variable est un concept mathématique et probabiliste qui réfère à la valeur la plus probable de cette variable si elle était mesurée un très grand nombre de fois. Elle peut être estimée par la moyenne des données dont dispose l'analyste sur cette variable. Cette moyenne sera d'autant plus près de la valeur attendue qu'est élevé le nombre d'observation.

[5] Le signe | signifie "à *la condition que"*.

[6] Les écrits sur les effets de traitement nomment ce résultat contrefactuel.

En pratique, la condition qui précède n'est pas simple à respecter et même peu réaliste dans un univers où de nombreux facteurs peuvent différencier naturellement les participants et les non-participants à un programme. Certaines hypothèses minimales permettent néanmoins de s'en approcher. L'une d'elles est l'hypothèse d'indépendance conditionnelle (HIC). Elle stipule que l'effet du programme sur les participants peut être identifié sans biais si les résultats attendus des participants et des non-participants sont comparés à des valeurs identiques de tous les facteurs qui déterminent la sélection au programme[7] auquel cas l'effet du programme sur les participants est donné par :

$$E(\delta_i \mid D_i = 1, Z) = E(Y_i^1 \mid D_i = 1, Z) - E(Y_i^0 \mid D_i = 0, Z)$$
$$\text{à la condition que :} \qquad (5)$$
$$E(Y_i^0 \mid D_i = 1, Z) = E(Y_i^0 \mid D_i = 0, Z) = E(Y_i^0 \mid Z)$$

où Z représente le vecteur de tous les facteurs de sélection. Le terme $E(Y_i^0 \mid Z)$ montre précisément que le résultat attendu en l'absence du programme est invariant à l'égard du statut de participation.

L'hypothèse d'indépendance conditionnelle s'avère utile et nécessaire afin d'estimer l'effet d'un programme dont la participation dépend d'un seuil prédéterminé du facteur de sélection. Elle est aussi très avantageuse puisqu'elle réduit considérablement les risques d'un biais majeur dans les estimations d'impact. En effet, les facteurs de sélection sont parfaitement connus[8] pour les programmes régis par des seuils de participation. C'est le cas notamment des programmes cités précédemment où la médiane des richesses foncières et les différents seuils de revenu parental agissent à titre de règles de participation aux programmes de la péréquation et des prêts et bourses respectivement.

L'équation (5) reste toutefois incomplète afin d'identifier l'incidence d'un programme avec seuil de participation. Bien que les valeurs attendues de Y soient conditionnées sur le(s) facteur(s) de sélection connu(s), les participants et les non-participants à ce type de programme ne partagent pas le même support[9] du facteur de sélection. Autrement dit, il est impossible de constituer un groupe de non-participants qui soit comparable aux participants en regard du facteur de sélection étant donné la règle d'inéquation qui régit la participation au programme. Il est possible néanmoins de modifier le problème d'évaluation décrit plus haut de telle sorte que la comparaison souhaitée puisse être réalisée dans une région « avoisinante » au seuil de participation où les unités participantes et non participantes partagent des valeurs relativement semblables, et à la limite identiques, du facteur de sélection. C'est précisément cette idée qui sous-tend la méthode de régression par discontinuité.

---

[7] Ce problème est aussi connu par l'appellation « *sélection sur les observables* » (Heckman et Hotz (1989)).

[8] La méthode de régression par discontinuité peut également tenir compte des facteurs de sélection inobservables. Cette préoccupation déborde toutefois les objectifs de cette note.

[9] Le support d'une variable représente une région particulière de la distribution de l'ensemble de ses valeurs.

3 Approche de régression par discontinuité

L'idée générale de la méthode de régression par discontinuité (RD) est d'exploiter l'information offerte par la règle de participation à un programme lorsque cette règle conditionne la participation en fonction d'une valeur préfixée d'un facteur de sélection[10] (seuil de participation). Par exemple, la participation au programme de péréquation municipale dépend de la richesse foncière uniformisée par habitant (facteur de sélection) d'une municipalité par rapport à la médiane de ces richesses (seuil de participation). De la même façon, l'admissibilité au programme des prêts et bourses est assujettie à une valeur préfixée (seuil de participation) des revenus parentaux (facteur de sélection). Ces règles de participation peuvent créer une discontinuité dans la relation de la variable dépendante au facteur de sélection et cela, précisément au seuil de participation, si le programme comporte un effet sur la variable dépendante. Or, c'est précisément cette discontinuité que vise à tenir compte la méthode RD afin d'identifier l'effet du programme.

En clair, au seuil de participation du programme, les écarts de la variable dépendante entre les participants et les non-participants doivent être associés à d'autres explications que celui du facteur de sélection, identique par construction[11], pour toutes les unités. Le seul facteur dans ce cas susceptible de déterminer les différences de résultats au seuil de participation est le programme. Autrement dit, la distribution des résultats à ce seuil ne dépend plus du facteur de sélection. Cette variable est distribuée de manière aléatoire et toutes différences statistiquement importantes de cette variable entre les participants et les non-participants doivent provenir des effets du programme.

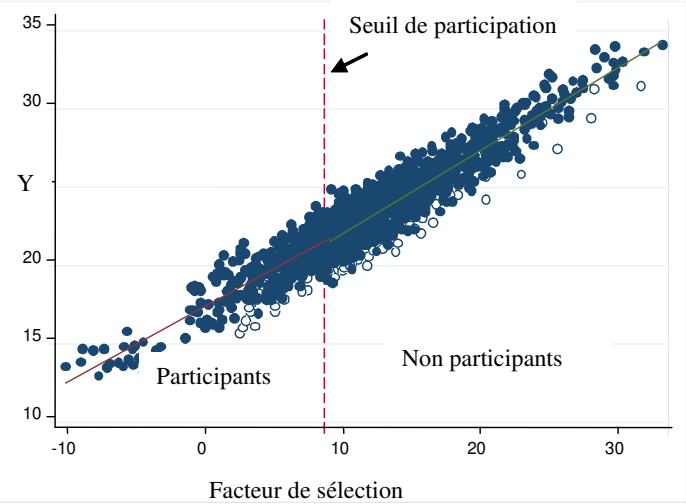
Pour l'instant, cette approche est purement théorique puisque les participants et les non-participants sont différents pour toutes les valeurs du facteur de sélection en l'occurrence celle qui est rattachée au seuil de participation. L'approche RD tire toutefois son épingle du jeu à l'aide du concept de limite mathématique. En bref, elle vise à rendre mathématiquement identiques les participants et les non-participants dans une région infiniment petite autour du seuil de participation au programme. Dans cette région, le facteur de sélection n'a plus d'incidence sur la variable dépendante et tout écart de résultats peut être attribué à l'effet du programme. Bien entendu, le nombre de participants et de non-participants dont les valeurs du facteur de sélection sont localisées dans cette région est tout compte fait nul. La méthode RD considère alors différentes hypothèses afin d'accroître le nombre d'unités participantes et non-participantes selon des voisinages plus réalistes du seuil de participation de telle sorte qu'un nombre suffisant d'observations soit utilisé.

Les principes élémentaires de la méthode sont illustrés au graphique 1 où l'on y retrouve les valeurs simulées d'une variable dépendante et d'un facteur de sélection[12]. La

[10] La méthode peut également s'appliquer dans le cas de plus d'un facteur de sélection.
[11] C'est-à-dire en conditionnant sur une seule valeur du facteur de sélection.
[12] Les données simulées sont produites par une équation linéaire et additive où une série de facteurs explicatifs déterminent la variable dépendante Y dont un facteur de sélection et la participation à un programme. Les données sont générées à l'aide d'une fonction de distribution normale selon des moyennes et écarts-types fixés pour chacune des variables.

variable dépendante est représentée par l'ordonnée du graphique et le facteur de sélection, par l'abscisse. Les observations sur les unités non participantes sont localisées à droite du seuil de participation (ligne verticale en trait discontinu) et celles sur les unités participantes, à gauche. La dispersion des points indique que la variable dépendante s'accroît en fonction des hausses progressives du facteur de sélection. En bref, la relation observée est linéaire et dotée d'une pente positive. On n'observe aucune variation brusque ou anormale (discontinuité) de la relation peu importe les valeurs du facteur de sélection, en particulier au seuil de participation, ce qui indique que le programme n'a pas d'incidence « apparente »[13] sur la variable dépendante. Tel est le cas puisque nous avons simulé une incidence nulle du programme pour cette première analyse.

Graphique 1 – Dispersion des données simulées d'une variable dépendante Y et d'un facteur de sélection, effet nul du programme



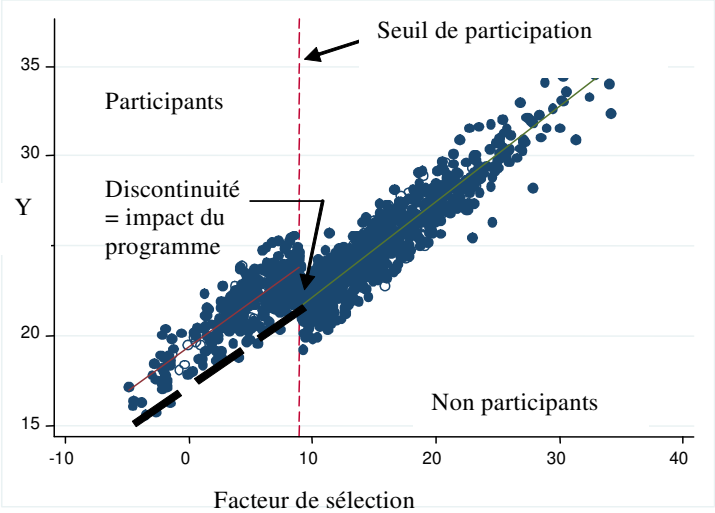
Le graphique 2 illustre la relation entre la variable dépendante et le facteur de sélection si le programme comporte véritablement un effet[14]. Un saut du nuage de points est alors observé au seuil de participation du programme. Puisque cette discontinuité survient à ce seuil très précisément, elle est alors directement associée à l'effet du programme. Cet effet peut être estimé par la différence entre le résultat moyen des participants et celui des non-participants à la « marge » du seuil de participation. Si le nombre d'observations est suffisant, l'effet estimé du programme sera non biaisé puisqu'il s'appuie sur la comparaison des participants et des non-participants à une valeur presque identique du facteur de sélection.

[13] Seule une analyse économétrique peut conclure sur un effet statistiquement significatif d'un programme.

[14] L'effet du programme a été simulé de telle sorte que la variable dépendante est haussée d'une valeur constante de 2,25 pour tous les participants.

En principe, l'impact ainsi estimé est strictement attribuable aux unités localisées très près du seuil de participation[15]. Il est cependant possible de l'extrapoler aux autres unités participantes à la condition que la relation entre la variable dépendante et le facteur de sélection soit identique ou présumée comme tel de manière raisonnable pour l'ensemble des unités peu importe leur statut de participation. Cette condition permet d'une certaine façon d'estimer le résultat moyen des participants s'ils n'avaient pas participé au programme et de le comparer ensuite à leur résultat observé. Cette comparaison est équivalant à l'écart entre la droite de régression des participants (résultat moyen de *Y* pour les participants) et le prolongement de la droite estimée pour les non-participants (trait prolongé discontinu). L'impact est estimé avec beaucoup plus d'efficacité et de précision puisque la totalité des observations est sollicitée dans l'estimation. Par ailleurs, cette approche s'appuie sur l'une des hypothèses (linéarité de la relation de la variable dépendante avec le facteur de sélection) qu'il est possible de considérer si le nombre d'observations au seuil de participation est insuffisant afin d'estimer sans biais l'impact du programme.

Graphique 2 – Dispersion des données simulées d'une variable dépendante Y et d'un facteur de sélection, effet constant et positif du programme

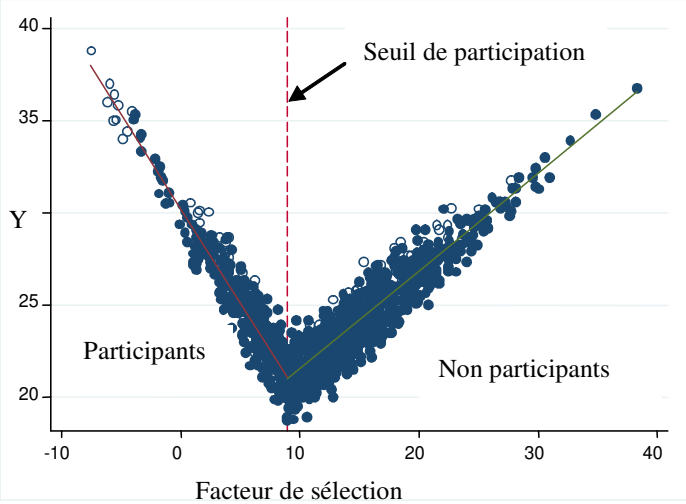


Quelques mises en garde s'imposent toutefois afin d'utiliser judicieusement l'approche RD. La première est associée au type de discontinuité susceptible de se présenter au seuil de participation selon la nature (constante ou variable) de l'effet du programme. Un exemple est celui d'un revirement de la relation entre la variable dépendante et le facteur de sélection au seuil de participation telle qu'illustrée au graphique suivant. La

[15] Les écrits sur les effets de traitement nomment "Effet moyen de traitement local" l'effet estimé pour le groupe de participants dont les valeurs du facteur de sélection sont localisées très près du seuil de participation.

discontinuité est peu importante près du seuil, mais s'intensifie graduellement au fur et à mesure de la réduction des valeurs du facteur de sélection. Ce type de discontinuité peut

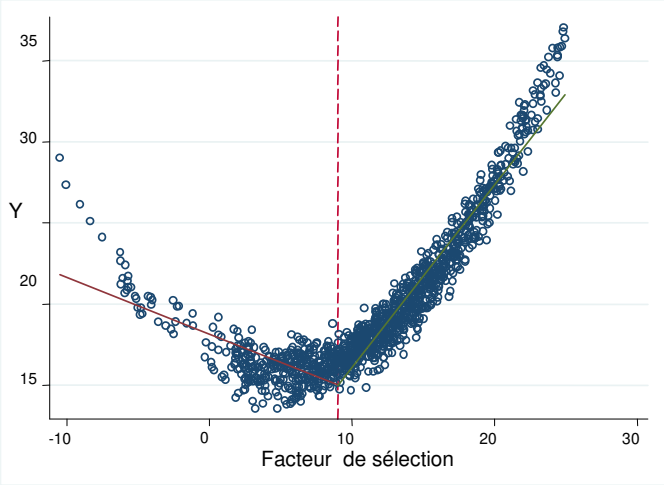
Graphique 3 – Dispersion des données simulées d'une variable dépendante Y et d'un facteur de sélection, effet inversement proportionnel du programme en fonction du facteur de sélection



se produire en outre pour des programmes comme la péréquation ou les prêts et bourses où l'aide financière accordée s'accroît en fonction de la réduction de la richesse foncière ou des revenus parentaux (facteurs de sélection). Il est donc légitime de penser que l'effet attendu du programme sur la variable dépendante augmente également en fonction de l'aide financière reçue par les unités participantes. Si nous tentons d'estimer l'impact du programme qu'à partir des observations localisées en marge du seuil de participation alors nous risquons de n'identifier aucun effet puisque l'aide financière est peu importante pour les unités concernées. Il faut se rabattre alors sur l'ensemble des observations en supposant à nouveau que la relation estimée entre la variable dépendante et le facteur de sélection pour les non-participants s'applique de manière identique aux unités participantes. De plus, il importe de considérer l'effet variable du programme en fonction du facteur de sélection.

Une attention additionnelle doit être portée à la relation qui unit la variable dépendante et le facteur de sélection. Le graphique 4 s'apparente bien au précédent. Toutefois, la simulation ne comporte aucun effet du programme sur la variable dépendante. Pourtant, le graphique montre une relation en forme de « U » entre la variable de sélection et la variable dépendante. Cette relation est générée par un effet non linéaire (quadratique) de la variable de sélection sur la variable dépendante. Une analyse précipitée qui

Graphique 4 – Dispersion des données simulées d'une variable dépendante Y et d'un facteur de sélection, effet quadratique du facteur de sélection sur la variable dépendante et effet nul du programme



présumerait que la variable de sélection détermine linéairement la variable *Y* pour l'ensemble des unités conclurait à tort sur un effet positif du programme. Il s'agit donc d'être prudent sur le type de relation qui unit la variable dépendante et la variable de sélection en considérant d'une part les indications apparentes fournies par le graphique et d'autre part, l'estimation de quelques formes fonctionnelles plausibles de cette relation lors de l'analyse économétrique.

Description formelle de l'approche RD

Hahn, Todd et Klaauw (1999) ont formalisé mathématiquement l'approche de régression par discontinuité. Les étapes de leur démarche sont ici reprises afin de comprendre les principaux rouages de la méthode. Il est d'abord utile de redéfinir la variable indicatrice de participation *D* selon la localisation des valeurs d'un facteur de sélection *Z* autour d'un seuil de participation Z_0. Soit :

$$D_i = \begin{cases} 1 & \text{si } Z_i < Z_0 \\ 0 & \text{si } Z_i >= Z_0 \end{cases} \qquad (6)$$

où l'unité *i* participe au programme ($D_i = 1$) si le facteur de sélection est inférieur au seuil de participation Z_0 et n'y participe pas dans le cas contraire[16]. L'approche RD s'intéresse très précisément aux unités participantes et non participantes dont les valeurs du facteur de sélection sont localisées dans une région infiniment petite à proximité du seuil de participation au programme. Soit e > 0, un nombre arbitraire infiniment petit. La règle de participation de l'équation (6) peut être réécrite afin de redéfinir la participation en fonction d'un voisinage très petit du seuil de participation :

$$D_i = \begin{cases} 1 & \text{si } Z_i \in Z_0 - e \\ 0 & \text{si } Z_i \in Z_0 + e \end{cases} \qquad (7)$$

où $\in$ est le symbole d'appartenance à un ensemble ou un intervalle. Il s'agit ensuite d'évaluer l'effet du programme pour les unités concernées par cette règle de participation. L'équation (5) est donc reformulée en fonction de cette règle lorsque le nombre *e* tend vers zéro. L'effet localisé du programme pour les unités participantes infiniment près du seuil de participation s'écrit :

$$\begin{aligned} &\lim_{e\to 0^-}\left[E(\delta_i \mid Z_0 - e) = E(Y_i^1 \mid Z_0 - e) - E(Y_i^0 \mid Z_0 + e)\right] \\ &\text{à la condition que :} \\ &\lim_{e\to 0^-}\left[E(Y_i^0 \mid Z_0 + e) = E(Y_i^0 \mid Z_0 - e)\right]. \end{aligned} \qquad (8)$$

Hahn et al. (op.cit.) posent cependant deux conditions afin d'estimer l'impact d'un programme à partir de cette expression :

Condition C1 : $E(Y_i^0 \mid Z_i = Z)$ est continue à la valeur Z_0 du facteur de sélection.
Condition C2 : $\lim_{e\to 0^-}\left[E(\delta_i \mid Z_i = Z_0 - e)\right]$ est bien définie.

La condition C1 stipule que la variable dépendante doit être bien définie c'est-à-dire sans discontinuité à la valeur du seuil de participation en l'absence du programme. La deuxième condition stipule que l'impact du programme est calculable à une proximité infiniment petite du seuil de participation.

Si les conditions qui précèdent sont respectées alors l'expression suivante identifie sans biais l'incidence localisée du programme pour les unités à une proximité arbitrairement et infiniment petite du seuil de participation :

$$\lim_{e\to 0^-}\left\{E(Y_i^1 \mid Z_i = Z_0 - e) - E(Y_i^0 \mid Z_i = Z_0 + e)\right\} = E(Y_i^1 \mid Z_0) - E(Y_i^0 \mid Z_0) = \delta_{|Z_0} \qquad (9)$$

où l'indice $| Z_0$ signifie que l'effet du programme $\delta_{|Z_0}$ est strictement déterminé au seuil de participation Z_0. Sans l'hypothèse d'un effet constant, l'impact du programme est identifiable seulement à ce seuil.

[16] L'analyse peut aussi bien s'appliquer dans un ordre inverse des inégalités.

Les méthodes d'estimation

Le paramètre $\delta_{|Z_0}$ de l'équation (9) peut être estimé par la simple différence entre les valeurs moyennes de Y des unités participantes et non participantes dont les valeurs du facteur de sélection sont localisées près du seuil de participation. L'estimateur peut s'écrire :

$$\hat{\delta}_{|Z_0} = \bar{Y}^{1}_{Z_i \in (Z_0 - e)} - \bar{Y}^{0}_{Z_i \in (Z_0 + e)} \qquad (10)$$

où la valeur *e* détermine la proximité au seuil de participation. Le biais et l'efficacité[17] de l'estimateur dépendent respectivement du nombre *e* et du nombre d'observations dans l'intervalle $[Z_0 - e, Z_0 + e]$. Ainsi, plus le nombre *e* est petit moins l'estimateur risque d'être affecté d'un biais important. En revanche, le nombre potentiellement moins élevé d'observations peut compromettre l'efficacité de l'estimateur. Au moment des estimations, l'évaluateur doit donc trouver un juste compromis entre ces deux critères à l'aide d'une analyse de sensibilité des résultats[18].

Cette analyse peut se dérouler en deux étapes. La première peut porter sur la réalisation d'une série d'estimations pour différentes valeurs relativement « petites » du nombre *e*[19]. Cette approche a l'avantage de réduire les risques de biais dans les effets estimés du programme en présumant d'un effet linéaire du facteur de sélection *Z* sur la variable dépendante. L'équation suivante modélise la variable dépendante *Y* selon cet effet linéaire et l'effet du programme décrit à l'équation (10) :

$$Y_i = \mu + X_i\beta + D_i\delta_{|Z_0} + Z_i\gamma + U_i \ . \qquad (11)$$

La variable *D* indique si l'unité *i* participe au programme. Le paramètre $\delta_{|Z_0}$ capte l'incidence moyenne du programme pour les unités participantes concernées c'est-à-dire dont les valeurs du facteur de sélection *Z* appartiennent à l'intervalle $[Z_0 - e, Z_0 + e]$. Le paramètre γ mesure l'effet causal du facteur de sélection *Z* sur la variable dépendante en autant que le vecteur X_i de facteurs explicatif élimine toute corrélation entre le terme d'erreur et le facteur de sélection. En effet, seule suffit l'hypothèse d'indépendance conditionnelle[20] afin d'estimer sans biais l'effet du programme. Par

[17] Le biais représente l'écart moyen attendu entre l'effet estimé et l'effet véritable du programme si les estimations étaient répétées un très grand nombre de fois. L'efficacité de l'estimateur désigne la variabilité moyenne des effets estimés autour de leur espérance mathématique. Généralement, l'efficacité d'un estimateur s'accroît avec le nombre d'observations impliquées dans les estimations.

[18] L'analyse de sensibilité vise à changer certains paramètres de l'estimation d'impact du programme afin d'y observer tout changement important.

[19] Sur le plan des estimations, il n'est plus question bien entendu de valeurs infinitésimales du nombre *e*. Celui-ci doit être fixé de telle sorte qu'un nombre suffisant d'observations soit disponible pour l'inférence statistique.

[20] Dans la littérature sur les effets de traitement, l'hypothèse d'indépendance conditionnelle stipule que la participation à un programme devient aléatoire une fois que l'on tient compte des facteurs de sélection.

conséquent, la présence de X_i n'est pas obligatoire, mais améliore l'efficacité des estimations (Pettersonn-Lidbom (2003)).

L'équation (11) peut être estimée par moindres carrés ordinaires selon les hypothèses habituelles. Les effets estimés du programme sont susceptibles de varier selon les valeurs du nombre *e* ce qui permet d'évaluer la robustesse des résultats. Cet exercice n'est pas simple puisqu'une valeur trop petite de *e* réduit le nombre d'observations sur les unités participantes et non-participantes utilisées dans les estimations et du même coup, accroît la variabilité de ces estimations. En revanche, une valeur élevée de *e,* en dépit d'une meilleure efficacité des estimations, réduit la véracité de l'hypothèse sur la linéarité des effets du facteur de sélection. De plus, elle accroît le risque d'homogénéiser les effets variables du programme selon les valeurs du facteur *Z*[21]. L'évaluateur peut donc s'appuyer sur une analyse graphique des données afin d'obtenir des indications sur la relation qui unit la variable dépendante et le facteur de sélection et le type apparent de discontinuité au seuil de participation. S'il suspecte une variation des effets du programme en fonction de Z, l'équation suivante sera nécessairement plus appropriée et son estimation participera également à l'analyse de sensibilité des résultats :

$$Y_i = \mu + X_i\beta + D_i\delta_{|Z_0} + D_i Z_i \delta'_{|Z_0} + Z_i\gamma + U_i . \qquad (12)$$

La seconde étape dans la stratégie d'estimation de l'effet du programme et de l'évaluation de sa robustesse consiste à exploiter l'ensemble des observations sur les unités participantes et non participantes et à ne plus localiser les estimations à des régions particulières des valeurs du facteur de sélection autour du seuil de participation. Cette approche vise principalement à maximiser l'efficacité des estimations. Les équations (11) et (12) peuvent être réutilisées en autant que l'effet de *Z* sur *Y* soit linéaire pour l'ensemble des unités participantes et non participantes. Le graphique 2 a bien illustré cette relation où la dispersion des observations sur les unités non participantes suggère une relation linéaire entre la variable dépendante et le facteur de sélection. À la lumière de cette indication, il peut s'avérer raisonnable de postuler une relation identique pour les participants au programme[22].

L'hypothèse de linéarité des effets du facteur de sélection pourrait néanmoins s'avérer l'exception plutôt que la règle. En effet, cette hypothèse est souvent privilégiée en vue de simplifier la modélisation des phénomènes socio-économiques sans qu'elle soit la plupart du temps conforme à la réalité. Les graphiques 3 et 4 ont illustré l'une des erreurs susceptibles d'être commises en présumant à tort un effet linéaire de la variable *Z* alors que survient un revirement de cet effet au seuil de participation dans un contexte d'un impact nul du programme. L'évaluateur peut tester quelques formes fonctionnelles de la variable *Z* afin de vérifier la nature de sa relation avec la variable dépendante. Les formes polynomiales d'ordre 2 (quadratique) ou 3 (cubique) peuvent représenter le point de départ de la modélisation des effets potentiellement non linéaire

[21] Voir le graphique 3.
[22] Bien entendu, une analyse graphique en deux dimensions ne peut qu'offrir des indications sur la relation entre la variable dépendante et le facteur de sélection.

du facteur de sélection. Dans le cas d'un effet de *Z* sur *Y* présumé quadratique, l'équation (12) se réécrit :

$$Y_i = \mu + X_i\beta + D_i\delta_{|Z_0} + D_i Z_i \delta'_{|Z_0} + Z_i\gamma_1 + Z_i^2\gamma_2 + U_i \qquad (13)$$

où Z_i^2 est le facteur de sélection élevé au carré et γ_2, l'effet de ce facteur sur la variable *Y*.

4 Conclusion

L'approche de régression par discontinuité est privilégiée lorsque la participation à un programme est établie en fonction d'un seuil de participation c'est-à-dire une valeur prédéterminée d'un facteur qui explique également la variable dépendante. Le but de la méthode consiste à identifier l'effet d'un programme sur la base des discontinuités susceptibles de survenir entre la variable dépendante et le facteur de sélection au seuil de participation du programme. En clair, tout écart entre les résultats moyens des participants et des non-participantes dans un voisinage restreint du seuil de participation est attribuable à l'effet du programme.

Une limite importante de la méthode de régression par discontinuité est que l'effet identifié n'est en théorie applicable qu'aux unités participantes dont les valeurs du facteur de sélection avoisinent le seuil de participation. Cette limite se traduit donc par l'identification d'un effet localisé du programme. Sur le plan de l'inférence statistique, cet effet risque d'être estimé de façon peu efficace en raison d'un nombre insuffisant d'observations au voisinage du seuil de participation. En revanche, l'extension de ce voisinage afin d'accroître le nombre d'observations peut compromettre l'hypothèse d'un effet linéaire du facteur de sélection sur la variable dépendante et ainsi, biaiser les estimations.

Une démarche proposée afin de minimiser les biais et maximiser l'efficacité des estimations consiste d'abord, à l'aide d'analyses graphiques, à recueillir des indications d'une part, sur la relation apparente qui unit la variable dépendante et le facteur de sélection et d'autre part, sur la nature de sa discontinuité au seuil de participation. La modélisation économétrique peut ensuite s'appuyer sur les indications recueillies. Cette modélisation peut être réalisée en deux étapes. La première modélise la variable dépendante en fonction d'un effet linéaire de la variable de sélection. L'estimation de cette équation portera sur les observations au voisinage du seuil de participation. La seconde étape exploite l'ensemble des observations et modélise la variable dépendante en fonction d'effets variables du programme et d'une incidence non linéaire du facteur de sélection. La combinaison de ces deux approches permet d'analyser la sensibilité des résultats aux changements d'hypothèses et du nombre d'observations.

Bibliographie

Angrist, Joshua D. et V. Lavy (1999) “Using maimonides'rule to estimate the effect of class size on scholastic achievement”. *The Quarterly Journal of Economics*, mai 1999, 533-574.

Battistin, Erich et E. Rettore (2003) “Another look at the Regression Discontinuity Design”. *Cahier de recherche*, Institut des études fiscales de Londres et Université de Padova.

Hahn, Jinyong, P. Todd et W. Van Der Klaauw (1999). “Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design.” *NBER* cahier de recherche 7131.

Hahn, Jinyong, P. Todd et W. Van Der Klaauw (2001) “Identification and estimation of treatment effects with a regression-discontinuity design”. *Econometrica* 69 no 1, 201-209.

Heckman, J.J., Hotz, V.J. (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 (408), 862-874.

Hoxby, Caroline M. (1998) “The Effects of Class Size and Composition on Student Achievement : new Evidence From Natural Population Variation” *NBER* cahier de recherche 6869.

Ockert, Björn (2002) “Do university enrollment constraints affect education and earnings ?”. Cahier de recherche 2002:16, Institut d'évaluation des politiques du marché du travail.

Pettersson-Lidbom, Per (2003) “Do Parties Matter for Fiscal Policy Choices ? A Regression-Discontinuity Approach ». *Cahier de recherche*, département d'économie (Université de Stockholm).

Quandt, R.E. (1958). “The estimation of the parameters of a liinear regression system obeying two separate regimes”. Journal of the American Statistical Association 53 (284), 873-880.

Quandt, R.E. (1972). “A new approach to estimating switching regressions”. Journal of the American Statistical Association 67 (338), 306-310.

Rubin, D.B. (1974). “Estimating causal effects of treatments in randomized and nonrandomized studies”. Journal of Educational Psychology 66 (5), 688-701.

Trochim, William M.K. (1982) “Methodologically based discrepancies in compensatory education evaluations” *Evaluation Review* vol 6 no 4, 443-480.

Van Der Klaauw, Wilbert (2001) “Estimating the effect of financial aid offers on college enrolment: a regression-discontinuity approach”. Cahier de recherche, Département d'économie (UNC).