Méthodes économétriques pour l'évaluation de politiques publiques

Pauline Givord

Dans Économie & prévision 2014/1-2 n° 204-205 , pages 1 à 28
Éditions Direction générale du Trésor

ISSN 0249-4744
DOI 10.3917/ecop.204.0002
Date de mise en ligne : 14/01/2016

Article disponible en ligne à l'adresse
https://shs.cairn.info/revue-economie-et-prevision-2014-1-page-1?lang=fr

CAIRN.INFO

Découvrir le sommaire de ce numéro, suivre la revue par email, s'abonner...
Scannez ce QR Code pour accéder à la page de ce numéro sur Cairn.info.

Distribution électronique Cairn.info pour Direction générale du Trésor.
Vous avez l'autorisation de reproduire cet article dans les limites des conditions d'utilisation de Cairn.info ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Détails et conditions sur cairn.info/copyright.
Sauf dispositions légales contraires, les usages numériques à des fins pédagogiques des présentes ressources sont soumises à l'autorisation de l'Éditeur ou, le cas échéant, de l'organisme de gestion collective habilité à cet effet. Il en est ainsi notamment en France avec le CFC qui est l'organisme agréé en la matière.

Méthodes économétriques pour l'évaluation de politiques publiques

Pauline Givord(*)

La demande d'évaluation est de plus en plus présente dans le débat public. Cette évaluation recouvre plusieurs dimensions, la question de l'efficacité des politiques étant essentielle. Il s'agit de déterminer si une mesure a rempli les objectifs qui lui étaient assignés : par exemple, évaluer si une politique d'aide aux retours à l'emploi s'est effectivement traduite par des sorties du chômage plus fréquentes ou si une politique de diminution de la taille des classes a permis d'augmenter le niveau scolaire. De nombreuses méthodes d'évaluation empiriques pour l'évaluation quantitative de l'efficacité des politiques ont été proposées depuis les années 1990. Cet article présente les principales d'entre elles, à destination des économistes souhaitant les mettre en œuvre ou simplement disposer d'une grille de lecture critique pour utiliser à bon escient des études d'évaluation. L'exposé des principes statistiques insiste sur les intuitions et les aspects pratiques. Il est illustré par de nombreux exemples issus de la littérature économique récente.

Ces méthodes empiriques se placent dans le cadre d'analyse dit "de Rubin", qui systématise le recours à un contrefactuel hypothétique pour comparer la situation des bénéficiaires de la mesure que l'on souhaite évaluer avec celle qui aurait prévalu en son absence. Ce contrefactuel n'étant pas observable, il est nécessaire de l'estimer, à partir d'un groupe de contrôle. Celui-ci correspond aux unités qui n'ont pas bénéficié de la politique mais qui sont par ailleurs comparables aux bénéficiaires. En général, la difficulté de l'exercice tient à la présence d'effets de sélection liés au ciblage du programme. Les bénéficiaires sont justement ceux pour qui la mesure est particulièrement utile. Ils peuvent avoir des profils socio-économiques spécifiques, indépendamment de la mise en place de la mesure évaluée. Le risque existe d'attribuer à l'effet de la mesure évaluée ce qui relève de ces particularismes des bénéficiaires.

La méthode la plus directe pour s'affranchir de ces effets de sélection est celle des expérimentations contrôlées. Celles-ci consistent à tirer au hasard les personnes qui vont bénéficier du programme dont on souhaite évaluer l'efficacité et d'autres qui n'en bénéficieront pas. Ces méthodes, issues du champ médical, sont fréquentes dans le champ social en Amérique du Nord depuis les années 1960. Elles constituent le paradigme des méthodes économétriques d'évaluation présentées ici. Toutefois, elles sont encore assez rares en France. De fait, elles nécessitent une intervention très en amont de la mise en place de la politique qu'il s'agit d'évaluer. Souvent, le statisticien est plutôt conduit à estimer ex post *des politiques déjà mises*

(*) Institut National de la Statistique et des Études Économiques (Insee).
E-mail : pauline.givord@insee.fr

Cet article reprend dans ses grandes lignes un cours assuré à l'École Nationale de la Statistique et de l'Administration Économique (Ensae) avec Bruno Crépon entre 2006 et 2015. Il s'est enrichi des questions souvent stimulantes des étudiants qui ont assisté à ce cours, ainsi que des discussions avec les membres du Département des Études Économiques de l'Insee. L'auteure remercie tout particulièrement Magali Beffy, Didier Blanchet, Nila Ceci-Renaud, Laurent Davezies, Xavier d'Haultfœuille et Lionel Wilner, ainsi qu'Hélène Erkel-Rousse et deux relecteurs anonymes, qui lui ont permis par leur lecture attentive et leurs commentaires constructifs d'améliorer significativement les premières versions de cet article. Elle reste seule responsable des erreurs et omissions qu'il pourrait contenir.

Cet article n'engage que son auteure et non les institutions auxquelles elle appartient. Il n'engage *a fortiori* ni la Direction générale du Trésor, ni le ministère des Finances et des Comptes publics, ni le ministère de l'Économie, de l'Industrie et du Numérique.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

en place. De nombreuses autres méthodes empiriques ont été développées pour estimer dans ce contexte l'effet propre d'un programme, qui sont détaillées dans cet article. Ces différentes méthodes sont parfois appelées expériences naturelles ou quasi-expériences, car elles tentent de reproduire le cadre des expériences contrôlées.

Parmi celles-ci, la méthode dite des différences de différences est classique. Il s'agit de comparer non pas les bénéficiaires avec les non bénéficiaires, mais l'évolution de la situation des bénéficiaires avant et après la mise en place du programme avec celle de la situation d'un groupe de non bénéficiaires. Dans ce cas, l'hypothèse est qu'en l'absence du traitement, l'évolution temporelle aurait été la même dans les deux groupes. Même s'il existe des écarts dans la composition des groupes des bénéficiaires et des non bénéficiaires, ces écarts sont supposés constants dans le temps. La différenciation temporelle est supposée permettre d'éliminer ces écarts systématiques.

Une autre méthode classique est celle des variables instrumentales. Elle repose sur l'existence de facteurs qui ont un effet sur la probabilité de bénéficier de la mesure mais pas sur les déterminants inobservés de la variable de résultat à laquelle on s'intéresse. Malheureusement, de tels facteurs existent rarement. En revanche, la présence fréquente de seuils d'éligibilité permet parfois d'utiliser une méthode proche, dite de régressions sur discontinuités. Les personnes qui se situent aux voisinages inférieur et supérieur de ces seuils ont des caractéristiques très similaires mais des probabilités très différentes de bénéficier de la mesure.

Dans de nombreux cas, il n'est pas possible de disposer des sources d'identification exogènes qui sont au fondement des méthodes de différence de différences, de variable instrumentale ou de régression sur discontinuités. On peut alors tenter de corriger a minima *les effets de composition liés au caractéristiques observables, par exemple par des méthodes d'appariement (appelées aussi méthodes de* matching*).*

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

La demande d'évaluation est de plus en plus présente dans le débat public. Quel est l'impact d'un programme de suivi des chômeurs, du classement de certaines zones urbaines en zones franches, d'une réduction de la taille des classes ou de la suppression de la carte scolaire ? Ceci constitue des exemples de questions sur lesquelles le statisticien pourrait être sollicité pour apporter des éléments de réponse. À ces questions, une approche purement "gestionnaire" prévaut souvent. On cherche par exemple à compter le nombre de bénéficiaires d'un dispositif ou à estimer le coût budgétaire d'une mesure. Indispensables, ces indicateurs de suivi ne sont toutefois pas suffisants pour juger l'*efficacité* d'une politique. Le coût social d'une mesure ne se limite pas à son coût budgétaire initial. Une politique de formation des chômeurs, si elle est efficace pour faciliter le retour à l'emploi des chômeurs qui en bénéficient, est susceptible de permettre une réduction du nombre de chômeurs indemnisés. Bien que coûteuse en soi, elle peut donc être en partie autofinancée. Une *étude d'impact* est une étape essentielle pour estimer la rentabilité économique de la mesure à évaluer.

Cette analyse est cependant complexe. Il s'agit de déterminer si les éventuelles améliorations constatées peuvent être attribuées à la mise en place de cette mesure. Une mesure est en général ciblée sur des populations particulières, dont les caractéristiques socio-économiques auraient en tout état de cause été différentes de celles du reste de la population, que la mesure ait été mise en place ou non. Par exemple, les classes à effectif réduit regroupent généralement les enfants les plus en difficulté. Il n'est alors pas surprenant que les résultats moyens obtenus par les élèves de ces classes soient plus faibles que dans la moyenne des autres classes. En conclure qu'augmenter la taille des classes serait efficace pour lutter contre l'échec scolaire serait cependant abusif... Ces effets de sélection, ou de population, rendent l'évaluation difficile.

De nombreuses méthodes économétriques ont été développées pour traiter ces questions depuis les années 1990. Elles se distinguent des usages classiques de l'économétrie appliquée sur au moins deux points : d'une part, par une grande attention aux hypothèses nécessaires pour identifier un effet causal, qui pousse à plus de rigueur dans l'analyse empirique ; d'autre part, par la prise en compte explicite du fait que les effets puissent être différents selon les personnes. Cette littérature est encore en développement et il n'existe pas toujours de consensus sur certains points.

Cet article présente les principes généraux de ces méthodes, illustrés à l'aide d'exemples issus de la littérature empirique récente. Il s'agit de fournir un guide pratique de ces méthodes, qui constituent des outils utiles pour l'économiste devant procéder à une évaluation *ex post* d'une politique. L'article propose également une grille d'interprétation critique à des utilisateurs éventuels de telles évaluations. Pour faciliter la lecture, le choix a été fait de se limiter aux formes les plus courantes de ces méthodes. On en trouvera certaines extensions dans Givord (2010). Les résultats sur les propriétés statistiques des différents estimateurs ne sont ni exposés en détails ni démontrés. En revanche, les références dont ils sont issus sont systématiquement détaillées. Pour une présentation plus formalisée, le lecteur intéressé peut se référer en particulier à Imbens et Wooldridge (2009).

Soulignons, en préambule, qu'il existe deux approches assez différentes pour traiter de ces questions. La première, qualifiée de "structuraliste", consiste à expliciter un modèle complet du comportement des agents économiques. Ce type d'approche est particulièrement utile *ex ante*, pour prévoir le résultat d'une politique qui n'a jamais été appliquée auparavant. Une autre approche, qui est plus "empirique", consiste à tester directement l'effet d'une mesure sur ce qui intéresse l'économètre, sans forcément expliciter tous les mécanismes à l'œuvre. C'est clairement à ce second courant que se rattachent les méthodes étudiées ici.

Ces deux approches sont plus complémentaires que concurrentes. Pour "calibrer" un modèle théorique, c'est-à-dire estimer les paramètres clés du modèle (comme par exemple les élasticités-prix des agents), il est nécessaire de recourir à des estimations empiriques. Inversement, pour être généralisés, les résultats d'une étude empirique requièrent souvent de passer par une explicitation des comportements économiques des agents concernés. Un courant récent tente de prendre en compte les apports de la modélisation économique en les validant à partir des résultats des expériences aléatoires (pour un exemple, voir Todd et Wolpin, 2006).

L'article est structuré comme suit. Nous commençons par rappeler le cadre classique dit "de Rubin" pour définir une inférence causale. Puis nous décrivons en détail les quatre principales méthodes empiriques d'évaluation : la méthode d'évaluation par différence de différences, la technique de variables instrumentales, la régression sur discontinuités et l'approche par sélection sur observables.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

La formalisation du problème d'évaluation : le cadre de Rubin

Dans toute la suite, nous nous plaçons dans le cadre dit "de Rubin", du nom du statisticien qui l'a popularisé au cours des années 1970. Le vocabulaire de cette approche emprunte beaucoup au champ de l'expérimentation médicale, dont il est issu. Ainsi, on parle souvent de "traitement" pour désigner la variable dont on cherche à mesurer l'effet.

En pratique, on s'intéresse donc à l'évaluation d'une mesure (ou traitement) *T*. Dans la version la plus simple, on distingue les personnes qui bénéficient de cette mesure ($T = 1$) de celles qui n'en bénéficient pas ($T = 0$). On s'intéresse alors à l'effet de cette mesure sur une grandeur d'intérêt, qu'on appellera souvent dans la suite le *revenu*. L'originalité ici est de considérer que chacun a "virtuellement" deux revenus potentiels, selon qu'il bénéficie ou non de la mesure. On note donc Y^0 le revenu associé à $T = 0$ (pas de traitement) et Y^1 le revenu associé à $T = 1$ (traitement).

La définition de la variable d'intérêt n'est jamais évidente et demande de s'interroger sur les objectifs que l'on assigne à la politique qu'il s'agit d'évaluer. Par exemple, s'intéresse-t-on à l'effet d'un stage de formation sur le retour à l'emploi, le salaire, la récurrence d'épisodes de chômage ? Si l'on diminue la taille des classes, cherche-t-on à augmenter les résultats moyens des élèves dans leur ensemble ou des élèves les plus faibles ; ou bien cherche-t-on plutôt à réduire les inégalités ? Un bilan complet d'une mesure ne peut être fourni par un chiffre unique. L'analyse de ses multiples facettes est importante.

Le problème fondamental de l'inférence causale

Pour une personne *i*, on peut définir l'effet propre (ou causal) de la mesure Δ_i à partir de ses deux revenus potentiels :

$$\Delta_i = Y_i^1 - Y_i^0$$

Le problème est que l'on n'observe pas simultanément Y_i^0 et Y_i^1. Pour un bénéficiaire de la mesure, on observe Y_i^1 mais pas Y_i^0. Pour un non bénéficiaire, on observe Y_i^0 mais pas Y_i^1.

Par exemple, il est possible de savoir si un chômeur a finalement obtenu un emploi après un stage de formation. Il est difficile de déterminer si cette reprise d'emploi est due au stage en lui-même : pour cela, il faudrait connaître les chances de retour à l'emploi de ce chômeur *s'il n'avait pas suivi ce stage*.

Tout l'esprit des méthodes d'évaluation est justement d'estimer cette situation *contrefactuelle*, c'est-à-dire ce qui aurait été observé si la mesure n'avait pas été mise en place. Pour cela, il suffit de trouver des personnes non affectées par la mesure mais totalement comparables par ailleurs à celles qui en ont bénéficié (personnes dites *traitées*). En pratique, il s'agit de trouver des personnes non bénéficiaires les plus comparables possibles aux personnes ayant bénéficié de la mesure. Ces personnes non bénéficiaires constituent ce qu'on appelle le *groupe de contrôle*.

Définition des paramètres d'intérêt : les effets moyens du traitement

Ce que l'on cherche le plus naturellement à estimer pour évaluer l'impact de la mesure est l'effet moyen de la mesure sur les individus qui en ont effectivement bénéficié, soit l'effet moyen du traitement sur les traités (*Average Treatment effect on the Treated* ou *ATT* dans la suite). Formellement, il s'écrit :

$$\Delta^{ATT} = E(Y_i^1 - Y_i^0 \mid T_i = 1)$$

Au-delà de l'effet de la mesure sur les seuls bénéficiaires, on peut se demander quel serait l'effet de cette mesure si elle était étendue à l'ensemble de la population. Il s'agit alors d'estimer l'effet moyen du traitement sur l'ensemble de la population (*Averaged Treatment Effect*, *ATE*), soit formellement :

$$\Delta^{ATE} = E(Y_i^1 - Y_i^0)$$

Ces effets sont susceptibles de dépendre des caractéristiques des personnes. Par exemple, les jeunes chômeurs peuvent davantage bénéficier d'un programme de formation que des chômeurs plus âgés (ou inversement). De même, il est possible que les élèves issus de milieux défavorisés soient plus sensibles à la taille des classes. Pour ces raisons, il peut être pertinent d'estimer des effets moyens, à caractéristiques observables données (par exemple : l'âge, le genre ou la nationalité). En pratique, cela revient à estimer des effets moyens conditionnels à ces caractéristiques observables. Nos deux paramètres d'intérêt s'écrivent donc plutôt :

$$\Delta^{ATT}(x) = E(Y_i^1 - Y_i^0 \mid T_i = 1, X_i = x)$$

et :

$$\Delta^{ATE}(x) = E(Y_i^1 - Y_i^0 \mid X_i = x)$$

Effets de sélection, hétérogénéité de l'effet du traitement

L'effet du traitement sur les traités correspond à la différence entre le revenu moyen avec traitement $E(Y^1 \mid T = 1)$ et le revenu que les traités auraient perçu

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

s'ils n'avaient pas été traités $E(Y^0|T=1)$. Si le premier terme est observé, le second ne l'est évidemment pas. Il faut donc trouver un estimateur crédible de ce revenu potentiel contrefactuel.

L'estimateur le plus simple, qu'on peut qualifier de "naïf", de l'effet moyen du traitement consiste à comparer les revenus des bénéficiaires $E(Y|T=1)$ avec ceux des non bénéficiaires $E(Y|T=0)$. Comme souligné plus haut, le problème est qu'en général, les bénéficiaires ne le sont pas par hasard. Le fait qu'ils ont été choisis pour bénéficier du traitement est probablement lié aux revenus qu'ils étaient en mesure d'espérer, avec ou sans traitement. On peut supposer par exemple que les chômeurs qui choisissent de ne pas suivre un stage de formation sont justement ceux qui ont des chances de retour à l'emploi (sans ce stage) suffisamment élevées pour se passer d'un tel stage. Dans ce cas, l'estimateur naïf sous-estimerait l'impact du stage. On peut imaginer, au contraire, que ceux qui choisissent de suivre le stage sont plus motivés et auraient en tout état de cause plus de chances de trouver un emploi. Dans ce cas, l'estimateur naïf surestime l'effet du stage[1].

Si Y désigne le revenu observé, alors la comparaison naïve des moyennes observées dans les deux groupes, qui correspond à un estimateur de $E(Y|T=1) - E(Y|T=0)$, constitue *a priori* un estimateur biaisé de l'effet moyen du traitement :

$$\begin{aligned} E(Y|T=1) - E(Y|T=0) &= E(Y^1|T=1) - E(Y^0|T=0) \\ &= E(Y^1|T=1) - E(Y^0|T=1) \\ &\quad + E(Y^0|T=1) - E(Y^0|T=0) \end{aligned}$$

On cherche à mesurer $E(Y^1|T=1) - E(Y^0|T=1)$, c'est-à-dire la variation de revenu due au traitement pour les personnes traitées. La comparaison naïve conduit donc à un biais, égal à $E(Y^0|T=1) - E(Y^0|T=0)$. Ce biais correspond à la différence de revenu moyen qui serait enregistrée en l'absence de la mesure, entre ceux qui ont choisi d'en bénéficier et les autres. Il n'est pas nul dès lors que les personnes choisissent de bénéficier de la mesure en fonction de leur revenu car, dans ce cas, la population des non bénéficiaires n'est pas représentative de celle des bénéficiaires. En termes statistiques, on dira que le revenu sans traitement Y^0 n'est pas indépendant du fait de suivre le traitement, soit formellement que l'hypothèse suivante n'est pas vérifiée :

(1) $Y^0 \perp T$

Le problème est encore plus sévère lorsque l'on veut estimer l'effet moyen du traitement sur toute la population (*ATE*). Dans ce cas, on voudrait connaître la situation contrefactuelle des bénéficiaires s'ils n'avaient pas été affectés par la mesure (comme précédemment), mais également la situation contrefactuelle des non bénéficiaires, s'ils avaient été affectés par cette mesure. Une décomposition simple montre en effet que :

$$\begin{aligned} \Delta^{ATE} &= P(T=1)[E(Y^1|T=1) - E(Y^0|T=1)] \\ &\quad + [1-P(T=1)]\,[E(Y^1|T=0) - E(Y^0|T=0)] \end{aligned}$$

L'effet moyen sur la population correspond à la moyenne de l'effet mesuré du traitement sur les bénéficiaires et les non bénéficiaires pondérée par leurs proportions respectives. Pour que l'effet de la mesure puisse être évalué à partir des bénéficiaires, il faut que l'effet moyen sur les bénéficiaires soit le même que celui qu'on aurait enregistré en moyenne si on avait appliqué le traitement aux non bénéficiaires. Cette décomposition montre également qu'on ne peut estimer, au mieux, qu'un impact moyen, alors que l'effet d'une mesure peut différer selon les individus. Le fait d'estimer un effet moyen sur une sous-population plutôt que sur une autre (par exemple, sur les bénéficiaires plutôt que sur les non bénéficiaires) n'est donc pas neutre si cette sous-population a des caractéristiques différentes du reste de la population. L'attention portée sur cette hétérogénéité constitue l'un des apports de cette littérature (nous y reviendrons par la suite).

Une hypothèse importante : l'absence d'externalités

Une hypothèse courante dans le cadre de l'évaluation est d'exclure l'existence d'externalités. Le traitement de l'individu n'a d'effet que sur son revenu, non sur ceux des autres personnes. Cette hypothèse est connue dans la littérature sous la terminologie d'hypothèse *SUTVA*, pour *Stable Unit Treatment Value Assumption*.

Cette hypothèse paraît assez naturelle. Toutefois, elle exclut des cas importants. Par exemple, une campagne de vaccination a un impact sur la propagation de la maladie et peut donc indirectement protéger des personnes qui n'ont pas été vaccinées. Une politique de soutien scolaire aux plus faibles peut avoir des effets d'entraînement au sein d'une classe. Lorsqu'une mesure d'aide à l'emploi des chômeurs est mise en œuvre, il est difficile de croire que le marché de l'emploi est indéfiniment extensible et sera capable d'absorber à court terme l'ensemble des personnes aidées, même si le programme est individuellement efficace. Selon

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

l'ampleur du dispositif, il est plus ou moins plausible d'ignorer ces effets de bouclage.

Lorsqu'ils ne peuvent pas être ignorés, ces effets de bouclage peuvent être explicitement pris en compte en intégrant les résultats obtenus dans un modèle structurel (voir, par exemple, Lise *et alii*, 2004). En conservant une approche strictement empirique, une solution pour prendre en compte la présence d'externalités est de raisonner à un niveau plus agrégé quand cela a un sens. Une illustration classique de cette démarche est l'article de Miguel et Kremer (2004). Celui-ci évalue l'impact de traitements anti-parasitaires fournis à des enfants scolarisés dans certaines écoles du Kenya (désignées au hasard au cours d'une expérience contrôlée). Ces parasites sont très contagieux : traiter certains enfants seulement peut avoir un effet sur les camarades, même si ces derniers ne bénéficient pas directement du traitement. Comparer simplement l'état de santé des enfants traités à celui des autres conduirait donc à sous-estimer l'effet réel du traitement. Pour prendre en compte ce biais, Miguel et Kremer (2004) se placent à un niveau plus agrégé. Le principe de leur approche consiste à comparer l'état de santé moyen de l'ensemble des enfants scolarisés dans les écoles ayant bénéficié du programme (même si tous les enfants n'en ont pas effectivement bénéficié, par exemple parce que certains étaient absents le jour du traitement) par rapport à l'état de santé moyen des enfants scolarisés dans les écoles n'en ayant pas bénéficié.

Méthodes empiriques d'évaluation

Tout l'enjeu des méthodes empiriques d'évaluation réside donc dans la possibilité de trouver des données et des situations permettant de neutraliser l'effet de sélection. La manière la plus directe consiste à se placer dans le cadre d'expériences aléatoires contrôlées. Dans ce cadre, l'expérimentation consiste à déterminer aléatoirement les personnes qui bénéficieront du traitement, par exemple un accompagnement renforcé pour l'aide à la recherche d'emploi pour des chômeurs, et celles qui n'en bénéficieront pas (le groupe de contrôle). Si la répartition aléatoire est bien vérifiée, le choix d'être traité ou non est indépendant des revenus et :

$$\Delta(X) = E(Y|T = 1, X) - E(Y|T = 0, X)$$

$$= E(Y^1|T = 1, X) - E(Y^0|T = 0, X)$$

$$= E(Y^1|T = 1, X) - E(Y^0|T = 1, X)$$

correspond bien à l'effet du traitement sur les traités.

Les expériences aléatoires constituent le paradigme des méthodes économétriques d'évaluation. Elles sont présentées en détail dans Duflo *et alii* (2008). Cependant, même si leur utilisation commence à se développer en France, elles restent encore rares. De fait, elles nécessitent une intervention très en amont de la mise en place de la politique qu'il s'agit d'évaluer. Souvent, le statisticien est plutôt conduit à estimer *ex post* des politiques déjà mises en place. Plusieurs méthodes ont été développées pour ce faire, que nous présentons ici.

Un premier ensemble de méthodes tente d'utiliser des variations exogènes de l'environnement économique, par exemple des réformes de législation créant une situation presque expérimentale. On parle dans ce cas d'"expériences naturelles" ou de "quasi-expériences". Les trois méthodes présentées ici qui entrent dans cet ensemble sont les différences de différences, les variables instrumentales et les régressions sur discontinuités. Elles se distinguent par les hypothèses sous lesquelles on estime correctement l'impact causal de la mesure qu'on souhaite évaluer. Le fait d'appliquer l'une ou l'autre est souvent tributaire des données.

Souvent, toutefois, aucune de ces méthodes n'est applicable. Pour réduire le biais de sélection, il faut au minimum que l'économètre contrôle les différences observables entre les bénéficiaires et les non bénéficiaires. Les estimateurs par appariement (ou *matching*) consistent pour ce faire à apparier chaque personne traitée avec un "jumeau" non traité possédant les mêmes caractéristiques par ailleurs. Ces estimateurs permettent d'identifier l'impact réel du traitement sous l'hypothèse, forte, que l'ensemble des observables est suffisamment riche pour que, conditionnellement à ces observables, le fait de choisir de suivre ou non le traitement soit indépendant des caractéristiques inobservables.

Les estimateurs de différence de différences

Le principe d'une évaluation est d'estimer un contrefactuel, c'est-à-dire la situation qui aurait prévalu en l'absence du programme que l'on souhaite évaluer. Comme on l'a vu, comparer les revenus des bénéficiaires d'une mesure avec ceux des non bénéficiaires risque de donner des résultats biaisés par la présence d'effets de sélection. Comparer les revenus après traitement à ce qu'ils étaient avant la mise en place de la réforme (estimateur "avant/après") n'est pas plus satisfaisant. De nombreux facteurs sont susceptibles d'expliquer des variations du revenu dans le temps, indépendamment des effets de la réforme que l'on souhaite évaluer : conjoncture économique plus ou

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

moins favorable, profil de revenu croissant avec l'expérience, etc.

L'idée, très simple, des estimateurs de différence de différences est de mélanger ces deux approches. Dans le cas le plus simple, on dispose de deux groupes observés avant et après l'introduction de la mesure à évaluer. Cette réforme ne concerne que l'un des deux groupes. L'estimateur consiste alors à comparer l'évolution des revenus des bénéficiaires avant et après le traitement à celle des revenus des personnes du groupe de contrôle sur la même période.

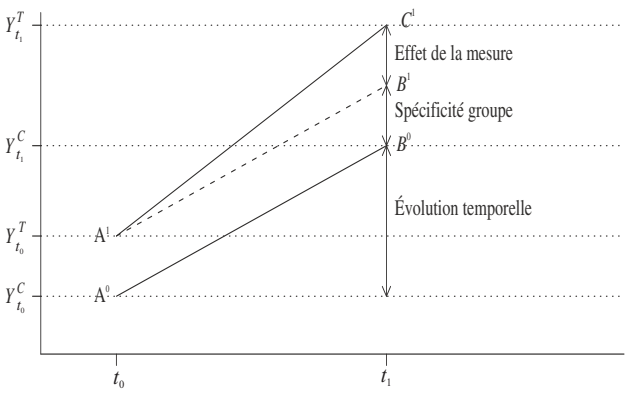
Son nom résume bien la méthode : on fait la différence entre les deux groupes des évolutions respectives (différences temporelles) de leurs revenus sur la période. L'une des différences doit permettre d'éliminer les écarts systématiques entre le groupe des bénéficiaires et les autres individus. L'autre différence doit permettre d'éliminer l'évolution temporelle, supposée identique pour les deux groupes en l'absence de la mesure.

Cet estimateur permet d'identifier l'effet propre, causal, de la mesure à évaluer, sous l'hypothèse que les revenus des deux groupes auraient effectivement évolué de manière identique en l'absence de la mesure.

Hypothèse identifiante

Dans le cas caricatural présenté sur la figure 1, les évolutions des revenus sans traitement des deux groupes sont parfaitement identiques. Du fait de la mesure, le revenu du groupe traité ($T = 1$), indiqué par l'exposant T, a cependant évolué plus vite que celui du groupe de contrôle ($T = 0$), indiqué par l'exposant C. L'estimateur naïf qui compare les revenus des ménages traités et non traités à l'issue de la réforme (donné par la distance B^0C^1) est biaisé par des effets de sélection, car les ménages traités sont intrinsèquement différents des ménages non traités (différence *ex ante* de revenu A^0A^1). L'effet réel du traitement est donné par B^1C^1.

Figure 1 : évolution attendue des revenus des deux groupes dans un cas de différence de différences



Lecture : en l'absence du programme, l'évolution du revenu dans le groupe de traitement entre les dates t_0 et t_1 aurait été identique à celle observée dans le groupe de contrôle (de A^0 à B^0). Du fait du programme, il est en réalité augmenté de A^1 à C^1. Seul l'écart entre B^1 et C^1 correspond à l'effet du programme.

Source : auteure.

Cette hypothèse peut s'écrire de manière plus formelle. Toujours dans le cas à deux groupes et deux périodes, elle revient à supposer que le revenu potentiel sans traitement moyen d'un individu i à la période t appartenant au groupe de contrôle, noté G^C, ou de traitement, noté G^T, s'écrit :

$$E(Y^0_{it}) = \alpha + \beta 1_{\{t=t_1\}} + \gamma_T 1_{\{i \in G^T\}}$$

où $1_{\{.\}}$ est la variable indicatrice valant 1 si la condition entre parenthèses est vérifiée et 0 sinon. α représente le revenu moyen sans traitement, avant la mise en place de la mesure en t_0, des personnes appartenant au groupe de contrôle. Ce revenu augmente de β entre les périodes t_0 et t_1. Même en l'absence de la mesure, les personnes du groupe de traitement ont un revenu potentiel moyen différent des personnes du groupe de contrôle. Cette différence, γ_T, est supposée constante entre les deux périodes. Cette hypothèse de séparabilité repose sur une double condition : la stabilité de la composition des groupes au cours du temps ; des effets conjoncturels identiques dans tous les groupes.

Revenons au revenu observé d'une personne i, à la période t. Il peut s'écrire :

$$(2)\quad Y_{it} = \alpha + \beta 1_{\{t=t_1\}} + \gamma_T 1_{\{i \in G^T\}} + \delta T_{it} + u_{it}$$

en notant δ l'effet causal moyen du traitement pour les traités $E(Y^1_{it} - Y^0_{it} \mid T_{it} = 1)$, u_{it} un terme de perturbation qui tient compte de la variabilité individuelle et T_{it} une indicatrice valant 1 lorsque l'individu i bénéficie de la mesure à la date t et 0 sinon. Dans ce cas, T_{it} correspond à l'interaction des indicatrices temporelles et de groupe : $T_{it} = 1_{\{t=t_1\}} 1_{\{i \in G^T\}}$. Une condition suffisante pour identifier sans biais l'effet du traitement δ est que l'indicatrice de traitement T_{it} soit exogène conditionnellement au fait d'appartenir à l'un ou l'autre des groupes et des périodes. Le fait de bénéficier du traitement ou non ne doit pas dépendre des chocs sur les revenus contrefactuels passés et/ou futurs. Cette hypothèse identifiante n'est pas vérifiée lorsque le choix d'en bénéficier ou pas se fait en fonction des revenus précédents (voir *infra*).

Déterminer un groupe de contrôle

La question consiste alors à déterminer un groupe de contrôle plausible, *i.e.* susceptible d'avoir effectivement connu des évolutions temporelles comparables à celles du groupe traité. Il faut insister sur le fait que, pour évaluer un dispositif avec libre entrée, le groupe des "ayants droit" qui ont choisi de

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

ne pas participer ne constitue pas *a priori* un bon groupe de contrôle. Il est possible que les problèmes classiques de sélection réapparaissent : le fait qu'ils aient renoncé au dispositif peut être lié aux bénéfices qu'ils anticipaient pouvoir en retirer... Si l'on pense qu'ils ont fait ce choix en connaissance de cause, cela signifie que l'évolution de leur revenus potentiels anticipés se distingue de l'évolution des revenus potentiels des bénéficiaires du dispositif. La double différenciation ne permettra pas *a priori* de régler ce problème. En revanche, des réformes ciblées sur certaines sous-populations offrent des situations de "quasi-expérience" si le ciblage n'est pas déterminé directement par le bénéfice que ces personnes peuvent tirer de ces réformes ou des modifications de l'environnement économique. Un exemple célèbre est l'évaluation par Card et Krueger (1994) de l'impact d'une augmentation du salaire minimum intervenue dans l'État du New Jersey sur l'emploi non qualifié. Les auteurs comparent pour ce faire l'emploi dans les établissements de restauration rapide de cet État et de l'État voisin de Pennsylvanie, juste avant et six mois après l'augmentation du salaire minimum dans le New Jersey. Le choix de se limiter à ce type d'établissements est justifié par le fait qu'ils sont des employeurs importants de main-d'œuvre non qualifiée, avec des niveaux de rémunération souvent proches du salaire minimum. Par ailleurs, ces établissements utilisent des modes de production et des matières premières assez standardisés, ce qui limite les écarts entre États. Néanmoins, il peut subsister des différences dans les niveaux d'emploi de ce type d'établissements entre les deux États, par exemple si les consommateurs n'ont pas les mêmes caractéristiques ou les mêmes habitudes alimentaires dans ces deux États. Une simple comparaison serait donc biaisée. L'hypothèse identifiante est ici que les *évolutions* de ces niveaux d'emploi auraient été identiques en l'absence d'augmentation du salaire minimum au New Jersey. De manière plus anecdotique, Card (1990) s'appuie sur l'afflux massif et rapide d'immigrants cubains à Miami suite à l'exode de Mariel[2] pour évaluer l'impact de l'immigration sur le marché du travail, en utilisant comme points de comparaison des villes américaines "proches". Ici encore, l'hypothèse identifiante ne suppose pas que le marché du travail de Miami était semblable à celui de ces autres villes, ce qui serait très restrictif, mais que les *évolutions* de ces marchés du travail auraient été identiques en l'absence de ce choc.

Ces exemples constituent des expériences "naturelles", encore appelées des "quasi-expériences". On utilise un choc sur l'environnement économique exogène pour identifier un effet causal (l'effet du salaire minimum sur l'emploi non qualifié pour Card et Krueger, 1994, l'impact de l'afflux d'une main-d'œuvre étrangère dans le cas de Card, 1990). On s'approche ainsi des conditions des expériences contrôlées, au sens où l'on peut raisonnablement supposer que le fait que certaines populations soient affectées ou non n'est pas lié à leurs caractéristiques propres (ce n'est pas parce que le marché de l'emploi était particulièrement dynamique qu'il y a eu une vague d'immigration à Miami).

Ces quasi-expériences proviennent souvent de dispositifs ciblés sur une tranche d'âges (on peut alors comparer les individus traités avec des populations d'âges différents), des zones géographiques, des tranches de revenus, etc. Insistons encore sur le fait que le ciblage ne doit pas être directement lié aux caractéristiques des populations visées. La plupart des études qui appliquent une méthode de différence de différences définissent le groupe de contrôle de manière *ad hoc*. En général, on compare la situation avec un groupe considéré comme semblable, mais il existe toujours une part d'arbitraire dans le choix de ce semblable. Une autre solution peut être de constituer un groupe de contrôle en appariant les bénéficiaires avec des non bénéficiaires qui présentent des caractéristiques observables identiques. Nous verrons cela en détail dans la partie sur l'appariement (ou *matching*). Une autre méthode est proposée par Abadie *et alii* (2010), dans le cas où l'on dispose de données agrégées, sur une région affectée par un choc ou une politique particulière par exemple. Cette méthode consiste à construire un contrefactuel à partir de l'ensemble des individus non traités (voir Givord, 2010 pour une description plus détaillée).

Estimation : quelles données, quelles méthodes ?

Considérons le cas le plus simple avec deux groupes et deux périodes. L'estimateur de différence de différences vise à estimer le paramètre d'intérêt :

$$(3)\ \Delta = E(Y_{it_1} - Y_{it_0} \,|\, T = 1) - E(Y_{it_1} - Y_{it_0} \,|\, T = 0)$$

où Y_{it_1} représente le revenu de l'individu *i* observé après l'introduction de la mesure et Y_{it_0} celui avant son introduction. En pratique, l'estimateur de différence de différences $\hat{\Delta}$ peut être obtenu à partir de l'équation (2) comme l'estimateur des moindres carrés ordinaires de δ.

On dispose souvent d'observations sur de plus nombreuses périodes, que ce soit avant ou après la mesure. On peut être également amené à distinguer plusieurs groupes, de manière à mieux capter des différences au sein de l'ensemble des bénéficiaires ou des non bénéficiaires. Par exemple, pour une politique s'appliquant à plusieurs régions, on souhaitera tenir compte d'éventuelles différences régionales. Dans ce cas, on introduira autant d'indicatrices temporelles et d'effets fixes par groupe rendant compte des différences permanentes entre les groupes.

Lorsque des données de panel (*i.e.* des données individuelles sur les même personnes aux différentes

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

dates) sont disponibles, il est dommage de ne pas utiliser l'information supplémentaire qu'elles apportent. Typiquement, on peut mieux contrôler l'hétérogénéité individuelle en utilisant des données répétées pour chaque individu. En pratique, cela revient à remplacer les indicatrices de groupes par des effets fixes individuels :

$$(4)\quad Y_{it} = c_i + \sum_{\tau=1}^{L} e_\tau 1_{\{\tau = t\}} + \delta T_{it} + u_{it}$$

où c_i est un effet fixe individuel, les termes e_τ correspondent à des effets temporels et T_{it} est, comme auparavant, une indicatrice valant 1 lorsque l'individu *i* bénéficie de la mesure à la date *t*. On peut alors estimer δ par des procédures standard d'économétrie des panels, comme des estimations en différences ou des estimateurs à effets fixes (*intra* ou *within*). Rappelons que la première procédure revient à différencier les variables selon la dimension temporelle tandis que la seconde utilise les écarts à la moyenne temporelle, ces deux opérations permettant d'éliminer les effets fixes c_i.

Il n'est cependant pas nécessaire de disposer de données de panel. Des données agrégées sur les différents groupes aux différentes périodes suffisent. Cependant, pour que l'hypothèse identifiante ait des chances d'être vérifiée, il faut que la composition des groupes au sein des échantillons dont on dispose soit stable au cours du temps.

Si l'on dispose de plus d'informations sur les groupes concernés, il peut être judicieux (et surtout plus efficace) de contrôler l'effet de ces variables observables, qui sont susceptibles de varier entre les groupes et au cours du temps. Dans ce cas, l'hypothèse d'identification est moins stricte que dans le cas général. Elle suppose que l'évolution des groupes qui n'ont pas bénéficié du traitement aurait été la même que celle des groupes de traitement, à l'évolution des déterminants observables *X* près. Cela signifie que toutes les évolutions différenciées des revenus entre les groupes traités et non traités qui ne s'expliquent pas par des variations de ces contrôles sont attribuées au traitement.

Quelle que soit la spécification retenue, il est important de souligner que l'estimation de la précision de l'estimateur de différence de différences $\hat{\Delta} = \hat{\delta}$ par les procédures classiques peut être biaisée. L'hypothèse d'indépendance des résidus et d'homoscédasticité peut être mise en cause par deux effets : l'autocorrélation temporelle (les variables de résultats sont persistantes dans le temps) et des effets groupés (les individus des différents groupes sont affectés par des chocs communs). Il est donc préférable d'utiliser pour le calcul de la précision des procédures de *bootstrap* par bloc ou des estimateurs de la matrice de variances-covariances adaptés (voir Bertrand *et alii*, 2002 pour une discussion détaillée, résumée dans Givord, 2010). Pour obtenir des estimateurs non biaisés de la précision, il est nécessaire de disposer de suffisamment de groupes différents.

Validité de l'hypothèse identifiante et tests de robustesse

L'hypothèse identifiante est impossible à tester, puisque cela signifierait qu'on pourrait observer l'évolution contrefactuelle du groupe traité en l'absence de traitement... Il peut être cependant utile de faire plusieurs tests, pour juger de sa vraisemblance.

Comparaison des évolutions passées du revenu

Le premier test consiste à vérifier que, sur les périodes antérieures, les groupes traités et de contrôle ont bien connu des évolutions similaires. Un graphique des évolutions des revenus observés pour les différents groupes peut permettre de se faire une idée sur la crédibilité de l'hypothèse identifiante. Un cas d'école où la condition n'est pas vérifiée est donné par Ashenfelter (1978). Ce dernier s'intéresse à l'impact d'une formation sur les revenus de salariés. Une analyse détaillée de l'évolution des revenus avant la formation montre que, l'année précédant la formation, les salariés formés avaient enregistré une croissance de leur revenu plus faible qu'auparavant. Il est vraisemblable que c'est la raison pour laquelle ils ont choisi de se former. Il est dès lors difficile de savoir si l'effet sur les salaires qu'on observe est bien l'impact de la formation ou celui d'un "rattrapage" des revenus sur un sentier de croissance moyen, après un choc négatif. Cet effet est passé à la postérité sous la dénomination d'*Ashenfelter's Dip*.

Tests "Placebo"

De manière plus systématique, on peut appliquer la même procédure de différence de différences, mais autour d'une date où il ne s'est rien passé. Si l'on mesure malgré tout un impact significatif de cette mesure *"placebo"*, on peut craindre qu'il existe une différence significative d'évolution entre les deux groupes et que le groupe de contrôle ne soit pas adapté.

Dans le cas où l'on dispose de séries temporelles longues et où il s'agit de tester une mesure appliquée à des dates différentes selon les groupes, Autor (2003) propose une procédure inspirée des tests de causalité au sens de Granger. L'idée est de vérifier que le fait d'être traité dans le futur n'a pas d'impact sur le revenu actuel. En effet, toute la question de la sélection est de savoir si c'est la mesure qui modifie le revenu auquel on s'intéresse ou si c'est pour

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

répondre à des évolutions du revenu que la mesure est mise en place : dit autrement, que les conséquences ne précèdent pas les causes.

En pratique, on peut estimer une relation du type :

$$(5)\quad Y_{it} = \sum_{g=1}^{K} e_g 1_{\{i \in G^g\}} + \sum_{\tau=1}^{L} e_\tau 1_{\{\tau = t\}} + \sum_{\tau=0}^{n} \alpha_\tau T_{i,t-\tau} + \sum_{\tau=1}^{m} \delta_\tau T_{i,t+\tau} + u_{it}$$

On s'attend à ce que les coefficients δ_τ, qui correspondent à l'impact de la mesure après sa mise en place, ressortent significativement de l'estimation de (5)[3]. Au contraire, il serait inquiétant d'observer que faire partie d'un futur groupe traité (coefficients α_τ) a un effet significatif sur le revenu. Enfin, lorsque l'on dispose de plusieurs groupes de contrôle possibles, il est judicieux de vérifier que l'effet mesuré est identique selon qu'on utilise l'un ou l'autre de ces groupes[4].

En résumé

En résumé, l'utilisation d'une méthode de différence de différences pour évaluer une mesure de politique publique nécessite de :

– déterminer un groupe de contrôle : ce groupe n'est "crédible" que si l'on peut supposer que les évolutions des revenus dans ce groupe sont les mêmes que celles qu'auraient connues les bénéficiaires de la mesure en l'absence de cette dernière. Attention, le groupe de contrôle ne peut être simplement constitué des individus non bénéficiaires, dès lors qu'il y a (auto) sélection dans le dispositif. Disposer de données cohérentes avant et après la mise en place de la mesure est indispensable. Il est préférable que les données soient relativement proches dans le temps, pour éviter de capter d'autres évolutions qui ne seraient pas liées à la mesure à évaluer ;

– estimer le paramètre représentatif de l'impact de la mesure : en pratique, cela peut se faire très simplement par une procédure de moindres carrés ordinaires. Le calcul de la précision de l'estimateur est nécessaire pour borner statistiquement la vraie valeur de l'impact de la mesure. Il faut tenir compte pour cela de l'auto-corrélation temporelle éventuelle, ainsi que de la présence de chocs "groupés" si des données individuelles sont utilisées.

Variables instrumentales

L'utilisation de variables instrumentales pour traiter des problèmes d'endogénéité est très ancienne et plus générale que leur utilisation pour des questions d'évaluation des politiques publiques. Rappelons que l'on se place dans le cadre standard de l'équation linéaire :

$$(6)\quad Y_i = \alpha + \delta T_i + u_i$$

où l'on soupçonne le régresseur T_i d'être endogène, c'est-à-dire d'être corrélé avec le terme d'erreur u_i. C'est typiquement le cas lorsqu'il existe des effets de sélection.

Un instrument est alors simplement une variable Z qui remplit deux conditions (éventuellement conditionnellement à des variables observables X) :

– Z est corrélée avec la variable endogène T : $\text{cov}(T, Z) \neq 0$;

– mais pas avec les résidus u : $\text{cov}(u, Z) = 0$.

Si une telle variable existe, il est possible d'obtenir une estimation convergente du paramètre d'intérêt δ. L'estimateur le plus classique est celui des doubles moindres carrés. Il s'obtient en régressant le revenu Y non plus sur le régresseur endogène T, mais sur sa prédiction par les moindres carrés ordinaires $\hat{T}$ issue de la régression de première étape de T sur Z :

$$(7)\quad \hat{\delta}_{VI} = (\hat{T}'\hat{T})^{-1}\hat{T}'Y$$

On peut donc l'obtenir en deux étapes, d'où le nom de doubles moindres carrés : on régresse le traitement sur l'instrument, puis le revenu sur la valeur prédite du traitement par cette première régression. Attention, il est fortement déconseillé de procéder ainsi. En effet, les écarts-types fournis par la seconde régression ne sont alors pas corrects (*cf.* par exemple Robin, 2002). Tous les logiciels de statistique standard permettent cependant de mettre en œuvre cette procédure (par exemple, sous le logiciel *SAS*, par la procédure *syslin*, option 2*sls*) ; ils fournissent en particulier une estimation correcte de la précision de cet estimateur.

Remarquons que, lorsque l'instrument est binaire, un estimateur courant dans la littérature de l'évaluation est l'estimateur de Wald :

$$(8)\quad \hat{\delta}_W = \frac{\hat{E}(Y|Z=1) - \hat{E}(Y|Z=0)}{\hat{E}(T|Z=1) - \hat{E}(T|Z=0)}$$

les contreparties empiriques des espérances théoriques E étant les moyennes empiriques ($\hat{E}$).

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

L'estimateur $\hat{\delta}_W$ correspond à la variation des revenus avec l'instrument (*i.e.* à la différence entre la moyenne empirique des revenus pour les observations telles que $Z = 1$ la moyenne empirique des revenus pour les observations telles que $Z = 0$), rapportée à la variation du traitement avec l'instrument.

On peut montrer simplement que $\hat{\delta}_W$ correspond exactement à l'estimation de δ par les doubles moindres carrés ordinaires dans l'équation (6), en l'absence de variables explicatives supplémentaires et avec Z comme instrument.

Quels instruments ?

Pour simple qu'elle soit, la définition d'une variable instrumentale peut laisser fort démuni un économètre averti. Un instrument valide est une variable qui est liée au fait de bénéficier d'un programme, pour une raison que l'économètre peut expliciter et expliquer, mais qui n'a par ailleurs pas d'impact direct sur les autres déterminants du revenu. Il est parfois d'usage d'utiliser les valeurs ou différences retardées. La validité de ces instruments est souvent sujette à caution. C'est pourquoi l'attention s'est déplacée ces dernières années sur des instruments dont l'exogénéité est plus transparente.

Affectation aléatoire

Les instruments les plus crédibles sont ceux tirés d'une affectation aléatoire, par exemple dans le cadre d'une expérience contrôlée. En général, les personnes affectées au groupe de traitement se voient proposer de participer à un programme ou sont plus incitées que celles du groupe de contrôle, mais peuvent ensuite choisir de participer ou non. Ceci introduit à nouveau de la sélection, puisqu'il est probable que celles qui choisissent *in fine* de participer en anticipent un bénéfice plus élevé que les autres. Néanmoins, on dispose d'un instrument "idéal", l'orientation au traitement Z (*i.e.* le fait d'affecter une personne au groupe des participants potentiels, auquel le programme qu'on souhaite évaluer sera proposé, ou au groupe de contrôle). Par définition de l'expérimentation, cette orientation est en effet aléatoire, donc :

$$(9)\quad Y^1, Y^0 \perp Z$$

Si l'expérience est bien menée, on peut légitimement espérer que cette orientation soit corrélée avec l'entrée effective dans le programme :

$$(10)\quad \mathrm{cov}(Z, T) \neq 0$$

Les conditions pour disposer d'un bon instrument sont donc remplies. On peut alors estimer l'impact du programme à évaluer par une méthode classique de doubles moindres carrés ou par un estimateur de Wald.

Expériences naturelles

En l'absence d'expérience aléatoire, de bons instruments peuvent être fournis par des "expériences naturelles". Comme décrit dans l'introduction, il s'agit d'utiliser des variations exogènes de l'environnement des individus pouvant expliquer des probabilités plus ou moins élevées de participer. Par exemple, Angrist et Evans (1998) utilisent la "préférence pour la différence" des parents américains pour évaluer l'impact de la fécondité sur la participation des femmes au marché du travail. On observe en général que les mères de famille nombreuse (plus de trois enfants) sont moins souvent actives sur le marché du travail. Il est difficile de déterminer le sens de la causalité : est-il plus difficile pour ces mères de travailler (par exemple du fait d'une prise en charge insuffisante de la petite enfance ou de discrimination de la part des employeurs) ou ces femmes privilégient-elles simplement la vie de famille à la vie professionnelle? Angrist et Evans proposent un instrument original pour déterminer la probabilité d'avoir une famille nombreuse. Sur les générations récentes, on observe plus souvent des familles nombreuses (plus de trois enfants) lorsque les deux aînés sont du même sexe. Le fait d'avoir deux aînés du même sexe est donc un prédicteur du fait d'avoir au moins trois enfants. Il s'agit d'un événement totalement hors de portée de la volonté des parents et, donc, indépendant des caractéristiques éventuelles des femmes concernées.

Un autre exemple célèbre est l'utilisation par Angrist et Krueger (1991) du trimestre de naissance pour évaluer les rendements de l'éducation (définis comme le surcroît de revenu assuré par une année d'études supplémentaire). En effet, du fait de la législation américaine sur la scolarité obligatoire, les enfants entrent à l'école en septembre de l'année de leur six ans (soit entre cinq ans et neuf mois et six ans et huit mois, selon leur mois de naissance). En revanche, l'obligation scolaire cesse le jour des seize ans, quel que soit le moment dans l'année scolaire. Les enfants nés en début d'année sont donc théoriquement astreints à une scolarité obligatoire plus courte que ceux nés en fin d'année. De fait, pour les générations nées dans les années 1930 et 1940 (pour lesquelles la scolarité obligatoire était vraiment contraignante), être né au premier trimestre réduisait de 0,6 année la durée d'étude finale. Là encore, on peut supposer que le trimestre de naissance n'est en revanche pas corrélé avec les autres déterminants du revenu, ce qui en fait un bon instrument potentiel (nous revenons plus loin sur cet exemple).

D'autres études utilisent la dispersion régionale de certains dispositifs. Par exemple, pour évaluer

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

l'efficacité d'un nouveau protocole de prise en charge des infarctus, McClellan *et alii* (1994) utilisent la proximité de la résidence des patients aux hôpitaux utilisant ce nouveau protocole[5].

Interprétation : un estimateur local

Que mesure l'estimateur des variables instrumentales ? Dans cette partie, nous nous plaçons dans le cadre simple d'un instrument binaire (*i.e.* valant 0 ou 1), tout comme le traitement.

Comme l'ont montré Angrist *et alii* (1996) dans un papier influent, cet estimateur ne permet de mesurer l'effet moyen ni sur l'ensemble des personnes "traitées" ni sur la population totale (*i.e.* ni le paramètre Δ^{ATT} ni le paramètre Δ^{ATE} définis dans la première partie de l'article). Il mesure l'effet moyen uniquement sur une sous-population, les *compliers*. Il s'agit des personnes pour lesquelles l'instrument a un impact sur le fait de suivre ou non le traitement.

L'exemple d'Angrist *et alii* (1996) repose sur la "loterie" utilisée pour la conscription durant la guerre au Vietnam, sur laquelle ils s'appuient pour instrumenter l'effet d'avoir été vétéran sur les revenus ultérieurs. Il s'agit *a priori* d'un instrument "parfait" : il est aléatoire, donc non corrélé avec les déterminants inobservés du traitement. En outre, selon le numéro obtenu à la loterie, les jeunes des tranches d'âges concernées ont eu plus ou moins de chances d'échapper à la conscription. Notons que cela ne signifie pas, loin s'en faut, que le numéro obtenu corresponde exactement au fait d'être parti au Vietnam. Certains ont pu s'engager alors qu'ils avaient un numéro leur permettant d'en être dispensés, tandis que d'autres ayant reçu un "mauvais" numéro ont été réformés, pour des raisons physiques par exemple.

Les individus adaptent (ou non) leur comportement à la valeur de l'instrument. Plus exactement, on a une fonction de réaction "potentielle" $T(Z)$. Dans l'exemple de la loterie du Vietnam, il s'agit de la manière dont le choix de s'engager ou pas dépend du numéro attribué. Évidemment, on est confronté au même problème observationnel que pour les revenus potentiels : on observe le choix d'un individu pour une valeur particulière de l'instrument, mais pas pour l'autre. Dans le cas où l'instrument et le traitement sont binaires, on a la partition de la population fournie par le tableau 1.

Tableau 1 : partition de la population selon le comportement vis-à-vis de l'instrument

	$T=1, Z=1$	$T=0, Z=1$
$T=1, Z=0$	*Always Takers*	*Defiers*
$T=0, Z=0$	*Compliers*	*Never Takers*

Lecture : les personnes qui choisissent toujours le traitement quelle que soit la valeur de l'instrument, donc telles que $T_i = 1$ que Z_i soit égal à 1 ou à 0, sont appelées *always takers*.

Plus précisément, on désigne par *compliers* les individus pour lesquels l'instrument a un effet (au sens où ils se comportent conformément à ce qui est attendu). Pour d'autres, l'instrument n'a pas d'effet. Soit ils ne choisissent jamais le programme, $T_i(1) = T_i(0) = 0$ (Angrist *et alii*, 1996, les nomment les *never takers*), soit ils le choisissent systématiquement $T_i(1) = T_i(0) = 1$ (*always takers*). Enfin, on pourrait en toute rigueur imaginer qu'il existe des "rebelles" (*defiers*), *i.e.* des personnes pour lesquelles l'instrument a un effet opposé à celui couramment attendu. Dans le cas de la loterie du Vietnam, il s'agirait des cas improbables de personnes qui se seraient engagées si elles avaient reçu un numéro leur permettant de ne pas partir et qui, à l'opposé, se seraient faites réformer en cas de numéro les soumettant à la conscription...

Angrist *et alii* (1996) montrent que l'estimateur des doubles moindres carrés permet d'estimer l'effet du traitement uniquement pour la sous-population des *compliers*, c'est-à-dire pour ceux pour qui l'instrument a un impact sur le traitement final. Dans l'exemple du Vietnam, il s'agit des personnes qui se sont engagées (ou non) uniquement en fonction du numéro qui leur a été attribué.

L'effet estimé est donc dit local : Angrist *et alii* (1996) l'appellent le *LATE*, soit *Local Average Treatment Effect*.

Ils se placent en effet dans le cadre d'une affectation aléatoire de l'instrument Z, indépendant du revenu (tout l'effet passe par le traitement) et supposent que les *defiers* n'existent pas. Cette dernière hypothèse peut se formaliser par la relation *monotone* suivante :

$$(11)\quad T_i(1) \geq T_i(0)$$

On peut alors montrer que l'estimateur des variables instrumentales s'écrit :

$$(12)\quad \hat{\delta}_{VI} = \hat{E}[(Y_i^1 - Y_i^0) | T_i(1) - T_i(0) = 1]$$

L'estimateur ne mesure donc que l'effet du traitement pour ces *compliers*. Si l'effet du traitement est identique pour tous, c'est-à-dire si $Y_i^1 - Y_i^0 =$ *constante*, ce n'est pas un problème. Sinon, l'effet n'est identifié que sur une sous-population particulière. Deux remarques sont importantes. La première est que cette population n'est pas connue : *a priori*, il n'est pas possible de déterminer si tel ou tel individu appartient ou non à la population des *compliers*[6]. La seconde est que l'estimation dépendra de l'instrument utilisé. Des estimateurs utilisant des instruments différents sont susceptibles de donner des résultats différents. Pour donner un sens aux résultats obtenus, il faut encore une fois tenter de caractériser la sous-population sur laquelle l'effet est identifié. La théorie économique peut permettre d'apporter des pistes d'interprétation. Par exemple, Rosenzweig et

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Wolpin (2000) proposent un modèle de choix d'éducation qui permet d'interpréter les résultats obtenus dans la littérature sur les rendements de l'éducation.

Instruments faibles

Revenons maintenant à une discussion plus classique sur la qualité des instruments. En général, la discussion porte sur la validité de l'instrument retenu : il s'agit de justifier que l'instrument n'est pas corrélé avec les déterminants inobservés du revenu. Cette propriété est essentielle : elle assure que l'estimateur obtenu par variables instrumentales sera asymptotiquement sans biais. Il n'existe, cependant, aucun test statistique permettant de valider rigoureusement cette propriété. Rappelons que les tests dits de Sargan ou *J-tests* ou tests de sur-identification qui sont utilisés lorsque l'on dispose de plus d'instruments que de variables endogènes ne testent que la cohérence globale des instruments.

Cependant, le premier critère (corrélation de l'instrument avec le régresseur endogène) ne doit pas être négligé, comme l'ont montré des études récentes. De fait, la recherche d'instruments dont l'exogénéité ne puisse être mise en cause les rend parfois faiblement corrélés avec le traitement. On parle dans ce cas d'instruments faibles. Dans ce cas, les propriétés des estimateurs obtenus ne sont pas satisfaisantes. Les estimations risquent d'être très imprécises, voire plus biaisées que les estimateurs par moindres carrés ordinaires "naïfs". En outre, les tests classiques (intervalle de confiance, tests de significativité) sont erronés[7]. Il est donc indispensable pour juger de la qualité des résultats obtenus par variables instrumentales de toujours faire figurer les résultats détaillés de l'équation de première étape. On trouvera à ce sujet une discussion plus approfondie dans Givord (2010).

En résumé

La mise en œuvre d'une méthode instrumentale pour évaluer l'impact d'une mesure requiert de :

– disposer d'un instrument, c'est-à-dire d'une variable qui explique le fait que certains individus aient une plus grande chance d'être affectés par cette mesure que d'autres, mais qui n'a pas d'impact direct sur les résultats de cette mesure. Une connaissance approfondie de la législation et de ses variations, par exemple, peut permettre d'obtenir de telles variables ;

– s'assurer de la validité et de la force de cet instrument. La première de ces qualités n'est pas testable statistiquement mais nécessite une attention particulière. S'assurer de la seconde (corrélation entre l'instrument et la variable de traitement) exige de s'intéresser à la régression de première étape du traitement sur l'instrument. Des valeurs élevées de la statistique de Fisher sont nécessaires pour éviter les problèmes d'instruments faibles ;

– estimer : en pratique, une méthode de doubles moindres carrés peut être appliquée. Mieux vaut utiliser une procédure préprogrammée. Attention, les propriétés présentées ici n'ont été établies que pour des estimations linéaires (en général avec des revenus continus). Il est incorrect d'utiliser cette procédure lorsque la variable d'intérêt binaire est modélisée par une fonction paramétrique (logit ou probit par exemple) ;

– interpréter les résultats obtenus au regard, en particulier, du fait qu'il s'agit d'une estimation locale. L'estimateur des variables instrumentales mesure les effets du traitement en moyenne pour une population particulière, celle qui réagit à l'instrument. Ceci limite la validité externe de l'estimation qui en découle : il n'est pas certain que le résultat obtenu sur cette population particulière puisse être généralisé. Il n'existe pas de méthode systématique pour répondre à cette question. Il est cependant possible de déterminer la proportion de cette sous-population, des *compliers*.

Régression sur discontinuités

Les premières utilisations de techniques correspondant à une procédure de régression sur discontinuités (*Regression Discontinuity Design*, RDD) remontent aux travaux précurseurs de Thistlethwaite et Campbell (1960), en psychologie. Ces derniers étudient l'impact de la réussite scolaire sur la carrière future en comparant le devenir d'étudiants ayant raté de peu un examen et de ceux l'ayant juste réussi. Ce n'est cependant qu'à la fin des années 1990 que cette technique a été utilisée dans des études économiques, suite à des études influentes. Citons en particulier celles d'Angrist et Lavy (1999) sur l'impact de la taille des classes sur la réussite scolaire et de Black (1999) sur l'incidence de la carte scolaire sur les prix des logements, ainsi que l'article de Hahn *et alii* (2001), qui précise rigoureusement les conditions d'identification nécessaires à l'utilisation de cette technique. On trouvera une présentation détaillée de ces méthodes dans Imbens et Lemieux (2008), dont s'inspire en grande part cette partie.

Les raisons du succès de cette technique tiennent à sa simplicité. La méthode repose sur l'existence d'une variable de sélection (notée *S*) ayant un impact discontinu sur la probabilité d'être traité. Plus précisément, il s'agit d'exploiter le fait que de nombreux dispositifs institutionnels présentent des règles d'affectation "à seuils", par exemple : allocation sous condition de ressources, contrainte d'âge pour bénéficier de certains dispositifs, carte scolaire (dans certaines rues, des enfants dépendent d'écoles plus ou moins "cotées"), etc. L'intuition est qu'autour de ces seuils, les personnes sont "presque" identiques ; pourtant seules certaines d'entre elles

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

bénéficieront du dispositif à évaluer. Il est donc possible d'identifier l'effet causal du traitement en comparant ces deux types de personnes.

De manière plus formalisée, on s'intéresse encore à l'effet d'un traitement T_i sur une variable de revenu Y_i :

(13) $Y_i = \alpha + \Delta_i T_i + u_i$

Comme *supra*, $\Delta_i = Y_i^1 - Y_i^0$ représente l'effet du traitement pour un individu i, $\alpha = E(Y_i^0)$ la moyenne des revenus potentiels sans traitement et le terme d'erreur vaut $u_i = Y_i^0 - E(Y_i^0)$.

Deux cas de figure

En pratique, deux cas doivent être distingués : celui du *sharp design* et celui du *fuzzy design*.

Cas du sharp design

Le traitement T dépend de manière déterministe de la variable de sélection S :

(14) $T_i = T(S_i) = 1_{\{S_i \geq \underline{S}\}}$

C'est par exemple le cas de dispositifs dépendant de l'âge. Lalive (2008) s'intéresse ainsi aux indemnités chômage qui sont plus généreuses en Autriche lorsque la perte d'emploi a lieu après cinquante ans. Il étudie l'effet de ces indemnités sur le retour à l'emploi en tenant compte de cette discontinuité. La perte d'emploi présente de fait une discontinuité marquée après 50 ans.

Cas du fuzzy design

Dans un *fuzzy design*, la variable de sélection S affecte bien la probabilité d'être traité, mais l'affectation n'est plus parfaite. Cette fois, c'est seulement la probabilité d'être traité conditionnellement à la variable de sélection $P(T_i = 1 | S_i)$ qui présente une discontinuité en $\underline{S}$. Deux exemples sont devenus classiques :

– Van der Klaauw (2002) s'intéresse aux montants des bourses offertes aux étudiants par une grande université américaine. Dans ce système universitaire, les universités se font concurrence pour recruter les meilleurs étudiants et le montant des bourses qu'elles leur allouent constitue l'un des instruments de cette concurrence. Van der Klaauw s'intéresse à l'efficacité de cette bourse pour attirer les étudiants. L'estimation directe est difficile, les meilleurs étudiants étant susceptibles d'avoir également des offres intéressantes dans les universités concurrentes. Pour déterminer l'impact causal de ces bourses, Van der Klaauw utilise le fait qu'en pratique, l'université classe les étudiants dans quatre groupes selon leur score à un examen. Le montant des bourses proposées à un étudiant diffère selon le groupe auquel il appartient. Cela signifie que des étudiants dont le niveau scolaire est très proche mais qui se situent respectivement en dessous et au-dessus des seuils utilisés pour construire ces groupes peuvent se voir proposer des montants différents. L'allocation n'est cependant pas parfaite, puisque d'autres variables (revenus des parents, appartenance à une minorité, lettres de recommandation, etc.) interviennent également pour déterminer le montant des bourses. Van der Klaauw observe de fait une très grande dispersion du montant réel des bourses en fonction du niveau scolaire. En estimant une courbe d'interpolation polynomiale moyenne, il met cependant en évidence des sauts à chaque seuil ;

– Angrist et Lavy (1999) utilisent les limites existantes sur le nombre d'élèves par classe pour étudier l'effet de la taille des classes sur la réussite scolaire. Dans une école, les classes sont "remplies" jusqu'à un certain seuil à partir duquel une nouvelle classe est créée, ce qui réduit brusquement le nombre moyen d'élèves par classe dans cette école. En Israël, la règle des *Maimonides* prévoit que le nombre d'élèves par classe ne doit pas dépasser quarante. Selon le nombre d'élèves dans l'école, le nombre d'élèves par classe peut être très différent. Par exemple, une école avec 38 élèves à un certain niveau scolaire n'aura théoriquement qu'une classe très nombreuse, tandis qu'une école avec 42 élèves devrait avoir deux "petites" classes. En pratique, cette règle n'est pas parfaitement appliquée (il est en particulier fréquent qu'une nouvelle classe soit créée avant que la limite soit atteinte), ce qui explique qu'on se trouve dans un *fuzzy design*. Cette particuliarité se retrouve évidemment dans de nombreux pays. Piketty et Valdenaire (2006) l'utilisent pour la France.

Identification

Dans quelle mesure pouvons-nous identifier un effet causal dans ces modèles ? Comme on va le voir, les hypothèses sont assez générales, surtout dans le cadre d'un *sharp design*.

Identification dans le cas d'un sharp design

Le principe de la régression sur discontinuité est qu'au point de discontinuité, on passe "brutalement" du revenu potentiel sans traitement Y^0 au revenu potentiel avec traitement Y^1, pour reprendre la terminologie des revenus potentiels. La discontinuité observée permet de mesurer l'effet du traitement à la condition que les revenus potentiels ne présentent pas eux-mêmes de discontinuités à ce point (ce serait le cas si un autre déterminant du revenu dépendait de manière discontinue de la variable de sélection). Dit autrement, il est nécessaire que les personnes juste en dessous ou juste au-dessus du seuil $\underline{S}$ soient vraiment comparables. Plus formellement, cela signifie que la composante inobservée du revenu $E(u_i | S_i)$ est continue en $\underline{S}$, ainsi que l'effet moyen du traitement $E(\Delta_i | S_i)$.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

L'effet du traitement au point de discontinuité correspond alors simplement à :

$$(15)\quad E(\Delta_i \mid S_i = \underline{S}) = \lim_{s \downarrow \underline{S}^+} E(Y_i \mid S_i = s) - \lim_{s \uparrow \underline{S}^-} E(Y_i \mid S_i = s)$$

car :

$$(16)\quad \lim_{s \downarrow \underline{S}^+} E(u_i \mid S_i = s) = \lim_{s \uparrow \underline{S}^-} E(u_i \mid S_i = s)$$

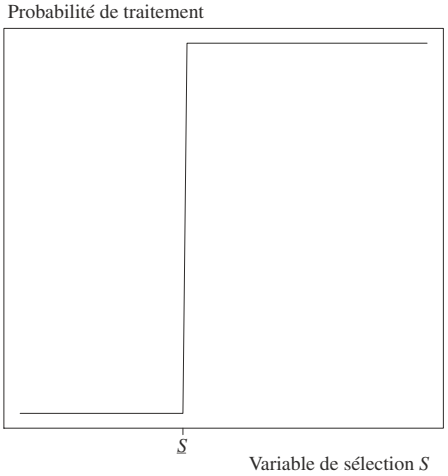
et :

$$(17)\quad \lim_{s \downarrow \underline{S}^+} E(\Delta_i T_i \mid S_i = s) = E(\Delta_i \mid S_i = \underline{S})$$

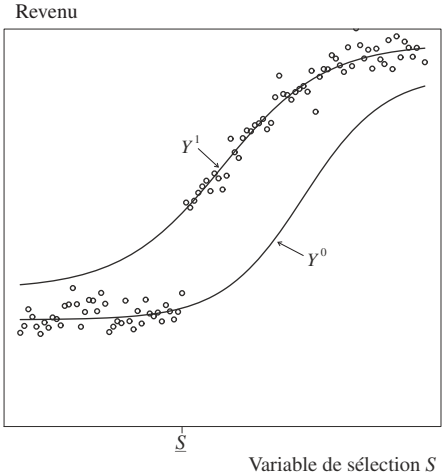
Il faut bien noter qu'on n'estime qu'un effet local du traitement, au point de discontinuité. Si le traitement n'est pas constant dans la population, l'interprétation de cette estimation sera limitée.

Figures 2 : traitement et revenus potentiels et observé en fonction de la variable de sélection *S*, *sharp design*

2.1 Probabilité de traitement



2.2 Revenus potentiels et observé



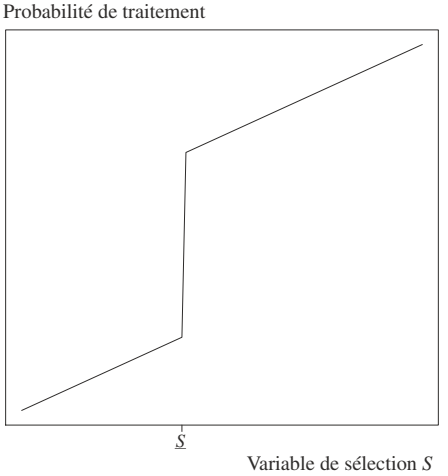
Lecture : sur cet exemple simulé et inspiré d'Imbens et Lemieux (2008), le traitement dépend d'une manière déterministe de la variable de sélection (figure 2.1), $T = 1_{\{S \geq \underline{S}\}}$. Pour toutes les personnes dont la valeur de cette variable de sélection est inférieure au point de discontinuité $\underline{S}$, on observe le revenu potentiel sans traitement, tandis qu'à l'inverse on observe le revenu potentiel avec traitement pour les autres (figure 2.2).

Source : auteure, d'après Imbens et Lemieux (2008).

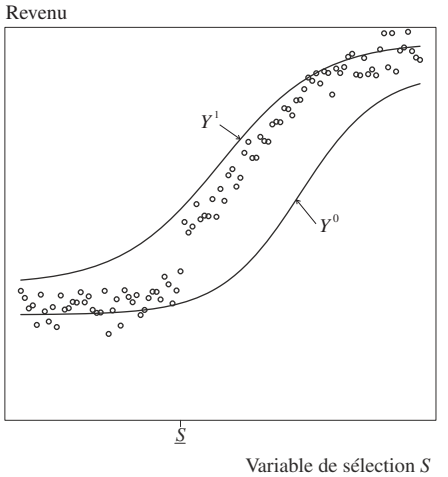
Les figures 2 présentent une illustration, à partir de données simulées, d'une situation où l'on pourrait utiliser un *sharp design*. La figure 2.1 représente la répartition de la probabilité du traitement en fonction d'une variable de sélection *S*. Aucune personne qui présente une valeur de la variable de sélection *S* inférieure à ce seuil $\underline{S}$ ne bénéficie du traitement, alors qu'en revanche toutes celles qui ont une valeur *S* supérieure à $\underline{S}$ en bénéficient. La figure 2.2 correspond aux revenus : les lignes pleines correspondent aux revenus potentiels (conditionnellement à une valeur de *S*) avec et sans traitement. Sur cet exemple, ces deux revenus potentiels moyens sont différents et varient de manière continue en fonction de la variable de sélection *S*. Rappelons que, pour une valeur particulière de la variable de sélection *s*, l'écart vertical entre ces deux courbes représente l'effet moyen du traitement pour les personnes dont la

Figures 3 : traitement et revenus potentiels et observé en fonction de la variable de sélection *S*, *fuzzy design*

3.1 Probabilité de traitement



3.2 Revenus potentiels et observé



Lecture : sur cet exemple simulé et inspiré d'Imbens et Lemieux (2008), la probabilité de traitement dépend de manière discontinue de la variable de sélection (figure 3.1). Pour chaque valeur *s* de la variable de sélection *S*, on observe à la fois des personnes traitées et non traitées : le revenu moyen conditionnel à cette valeur *s* se trouve entre les deux courbes de revenus potentiels (figure 3.2). La discontinuité se retrouve pour les revenus observés, la proportion de personnes traitées augmentant brusquement au point de discontinuité.

Source : auteure, d'après Imbens et Lemieux (2008).

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

valeur de sélection vaut s. Du fait de la règle de sélection dans le traitement (figure 2.1), le revenu observé correspond à gauche du seuil au revenu potentiel sans traitement, à droite du seuil au revenu potentiel avec traitement. Ceci aboutit à un saut au niveau du revenu observé (représenté par des points sur le graphique). Ce saut correspond à l'effet moyen du traitement pour les personnes dont la valeur de sélection est $\underline{S}$.

***Identification dans le cas d'un* fuzzy design**

L'hypothèse de continuité des revenus potentiels au voisinage du seuil $\underline{S}$ est-elle suffisante dans le cas d'un *fuzzy design* ? Oui si l'effet du traitement est constant au voisinage de $\underline{S}$.

D'après :

$$(18)\ \lim_{s \downarrow \underline{S}^+} E(Y_i | S_i = s) - \lim_{s \uparrow \underline{S}^-} E(Y_i | S_i = s)$$
$$= \lim_{s \downarrow \underline{S}^+} E(\Delta_i T_i | S_i = s) - \lim_{s \uparrow \underline{S}^-} E(\Delta_i T_i | S_i = s)$$
$$+ \lim_{s \downarrow \underline{S}^+} E(u_i | S_i = s) - \lim_{s \uparrow \underline{S}^-} E(u_i | S_i = s)$$

si l'effet du traitement Δ_i est constant (et donc non aléatoire) au voisinage du point de discontinuité, il est donné par :

$$(19)\ \Delta = \frac{\lim_{s \downarrow \underline{S}^+} E(Y | S = s) - \lim_{s \uparrow \underline{S}^-} E(Y | S = s)}{\lim_{s \downarrow \underline{S}^+} E(T | S = s) - \lim_{s \uparrow \underline{S}^-} E(T | S = s)}$$

Cependant, si l'effet du traitement varie selon les individus, il faut faire une hypothèse supplémentaire.

La première peut être l'indépendance "locale" du traitement (Hahn *et alii*, 2001). Conditionnellement à S_i, T_i est indépendant de Δ_i autour de $\underline{S}$:

$$(20)\ T_i \perp \Delta_i | S_i$$

Le ratio qui constitue le terme de droite de (19) est alors égal à $E(\Delta_i | S = \underline{S})$, l'effet local du traitement.

Si l'on n'est pas prêt à faire l'hypothèse (20), Hahn *et alii* (2001) montrent qu'une hypothèse moins restrictive suffit. Il s'agit d'une hypothèse de monotonicité : $T_i(S)$ est non décroissante en S, pour tout individu i, autour du point de discontinuité.

Cette hypothèse de monotonicité est directement l'analogue de celle utilisée par Angrist *et alii* (1996) pour les variables instrumentales. Elle signifie qu'il n'existe pas de *defiers*, *i.e.* de personnes qui choisissent le traitement si elles sont en dessous du seuil et le refusent si elles sont au-dessus. En revanche, on n'exclut pas l'existence de personnes qui sont systématiquement traitées ou non traitées (*always takers* ou *never takers* dans les termes d'Angrist *et alii*, 1996).

Que représente alors le ratio constituant le terme de droite de (19) ? Ici aussi, il s'agit d'un effet local (*LATE*) :

$$(21)\ \lim_{e \downarrow 0^+} E\left(\Delta_i | T_i(\underline{S}+e) - T_i(\underline{S}-e) = 1\right)$$

Plus précisement, il s'agit de l'effet du traitement pour les *compliers*, ceux pour qui le fait d'être en dessous ou au-dessus du seuil fait une différence dans l'acceptation du traitement.

Dans l'exemple présenté par la figure 3, la probabilité de traitement augmente régulièrement avec la variable de sélection (figure 3.1). Elle présente une augmentation brusque au point de discontinuité $\underline{S}$. Cette augmentation est cependant moins nette que dans le cas du *sharp design* : la probabilité ne passe pas de zéro à un. Certaines personnes caractérisées par une valeur de la variable de sélection inférieure à ce seuil bénéficient néanmoins du traitement. Inversement, des personnes associées à une valeur de S supérieure à ce seuil n'en bénéficient pourtant pas. Les revenus observés s'interprètent de ce fait moins simplement que dans le cas du *sharp design*. Il s'agit, aussi bien à droite et à gauche du point de discontinuité, d'une moyenne de ces revenus potentiels, pondérés par la probabilité de traitement. Même au niveau du point de sélection, les différences constatées entre les revenus observés ne correspondent pas aux différences entre les revenus potentiels avec et sans traitement (figure 3.2).

Manipulation du seuil

Attention, pour que l'hypothèse de continuité des revenus potentiels au voisinage du seuil $\underline{S}$ soit valide, il faut que les individus ne puissent pas "manipuler" la valeur du seuil de manière à disposer ou non du traitement. Sinon, l'hypothèse de continuité a peu de chances d'être respectée (les individus ne sont plus comparables aux voisinages inférieur et supérieur du seuil).

"Tests" graphiques

L'identification de l'effet du traitement repose sur deux hypothèses essentielles : l'existence d'une variable de sélection pour laquelle la probabilité de traitement est discontinue et la continuité des autres composantes du revenu en ce point.

Même si elle n'a pas statut de validation, une analyse graphique peut permettre de vérifier que la stratégie est bien adaptée. Plus exactement, on sera amené à vérifier plusieurs points :

– la variable de traitement présente bien une discontinuité au point prévu (si ce n'est pas le cas, par exemple parce que le dispositif institutionnel "officiel" n'est finalement pas appliqué en pratique, il est inutile de poursuivre) ;

– la variable de revenu présente une discontinuité au même point. Il est également plus rassurant qu'elle

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

ne présente pas d'autres sauts d'ampleur comparable en dehors de ceux correspondant aux discontinuités de la variable de traitement. Dans le cas contraire, s'il n'existe pas de justification plausible, on peut craindre qu'une discontinuité du revenu ne résulte pas uniquement de la discontinuité du traitement. L'estimateur fournira alors une mesure biaisée de l'effet causal ;

– il est également judicieux de vérifier que les autres déterminants éventuels du revenu ne présentent pas de discontinuité au point $\underline{S}$.

Estimation

L'esprit de la méthode de régression sur discontinuités requiert d'estimer des valeurs locales au point de discontinuité : les propriétés ne sont établies que localement. Cela suppose en toute rigueur de se concentrer sur les observations dont les valeurs de la variable de sélection sont proches de ce point de discontinuité. Cependant, les données disponibles sont parfois trop dispersées ou pas assez nombreuses pour permettre de se limiter au strict voisinage autour de ce point. On privilégiera une estimation semi-paramétrique, qui permet de tenir compte du fait que les composantes inobservées du revenu peuvent cependant varier en fonction de la variable de sélection en utilisant une forme polynomiale dans la variable de sélection.

Sous sa forme la plus simple (plus de précisions dans Givord, 2010) il s'agit, pour le *sharp design*, de mettre en œuvre une régression linéaire estimée par les moindres carrés ordinaires en se restreignant aux observations au voisinage immédiat du point de discontinuité $[\underline{S}-h, \underline{S}+h]$, où h est la taille du voisinage (fenêtre) :

$$Y_i = a + \Delta 1_{\{S_i \geq \underline{S}\}} + b_1(S_i - \underline{S}) + b_2(S_i - \underline{S})1_{\{S_i \geq \underline{S}\}} + b_3(S_i - \underline{S})^2 + b_4(S_i - \underline{S})^2 1_{\{S_i \geq \underline{S}\}} + \ldots + u_i$$

Δ étant le paramètre d'intérêt.

Dans le cadre d'un *fuzzy design*, on peut montrer que l'estimateur des régressions sur discontinuités correspond exactement à celui obtenu par une procédure de doubles moindres carrés où l'on instrumente la variable de traitement T_i par le seuil $1_{\{S_i \geq \underline{S}\}}$, en utilisant comme variables exogènes supplémentaires des fonctions polynomiales de la distance au seuil.

Ces estimations appellent plusieurs choix techniques : le degré des polynômes et la taille du voisinage autour du point de discontinuités. Ces choix seront, en partie, liés aux nombres d'observations dont on dispose, mais il n'existe pas de cadre bien établi pour les déterminer. En pratique, on pourra tester la sensibilité des résultats obtenus aux choix retenus. Il existe quelques méthodes pour déterminer la taille de la fenêtre optimale (Imbens et Kalyanaraman, 2012). Cependant, elles sont complexes et rarement mises en œuvre dans les articles appliqués[8]. Lorsque l'on utilise un échantillon très large, les estimations risquent d'être moins robustes, puisque plus sensibles à la pertinence de la spécification paramétrique retenue pour contrôler la dépendance du revenu à la variable de sélection. Black *et alii* (2007) comparent ainsi les résultats obtenus par une affectation aléatoire de personnes sans emploi à un dispositif d'aide à l'emploi aux résultats qui seraient obtenus par une procédure de régression sur discontinuités. Ils montrent que la méthode de régression sur discontinuités fournit des résultats assez convaincants, au sens où elle permet de bien reproduire les résultats de l'expérience aléatoire, *à condition de se limiter aux valeurs proches* des points de discontinuités. Les estimations qui utilisent plus de points et reposent sur une spécification paramétrique pour contrôler le fait que le revenu dépend aussi de la variable de sélection se révèlent au contraire peu robustes. Conformément à l'intuition, elles sont sensibles à la formalisation retenue, qui peut s'éloigner de la "vraie" fonction inconnue que l'on cherche à approcher.

Variable instrumentale ou régressions sur discontinuités ?

Pour évaluer un dispositif, plusieurs méthodes peuvent être mobilisées. Comme on vient de le voir, les techniques de régressions sur discontinuités sont susceptibles d'être utilisées lorsque le dispositif présente des seuils dans les modalités d'attribution. On peut cependant s'interroger sur l'intérêt de recourir à d'autres méthodes dans ce cas. En particulier, il serait envisageable d'utiliser la variable de sélection comme instrument : typiquement, $1_{\{S_i \geq \underline{S}\}}$ serait utilisable comme tel. De fait, comme on l'a vu, les estimateurs auxquels on a recours dans un cadre de régression sur discontinuités ressemblent beaucoup "techniquement" à des estimateurs par variables instrumentales, au sens où ils peuvent être calculés par des procédures de type doubles moindres carrés. Dans la littérature empirique, la frontière est assez floue entre les deux méthodes. On appelle aujourd'hui *régressions sur discontinuités* des méthodes d'identification qui auraient été caractérisées comme des approches de type variables instrumentales il y a quelques années. La différence essentielle entre les deux tient à l'hypothèse sous laquelle on identifie un effet causal : même si l'estimateur est parfois le même, la régression sur discontinuités n'est clairement pas un cadre où les variables instrumentales sont indiquées. On ne fait pas *a priori* l'hypothèse que les revenus potentiels sont indépendants de la variable de sélection. Dit autrement, les hypothèses nécessaires pour justifier que la procédure proposée mesure bien un effet causal sont moins fortes que dans le cas d'un instrument. La méthode des régressions sur discontinuités permet également d'insister sur le fait

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

que l'estimateur obtenu est encore plus local. Dans le cadre d'un *fuzzy design*, non seulement on n'identifie l'effet que sur les personnes pour lesquelles la variable de sélection a un impact (ce sont les mêmes *compliers* que dans le cadre de variables instrumentales) mais, en outre, on identifie l'effet seulement pour une valeur particulière des revenus (celle qui correspond au point de discontinuité).

La technique des régressions sur discontinuités a connu un rapide succès depuis le milieu des années 2000. Celui-ci s'explique, outre par la simplicité de l'approche, par sa forte validité interne. Les régressions sur discontinuités reproduisent très bien les données d'une expérience aléatoire contrôlée. Néanmoins, cela a un prix : les estimations ne sont valides que localement. À moins d'avoir de bonnes raisons de penser que l'effet est constant sur la population, les résultats sont donc difficilement généralisables. Dans les meilleurs des cas, on dispose de plusieurs points de discontinuités, ce qui permet d'estimer l'effet sur plusieurs points.

En résumé

Pour résumé, une procédure de régression sur discontinuités peut être utilisée pour évaluer une procédure ou une mesure dont le mode de désignation des bénéficiaires présente une discontinuité. Les principales étapes peuvent se décliner comme précédemment :

– disposer de données détaillées et d'échantillons suffisamment fournis pour pouvoir se restreindre aux observations proches du point de discontinuité ;

– vérifier la pertinence de la procédure. Ceci peut se faire par exemple graphiquement. Il faut que la probabilité de bénéficier de la mesure présente une discontinuité visible dans la variable de sélection. Il est également utile de vérifier que le revenu étudié est continu en dehors des points de discontinuités utilisés ;

– estimer : en pratique, on privilégiera les estimations locales. En d'autres termes, on se restreindra aux observations dont les valeurs se situent au voisinage du point de discontinuité. La manière la plus simple consiste à se limiter aux observations autour de la discontinuité et d'effectuer une régression linéaire sur ce sous-échantillon (simple régression linéaire pour un *sharp design*, doubles moindres carrés pour un *fuzzy design*). Idéalement, on peut choisir la taille du voisinage optimale (voir Imbens et Kalyanaraman, 2012). En pratique, l'immense majorité des études définit un voisinage de manière *ad hoc*, quitte à vérifier la sensibilité des résultats en faisant varier ce voisinage. Si on en élargit la taille, il faut tenir compte de la dépendance selon le revenu : les estimations risquent d'être moins robustes car sensibles à la spécification retenue ;

– interpréter : il s'agit d'un effet local, obtenu pour les individus ayant une valeur particulière au regard de la variable de sélection. Savoir si ces résultats sont généralisables à l'ensemble de la population demande, comme souvent, une discussion au cas par cas.

Méthodes de sélection sur observables (dont méthodes d'appariement)

Les différentes méthodes décrites jusqu'à présent nécessitent de disposer de caractéristiques particulières (disponibilité d'un groupe de contrôle pour appliquer des différences de différences, d'un instrument pour recourir à la technique des variables instrumentales, processus ayant une discontinuité pour permettre l'utilisation de la technique de régression sur discontinuités). Ces méthodes ne peuvent être utilisées qu'au cas par cas. Lorsque les conditions de validité d'aucune de ces méthodes n'est applicable et qu'il existe de la sélection probable dans les dispositifs, on peut corriger, au moins en partie, cet effet de sélection en contrôlant les différences observables entre les bénéficiaires et les non bénéficiaires. Les méthodes de sélection sur observables, dont les méthodes d'appariement (ou *matching*), ont été développées dans ce but. Elles ne permettent d'identifier l'effet causal de la mesure qu'on cherche à évaluer que sous l'hypothèse très forte que les différences observables entre les bénéficiaires et les non bénéficiaires captent l'ensemble des déterminants de la sélection des bénéficiaires (pour une présentation générale, voir également Brodaty *et alii*, 2007).

Formellement, l'hypothèse identifiante pour évaluer cet effet causal est que, conditionnellement à des variables observables X, l'affectation au traitement est indépendante des revenus escomptés :

$$(22)\ \ Y_i^0, Y_i^1 \perp T_i | X_i$$

Cette hypothèse est souvent appelée CIA, pour *Conditional Independence Assumption* (hypothèse d'indépendance conditionnelle). Supposons que l'on observe deux individus présentant des caractéristiques observables identiques, mais dont l'un est traité et l'autre pas. L'hypothèse CIA signifie que le fait qu'un individu soit traité et pas un autre n'est pas dû à des différences escomptées dans les revenus potentiels. Le revenu de la personne non traitée constitue donc un bon contrefactuel du revenu de la personne traitée si cette dernière n'avait pas été traitée, et inversement. La comparaison des deux permet donc d'obtenir un estimateur non biaisé de

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

l'effet du traitement (conditionnel à ces observables).

Si l'on ne souhaite évaluer que l'effet de la mesure sur les seuls bénéficiaires, on peut se contenter d'une hypothèse un peu moins forte. En effet, il "suffit" alors que :

(23) $Y_i^0 \perp T_i | X_i$

c'est-à-dire que, conditionnellement à ces caractéristiques observables, le revenu des personnes qui n'ont pas bénéficié de la mesure fournisse un bon contrefactuel du revenu potentiel des bénéficiaires si ceux-ci n'en avaient pas bénéficié.

Par exemple, pour évaluer l'impact du passage en Zone Franche Urbaine (ZFU) sur l'emploi et la création d'entreprises, Rathelot et Sillard (2009) se réfèrent aux critères utilisés pour classer ces zones de la politique de la ville. Ces zones étant par définition moins propices au développement économique, comparer directement l'évolution de l'emploi (par exemple) avec celle observée dans d'autres quartiers risque de donner une évaluation biaisée. En fait, les ZFU ont été définies à partir de certaines caractéristiques socio-économiques (taux de chômage, potentiel fiscal de la commune, etc.), mais également sur des considérations locales laissées à l'appréciation du préfet. Pour réduire ce biais, Rathelot et Sillard (2009) comparent les évolutions des zones avant leur passage en ZFU avec celles d'autres quartiers en difficulté présentant des caractéristiques socio-économiques comparables. L'hypothèse est que le fait que deux quartiers présentant des caractéristiques proches, mais dont l'un a bénéficié du statut de zones franches et pas l'autre, n'est pas lié à l'anticipation des effets escomptés du statut de ZFU.

Cette hypothèse identifiante est très forte. Elle signifie qu'en dehors des observables il n'existe pas d'autres caractéristiques qui influent à la fois sur les revenus potentiels et le choix du traitement. Le risque est donc grand qu'elle ne soit pas vérifiée. Néanmoins, il s'agit souvent de la seule hypothèse possible, ce qui explique que les méthodes s'appuyant sur cette dernière sont très répandues.

Par ailleurs, ces méthodes ne peuvent être appliquées que dans certains cas : pour pouvoir comparer un individu traité avec un individu non traité identique... il faut que ce dernier existe ! Si, pour certaines caractéristiques X, les individus sont assurés d'être traités, il ne sera pas possible de trouver des individus non traités comparables. Dans l'exemple des Zones Franches Urbaines ci-dessus, c'est le fait que la création des zones ne soit pas déterminée exactement par les caractéristiques socio-démographiques qui permet d'utiliser ces dernières comme variables de conditionnement. Si cela n'était pas le cas, il ne serait pas possible de trouver des quartiers comparables aux zones franches. Formellement, cela suppose que les méthodes d'appariement ne sont valables que sous l'hypothèse de l'existence d'un support commun, c'est-à-dire que, pour toutes les valeurs des observables, on puisse comparer des individus traités et non traités :

(24) $0 < P(T_i = 1 \mid X_i) < 1$

En pratique, sous l'hypothèse d'indépendance conditionnelle, l'effet peut être simplement estimé par une régression linéaire. Toutefois, de nombreuses autres méthodes se sont développées sur les deux dernières décennies. Les plus courantes sont les méthodes d'appariement (ou *matching*), qui tentent d'apparier chaque bénéficiaire à un non bénéficiaire ayant exactement les mêmes caractéristiques observables. Il n'existe cependant pas encore de cadre commun à l'ensemble de ces méthodes, qui se présente plus comme une collection de techniques que comme une théorie clairement édifiée et unifiée.

Limites de la régression linéaire

La méthode la plus simple sous l'hypothèse d'indépendance conditionnelle du traitement est la régression linéaire. Supposons par exemple que la distribution du revenu potentiel sans traitement conditionnelle aux caractéristiques observables X soit linéaire :

(25) $E(Y_i^0 | X_i) = \alpha + X_i' \beta$

où X_i' et β sont des vecteurs colonne de dimension K et X_i' est la transposée de X_i.

Si l'on suppose également que l'effet de la mesure $Y_i^1 - Y_i^0$ est constant (noté comme précédemment $\Delta, \Delta = E(Y_i^1 - Y_i^0)$), le revenu observé s'écrit :

(26) $Y_i = \alpha + \Delta T_i + X_i' \beta + u_i$

avec $u_i = Y_i^0 - E(Y_i^0 | X_i)$. On peut estimer l'effet du traitement Δ par les moindres carrés ordinaires dans une régression "contrôlant" l'effet des variables observables. Notons aussi qu'il est possible de relâcher l'hypothèse d'un effet constant de la mesure. On peut supposer que l'effet dépend des caractéristiques observables en ajoutant au modèle des termes croisés entre le traitement et ces variables observables. Le calcul de la précision de ces estimateurs est standard.

La régression linéaire présente des avantages indéniables. Elle repose sur un fondement théorique clair et son application pratique est bien balisée, tant pour l'estimation que pour l'inférence statistique. Même si la distribution conditionnelle des revenus

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

n'est pas exactement une fonction linéaire des caractéristiques observables, la régression linéaire fournit la meilleure approximation *linéaire* de la relation entre les deux.

La simplicité du modèle linéaire constitue également sa limite. Si la distribution conditionnelle s'éloigne trop d'une fonction linéaire, même la meilleure approximation linéaire restera peu adaptée et les estimateurs risquent d'être peu robustes.

Le problème est particulièrement aigu lorsque les deux échantillons (traitement et contrôle) sont très différents du point de vue des variables de contrôle, car c'est alors que la linéarité de l'impact de ces variables sur le résultat devient une hypothèse forte.

Pour mieux comprendre cela, rappelons que notre problème est d'estimer un contrefactuel du revenu sans traitement pour les bénéficiaires de la mesure. Si l'hypothèse d'indépendance conditionnelle est correcte, la meilleure estimation linéaire de ce contrefactuel est fournie par le revenu moyen du groupe de contrôle, corrigé des différences de composition entre les deux échantillons :

$$(27)\quad \hat{E}(Y_i^0 \mid T_i = 1) = \overline{Y}^0 + (\overline{X}^1 - \overline{X}^0)'\hat{\beta}$$

où $\overline{Y}^j$ et $\overline{X}^j$ sont les moyennes empiriques respectivement de Y et X dans chacun des deux échantillons $j = 0$ et 1.

L'estimateur de l'effet du traitement correspond simplement à :

$$(28)\quad \hat{\Delta} = \overline{Y}^1 - \overline{Y}^0 - (\overline{X}^1 - \overline{X}^0)'\hat{\beta}$$

Si les différences $\overline{X}^1 - \overline{X}^0$ entre les deux échantillons sont importantes, la correction le sera très probablement aussi et sera très sensible à des modifications même mineures de la spécification.

Imbens et Wooldridge (2009) suggèrent de mesurer la différence de composition à partir des K différences normalisées :

$$(29)\quad \Delta_X^k = \frac{\overline{X}_k^1 - \overline{X}_k^0}{\sqrt{V_k^1 + V_k^0}}, \qquad \forall k = 1, \ldots, K,$$

où $\overline{X}_k^j$ et $V_k^j = \sum_{i:T_i = j} (X_{ki} - \overline{X}_k^j)^2 / (N^j - 1)$ sont les moyenne et variance empiriques de la variable observable X_k dans l'échantillon j, N^j étant la taille de ce dernier[9].

Comme règle pratique, on considère en général que les estimations risquent d'être sensibles à la spécification dès lors que $\min_k \Delta_X^k > 0{,}25$ (Imbens et Wooldridge, 2009).

On trouvera une illustration de cet effet dans la critique de LaLonde (1986). Dans un article célèbre, celui-ci s'interroge sur la validité des méthodes non expérimentales pour évaluer l'effet de politiques publiques. Pour cela, il estime de plusieurs manières l'impact d'un programme expérimental de formation des chômeurs sur le salaire des personnes en ayant bénéficié. Pour estimer l'impact de ce stage, ces dernières ont été désignées de manière aléatoire parmi un groupe de chômeurs. Pour permettre l'évaluation, des données ont également été collectées sur les salaires des personnes n'ayant pas été retenues pour participer à cette formation. L'assignation étant aléatoire, la comparaison des revenus des bénéficiaires et des non bénéficiaires du stage permet d'obtenir l'effet "réel" de cette formation. LaLonde (1986) compare alors cette "vraie" valeur à celles obtenues en utilisant non les données sur le groupe de contrôle expérimental, mais des données administratives ou d'enquêtes (c'est-à-dire des données utilisées lorsque l'on ne dispose pas de données expérimentales). Il estime donc l'impact de la formation en contrôlant les caractéristiques observables pouvant affecter le revenu (sexe, âge, race, ancienneté dans le chômage et revenu passé). Les résultats de LaLonde (1986) font apparaître une très forte variabilité des estimations aux données et aux spécifications utilisées, notamment aux variables explicatives retenues pour contrôler les différences de composition dans l'échantillon. Ce n'est évidemment pas le cas lorsqu'il utilise le groupe de contrôle expérimental (pour une comparaison plus récente, voir l'article de Bléhaut et Rathelot dans ce numéro).

Cependant, comme le soulignent Imbens et Wooldridge (2009), les échantillons du groupe témoin (ayant reçu la formation) et du groupe de contrôle non expérimental sont très différents. LaLonde (1986) s'appuyant uniquement sur des spécifications linéaires, la variabilité qu'il observe est liée en grande partie à ces écarts de composition. S'appuyant sur les mêmes données que LaLonde (1986), Dehejia et Wahba (1999) obtiennent des résultats plus satisfaisants en utilisant une méthode d'appariement (*matching*).

L'appariement (ou *matching*)

Les méthodes d'appariement ont été très utilisées dans les deux dernières décennies. Il est nécessaire de les évoquer au pluriel, tant sont multiples les méthodes et variantes.

La méthode la plus simple consiste à comparer chaque individu avec son "jumeau". Ceci revient, pour évaluer l'effet moyen sur les individus traités, à comparer le revenu Y de chaque bénéficiaire avec celui d'un non bénéficiaire ayant exactement les mêmes caractéristiques observables. Pour un

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

bénéficiaire donné *i*, il est rare de trouver un individu du groupe de contrôle qui lui soit exactement identique. Par conséquent, on choisit son "plus proche voisin" *i'*. Cette notion nécessite de définir une métrique. À la simple métrique euclidienne (définissant la distance entre deux individus comme la somme de leurs écarts sur toutes les caractéristiques observables *X*), on préfère en pratique une métrique utilisant des variables normalisées. La plus courante est la distance de Mahalanobis, qui pondère par la matrice de variances-covariances Σ des variables *X* :

$$(30)\quad d(X_i, X_{i'}) = (X_i - X_{i'})' \Sigma^{-1} (X_i - X_{i'})$$

ou au moins sa version diagonale, où chaque covariable est pondérée par l'inverse de son écart-type.

Pour estimer l'effet moyen de la mesure sur les individus traités, on apparie chaque personne de l'échantillon des bénéficiaires avec son plus proche voisin dans l'échantillon des non bénéficiaires[10]. L'effet moyen estimé sur les traités $E(Y_i^1 - Y_i^0 \mid T_i = 1)$ correspond simplement à la moyenne de ces effets estimés :

$$(31)\quad \hat{\delta} = \frac{1}{N^1} \sum_{E^1} (Y_i^1 - \hat{Y}_i^0)$$

où N^1 est l'effectif du groupe de bénéficiaires E^1, Y_i^1 le revenu observé du bénéficiaire *i* et $\hat{Y}_i^0$ le revenu du plus proche voisin de *i* dans le groupe de contrôle.

L'appariement peut être fait sans remise (un individu du groupe de contrôle ne peut être apparié qu'une fois avec un individu du groupe de traitement) ou avec remise (on utilise l'ensemble de l'échantillon à chaque fois, ce qui autorise plusieurs appariements avec le même individu). L'appariement sans remise suppose que l'on dispose d'un large échantillon de contrôle. L'un de ses inconvénients est que l'estimation peut être sensible à l'ordre dans lequel l'appariement est effectué.

L'appariement avec le plus proche voisin est sans doute l'un des estimateurs de *matching* les plus utilisés. Il est assez simple à mettre en œuvre, son principe est intuitif et, contrairement aux variantes que nous allons évoquer, il ne requiert pas un choix de paramètres. Deux critiques peuvent toutefois lui être adressées. La première est qu'on ne contrôle pas la qualité de l'appariement. Or, la notion de plus proche voisin est par nature relative. Certains "plus proches" voisins peuvent être en fait éloignés. Or, la méthode du plus proche voisin traite de la même manière des couples proches et moins proches. La seconde critique est qu'apparier avec un seul individu prive de l'information apportée par tous les autres, ce qui réduit *a priori* la précision de l'estimation. Par exemple, certains bénéficiaires peuvent avoir plusieurs jumeaux très proches. On peut considérer qu'il est dommage d'en choisir (plus ou moins arbitrairement) un seul.

Il existe donc plusieurs variantes, qui tiennent à la manière dont on construit le contrefactuel de chaque bénéficiaire. En pratique, il s'agit donc d'estimer $\hat{Y}_i^0$ en utilisant un plus ou moins grand nombre d'individus du groupe de contrôle :

– *plus proches voisins* : plutôt que se limiter au plus proche voisin, on choisit d'apparier tout bénéficiaire avec un nombre fixe *M* de plus proches voisins. Le revenu contrefactuel du bénéficiaire *i* est alors simplement la moyenne des revenus de ces *M* voisins. Une autre variante consiste à exclure les couples trop éloignés : on n'utilise pas les bénéficiaires pour lesquels on ne peut pas trouver un (ou *M*) jumeau(x) à moins d'une certaine distance *d* à fixer ;

– *Radius* ou *caliper* : on sélectionne toutes les personnes *i'* du groupe de contrôle situées dans un proche voisinage fixé (*i.e.* telles que $\|X_i - X_{i'}\| < h$, pour un certain rayon de voisinage *h* ;

– *Kernel* ou noyau : ces méthodes vont au bout de cette logique. Le contrefactuel de l'individu *i* est calculé par une estimation à noyau (*Kernel*). Tous les individus du groupe de contrôle sont utilisés, mais pondérés par leur distance à l'individu traité *i* :

$$(32)\quad \hat{Y}_i^0 = \frac{\sum_{i' \in E^0} \mathrm{K}\left(\frac{\|X_i - X_{i'}\|}{h}\right) Y_k}{\sum_{i' \in E^0} \mathrm{K}\left(\frac{\|X_i - X_{i'}\|}{h}\right)}$$

où K est le noyau utilisé[11]. On utilise souvent un noyau gaussien, ou d'Eparechnikov. *h* est la taille de la fenêtre (*bandwidth*) du noyau. Les fonctions à noyau classique sont souvent rapidement décroissantes. La fenêtre mesure donc la taille du voisinage en dehors duquel les poids sont très faibles. Plus la fenêtre est petite, plus l'estimation du contrefactuel d'un bénéficiaire se concentre sur les personnes du groupe de contrôle dont les caractéristiques observables sont les plus proches de ce bénéficiaire. Il n'existe pas vraiment de règle établie pour le choix de cette fenêtre, dans le cas de ces méthodes d'appariement. En pratique, le choix est souvent fait de manière *ad hoc* ou à partir de "règles du pouce" classiques pour les estimations fonctionnelles semi-paramétriques, dont la validité n'a pas été démontrée dans le cas des méthodes par appariement. À partir de données simulées, Frölich (2004) montre que, même sur de petits échantillons, un choix de la fenêtre obtenu par validation croisée peut donner des résultats corrects. Pour présenter de

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

manière intuitive cette méthode, il s'agit de choisir la taille de la fenêtre qui permet d'obtenir la meilleure estimation du revenu. Cette taille est obtenue comme suit :

$$h^{VC} = \arg\min \sum_{i \in E^0} (Y_i - \hat{Y}_{-i})$$

où $\hat{Y}_{-i}$ correspond à l'estimation du revenu de la personne non bénéficiaire *i* obtenue par la relation (32) en sommant sur l'ensemble des personnes *i'* du groupe de contrôle hormis *i*.

Chacune de ces méthodes présente ses avantages et ses inconvénients. L'opposition entre la plus simple (appariement avec le plus proche voisin) et la plus complexe (noyau) reflète le dilemme classique entre biais et variance. Ne pas utiliser l'ensemble de l'information disponible, dans le cas de l'appariement avec le plus proche voisin, réduit théoriquement la précision. Frölich (2004) montre que, même sur de petits échantillons, les estimations par fonction à noyau sont toujours plus précises. On peut cependant craindre qu'elles augmentent aussi les risques de mauvais appariements (et donc les biais). En pratique, on recommande de tester la sensibilité des résultats à la méthode utilisée.

Au-delà du choix de la méthode, l'appariement exact est souvent compliqué à mettre en œuvre. *A priori*, pour que l'hypothèse d'indépendance conditionnelle soit vérifiée, on peut souhaiter utiliser le maximum d'information et, donc, apparier sur de très nombreuses variables. Toutefois, il peut alors s'avérer difficile de trouver un voisin proche. On montre qu'à distance finie, les estimateurs sont d'autant plus biaisés que le nombre de variables de conditionnement est élevé. Le problème s'avère encore plus aigu lorsque des variables de conditionnement sont continues. C'est pourquoi la grande majorité des études empiriques préfèrent l'appariement sur le score de propension.

Appariement sur le score de propension

Une propriété montrée par Rosenbaum et Rubin (1983) permet de régler théoriquement le problème de dimensionnalité. Ces auteurs montrent en effet que, si la propriété d'indépendance conditionnelle (22) est vérifiée, alors le revenu est également indépendant du traitement conditionnellement au *score de propension*, c'est-à-dire à la probabilité $p(X) = P(T = 1|X)$ d'être traité conditionnellement aux observables :

$$(33)\quad Y^0 \perp T|X \Rightarrow Y^0 \perp T|p(X)$$

Cette propriété simplifie en grande partie les estimations en réduisant la dimension des comparaisons[12]. Néanmoins, cette solution n'est pas miraculeuse : en pratique, le score de propension exact est rarement connu, il est donc nécessaire de l'estimer. Les méthodes décrites plus haut (appariement au(x) plus proche(s) voisin(s), radius ou *kernel matching*) peuvent toutes être appliquées en utilisant uniquement l'estimation du score de propension dans le calcul de la distance entre deux observations. D'autres estimateurs plus simples peuvent être utilisés à partir du score de propension (estimateurs par strate, estimateurs pondérés). On en trouvera une description notamment dans Imbens et Wooldridge (2009) ou Givord (2010).

Estimation du score de propension

S'il est toujours possible d'envisager des estimations non-paramétriques du score, en pratique, l'immense majorité des études utilisent une spécification de type logit ou probit :

$$(34)\quad \hat{p}(X) = \frac{\exp[f(X)]}{1+\exp[f(X)]}$$

où $f(X)$ est une fonction des observables X. La fonction la plus simple (et la plus courante) est une forme linéaire $f(X) = X'\beta$. Toutefois, pour davantage se rapprocher de la vraie distribution, il est recommandé d'utiliser une approximation de type polynomial dans les covariables : outre les covariables, on ajoute des termes de degré plus élevé, y compris des interactions. De manière formelle, Hirano *et alii* (2003) proposent d'estimer le score par :

$$(35)\quad \hat{p}(x) = \frac{\exp[\hat{\gamma}_0 h_0(X)+...+\hat{\gamma}_L h_L(X)]}{1+\exp[\hat{\gamma}_0 h_0(X)+...+\hat{\gamma}_L h_L(X)]}$$

où les fonctions $(h_l(X))_{l=1,...,L}$ sont des fonctions polynomiales des variables explicatives, dont le nombre L dépend de la taille de l'échantillon. En pratique, on se contente en général de coefficients de degré un ou deux[13].

Cette spécification a l'avantage de tenir compte du fait que le score est nécessairement borné entre 0 et 1 (c'est une probabilité). Il s'agit cependant d'une spécification paramétrique, ce qui réduit en partie l'intérêt de cette méthode d'appariement sur le score, alors que l'une des justifications de ces méthodes est justement de ne plus reposer sur des spécifications paramétriques de la distribution conditionnelle du revenu.

En fait, le résultat de Rosenbaum et Rubin (1983) est plus général : la relation (35) est vérifiée pour toutes les fonctions *b*, éventuellement différentes du score de propension *p*, vérifiant :

$$(36)\quad T \perp X|b(X)$$

On dit qu'il s'agit d'une fonction "équilibrée" (*balanced*) des covariables. Intuitivement, cela signifie que, conditionnellement à cette fonction des covariables, la composition des observables est la

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

même dans le groupe des bénéficiaires et des non bénéficiaires. L'avantage de cette propriété est que, même si l'on n'observe pas le vrai score, l'estimation obtenue en utilisant une spécification $\hat{p}$ de ce score est convergente dès lors que (36) est vérifiée.

À notre connaissance, il n'existe pas de règle clairement établie pour le choix optimal des termes. Dehejia et Wahba (1999) proposent une procédure de "test" intéressante. Celle-ci s'appuie sur le fait que le score vérifie la propriété (36) :

(37) $T \perp X | p(X)$

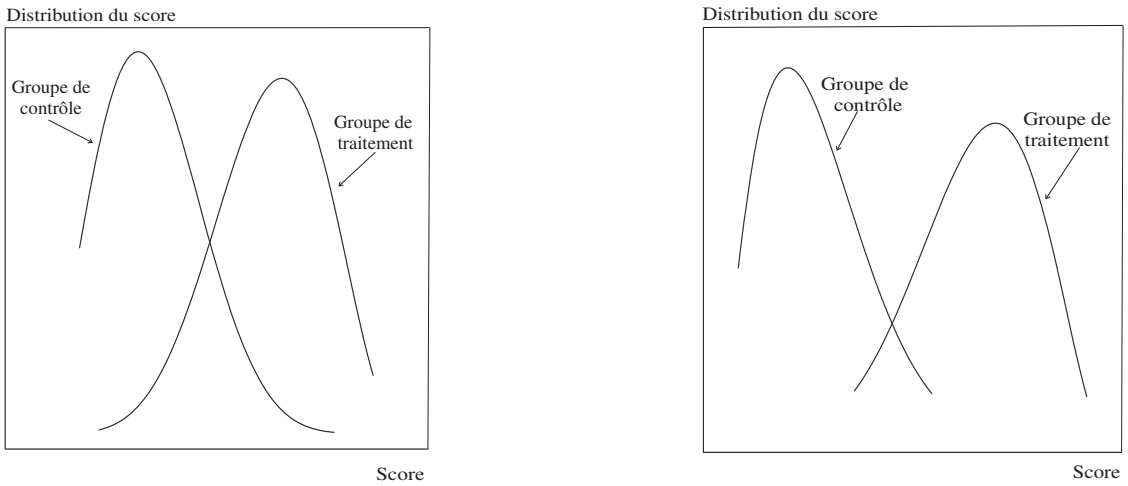
c'est-à-dire sur le fait que le score permet d'équilibrer dans les observables les deux échantillons d'individus traités et non traités. Dehejia et Wahba (1999) proposent donc de complexifier la modélisation du score jusqu'à obtenir des échantillons équilibrés. Partant d'une spécification parcimonieuse au regard des observables ($f(X)=X'\beta$), on enrichit la spécification par des termes de degrés plus élevés tant que la condition (37) n'est pas vérifiée. En principe, il faudrait pouvoir tester $\overline{X}^1 = \overline{X}^0$ conditionnellement à chaque valeur de $p(X)$. Dehejia et Wahba (1999) suggèrent de tester par strates du score. Pour ce faire, on subdivise les échantillons en plusieurs strates, selon leur score. Au sein de chaque strate, on vérifie que les moyennes des observables sont identiques dans les échantillons des individus traités et non traités, par exemple par un test d'égalité des moyennes. Le problème de cette règle pratique est qu'en cas de rejet, il est difficile de savoir si un rejet du test tient à des strates trop grossières (la propriété n'est censée être vérifiée qu'à $p(X)$ constant) ou parce que le score est effectivement mal spécifié. Becker et Ichino (2002) fournissent un programme Stata permettant de faire ce test. Huber (2009) propose une autre procédure de test de cette propriété.

Restriction au support commun

Les méthodes présentées sous l'hypothèse d'indépendance conditionnelle reposent sur l'hypothèse d'un support commun. Celle-ci signifie qu'il existe à la fois des bénéficiaires et des non bénéficiaires ayant des valeurs des observables identiques. Si ce n'est pas le cas, il sera par exemple impossible de trouver pour chaque bénéficiaire des non bénéficiaires comparables.

C'est pourquoi il est important de vérifier que le support commun, c'est-à-dire la zone sur laquelle cette condition est vérifiée, est suffisamment large. Pour cela, la manière la plus simple est de représenter la distribution du score sur les deux sous-échantillons. En pratique, on peut utiliser les histogrammes de la probabilité estimée d'être traité pour l'échantillon des bénéficiaires et des non bénéficiaires. Il est nécessaire de vérifier que le recouvrement est large, c'est-à-dire que, pour chaque valeur du score, on compte un nombre suffisant de personnes dans les deux sous-échantillons. Cela peut également se faire en représentant les densités. Les figures 4 illustrent deux cas caricaturaux. Dans le premier cas, les modes des deux distributions sont bien séparés. Cela exprime que la probabilité d'être traité est plus importante dans le sous-échantillon effectivement traité. Cependant, le support commun est large. Dans le second cas, les deux distributions sont pratiquement disjointes. Il sera impossible d'estimer un effet puisqu'on sera amené à comparer des individus très différents en terme de caractéristiques observables.

Figures 4 : distributions du score de propension



Lecture : la figure de gauche montre une situation factice dans laquelle les supports des distributions du score de propension du groupe de traitement et du groupe de contrôle se recouvrent largement, ce qui est un cas favorable pour permettre des appariements. Dans la figure de droite, le support commun est très étroit, ce qui rend l'appariement difficile.

Source : auteure.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

L'hypothèse d'indépendance conditionnelle suppose que l'on compare des personnes ayant des caractéristiques observables identiques. S'il n'existe pas d'individus comparables dans les deux groupes (traités et non traités) pour certaines valeurs des observables, il est incorrect d'utiliser ces individus pour l'estimation. Si l'on ne se limite pas au support commun, les estimations pourront être biaisées[14].

En pratique, plusieurs méthodes ont été proposées pour se restreindre au support commun. Il faut remarquer que ces méthodes changent la nature de ce qui est estimé. On estime un impact sur une partie seulement de la population (*i.e.* sur celle dont les caractéristiques observables sont telles qu'on constate un recouvrement pour les deux sous-échantillons, d'individus traités et non traités) :

– méthode du min/max : pour estimer l'effet du traitement sur les traités, Dehejia et Wahba (1999) éliminent les individus du groupe de contrôle dont le score est inférieur au minimum observé sur le groupe de traitement. Pour l'effet moyen sur l'ensemble de la population, il faut également éliminer les personnes du groupe de traitement dont la probabilité estimée d'être traitées est plus grande que le maximum observé dans le groupe de contrôle ;

– méthode d'écrémage (*trimming*) : cette méthode consiste à exclure les individus dont la probabilité d'être traités est trop élevée ou trop faible. En pratique, cela revient à se restreindre aux individus tels que :

$$\alpha \leq p(X) \leq 1 - \alpha$$

On trouvera dans Imbens et Wooldridge (2009) une condition sur α pour que l'estimateur de l'effet du traitement soit efficace, sous certaines conditions. Ces auteurs proposent comme règle empirique de retenir $\alpha = 0{,}01$[15].

Quelles variables de conditionnement ?

L'hypothèse d'indépendance conditionnelle, sous laquelle ces différents estimateurs correspondent à l'effet propre de la mesure étudiée, nécessite de disposer de suffisamment de caractéristiques observables dans les données pour que, conditionnellement à ces observables, le fait de bénéficier ou non de la mesure ne soit pas lié au bénéfice escompté. La question du choix de ces variables est évidemment primordiale. Comme souvent, il n'y a pas de règle précise pour les déterminer. En pratique, il doit s'agir de variables qui peuvent avoir un impact sur la variable d'intérêt et sur le choix de bénéficier du dispositif. Deux points doivent être notés. En premier lieu, il faut faire attention à ne pas utiliser des variables mesurées *après* la mise en place du dispositif et qui peuvent être également affectées par celui-ci : cela pourrait créer un problème évident d'endogénéité. Par exemple, supposons que l'on souhaite évaluer l'impact d'une aide aux entreprises sur le niveau d'emploi. Il ne serait pas correct de conditionner par des variables de bilan postérieures à l'obtention de l'aide. De telles variables peuvent en effet être également modifiées par le niveau de l'aide.

En second lieu, pour que l'hypothèse de support commun soit vérifiée, il ne faut pas que des variables explicatives expliquent "trop bien" le fait d'être traité, car cela pourrait rendre impossible un appariement avec des non bénéficiaires très proches. Pour prendre un exemple caricatural, supposons l'existence d'une aide aux entreprises généralisée à l'échelle d'un secteur d'activité. Il serait alors impossible d'utiliser l'appartenance à ce secteur comme variable de conditionnement puisqu'il n'existerait pas d'entreprise non bénéficiaire de cette aide à laquelle comparer la situation des bénéficiaires. Si l'on peut raisonnablement penser que le "revenu" ne dépend pas de cette variable, cela ne pose aucun problème. Si ce n'est pas le cas, cela signifie que l'effet mesuré sera biaisé, puisqu'il intégrera les différences systématiques liées à ce déterminant du revenu.

Inférence

Dans le cadre du modèle linéaire, l'estimation de la précision est tout à fait standard. Dans les autres cas, les méthodes sont loin d'être aussi balisées. Pour mener l'inférence, des méthodes de *bootstrap* sont souvent utilisées. Il faut cependant noter que les preuves de leur convergence n'ont pas été établies jusqu'à présent pour l'ensemble des estimateurs présentés ici. Abadie et Imbens (2008) montrent même qu'ils ne sont pas convergents en général pour les méthodes d'appariement avec remise sur un nombre fixe de voisins.

Traitements multiples

L'un des avantages de la régression linéaire est de permettre d'évaluer des traitements avec plusieurs modalités ou continus. Dans le cas de l'appariement sur le score de propension, Lechner (2002) propose une extension au cas de plusieurs traitements (par exemple, plusieurs types de formation pour les chômeurs). La solution la plus simple consiste à étudier l'impact des différents traitements deux à deux. Par exemple, pour étudier l'effet du traitement T^j comparé à celui du traitement $T^{j'}$, on se place sur le sous-échantillon des individus ayant suivi l'un ou l'autre de ces deux traitements et on applique une des procédures d'appariement vues plus haut. Notons que les résultats peuvent être à la fois plus difficiles à présenter et à interpréter : on ne peut évaluer l'impact d'un traitement que par rapport à un autre.

En résumé

En résumé, les étapes pour mettre en œuvre des méthodes sous une hypothèse d'indépendance conditionnelle sont les suivantes :

– sélectionner un groupe de contrôle ;

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

– sélectionner un ensemble de variables conditionnantes. Rappelons qu'en principe, des données en coupe sont suffisantes. Néanmoins, si l'on dispose de données individuelles antérieures à la mise en œuvre du programme, l'hypothèse d'indépendance conditionnelle est plus crédible lorsque l'on peut contrôler les effets fixes, en comparant les évolutions temporelles des revenus plutôt que les revenus en niveau ;

– déterminer une méthode d'estimation. Comme on l'a vu, on peut tenter une spécification linéaire. Cependant, si les compositions des échantillons d'individus traités et non traités sont trop différentes et que la dépendance réelle du revenu en fonction des caractéristiques observables s'éloigne trop de cette approximation linéaire, les estimations risquent d'être peu robustes. Il peut donc être préférable d'utiliser une procédure qui ne repose pas sur cette hypothèse de dépendance linéaire. Les méthodes sont multiples, la plus courante étant l'appariement sur le score de propension ;

– dans ce dernier cas, estimer le score de propension. Cela peut se faire par un logit ou un probit. En pratique, il est nécessaire d'utiliser une estimation du score qui "équilibre" les deux échantillons, c'est-à-dire telle que les personnes de groupes différents ayant des valeurs proches pour cette fonction présentent des caractéristiques observables identiques ;

– estimer les effets. L'estimation proprement dite se fait par une des méthodes présentées plus haut.

Toute la crédibilité des estimateurs proposés ici repose sur la validité de l'hypothèse d'indépendance conditionnelle. Il faut avoir conscience que cette hypothèse est très forte. Elle tient précisément aux caractéristiques observables disponibles utilisées pour la comparaison.

En particulier, si l'on dispose de trop peu d'informations dans ses données sur les choix entre les variables, il ne sera pas possible d'éliminer le biais de sélection.

Conclusion

Cet article présente différentes méthodes d'évaluation. En pratique, un chargé d'étude pourra se demander quelle méthode est la "meilleure". Lorsqu'on se limite à une évaluation *ex post*, il n'existe évidemment pas de réponse générale à cette question. Le plus souvent, on choisit la méthode "utilisable" en fonction du dispositif évalué et des données disponibles. Il faut aussi avoir conscience que les bases théoriques de la littérature traitant de l'évaluation se posent progressivement. Les usages évoluent aussi en fonction de ces avancées. Très schématiquement, les différences de différences, conceptuellement très simples mais qui reposent sur un choix du groupe de contrôle *ad hoc*, ont été "supplantées" au cours des années 1990 par les variables instrumentales issues d'expériences naturelles, dans la lignée des travaux d'Angrist, Krueger et Card notamment. Les limites de ce type d'estimateur, mises en évidence par exemple par Angrist, Imbens et Rubin (1996), ainsi que le papier influent de Heckman *et alii* (1997), qui présente une étude approfondie des estimateurs obtenus par des méthode d'appariement, ont beaucoup fait pour populariser ces dernières au tournant de la dernière décennie. Depuis le milieu des années 2000, c'est l'approche par les régressions sur discontinuités qui est considérée comme la plus convaincante, suite en particulier à l'article de Hahn *et alii* (2001), qui formalise et fonde théoriquement cette méthode d'identification. Celle-ci ne permet cependant que de donner des résultats très locaux. Les frontières entre les régressions sur discontinuités et les variables instrumentales sont floues : de nombreuses études antérieures à l'article de Hahn *et alii* (2001) utilisaient déjà, sous le terme de variables instrumentales, une méthode d'identification qui serait appelée aujourd'hui régression sur discontinuité. Comme on l'a vu, les mises en œuvre pratiques de ces deux méthodes sont identiques dans certains cas.

Plusieurs remarques s'imposent. Tout d'abord, il faut garder à l'esprit que toutes ces méthodes empiriques reposent sur des hypothèses fortes. S'affranchir des effets de sélection est difficile. Ceci signifie en particulier qu'il est extrêmement rare de juger de manière indiscutable de l'efficacité d'un dispositif à l'aide d'une telle évaluation. Ceci ne signifie pas que de telles évaluations soient inutiles. Même si elle n'élimine pas tous les biais liés aux caractéristiques inobservables, une évaluation peut apporter des éléments indispensables à l'analyse. Plutôt que s'interdire toute mesure en l'absence de l'outil idéal, une posture plus réaliste est sans doute d'appeler à la multiplication des évaluations utilisant des approches différentes. C'est souvent de la confrontation d'études aboutissant à des conclusions similaires (ou non) que le consensus peut se faire sur l'efficacité d'une politique.

Des utilisations naïves de telles évaluations tendent parfois à résumer leur conclusion à un bilan "ça marche/ça ne marche pas". L'efficacité d'un dispositif ne se laisse cependant pas enfermer dans une telle opposition simpliste. Par exemple, une évaluation d'un effet moyen nul peut masquer des effets contrastés selon les publics concernés. Piketty et Valdenaire (2006) montrent par exemple que la taille des classes a plus d'effet sur les élèves en difficultés que sur les autres. Diminuer le nombre d'élèves par classe peut donc n'avoir qu'un effet limité sur les performances moyennes des élèves, mais réduire une partie de l'échec scolaire. Déterminer si cette mesure est souhaitable relève ensuite d'un choix politique. Une même politique

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

peut avoir des effets différents selon le contexte. Par exemple, alors que Rathelot et Sillard (2009) mettent en évidence des effets négligeables sur l'activité économique des zones franches urbaines de deuxième génération, en utilisant une méthodologie quasiment identique, Givord *et alii* (2014) observent des effets de court terme beaucoup plus importants de la première vague de ce même dispositif. Une explication pourrait tenir à la différence de fiscalité sur le travail entre les deux vagues (voir Givord et Trévien, 2012, pour une discussion plus détaillée). Des effets de bouclage sont également susceptibles de modifier les résultats d'une expérimentation menée à une petite échelle ou sur un court laps de temps.

Enfin, ces évaluations empiriques permettent d'évaluer l'impact d'un dispositif sur un critère précis et quantifiable. Elles ne fournissent pas toujours les clés pour comprendre les raisons d'un échec ou d'un succès, une telle compréhension étant pourtant nécessaire pour juger de leur reproductibilité dans un autre contexte. Ces évaluations ne constituent également qu'un élément parmi d'autres dans une évaluation globale d'un dispositif.

Notes

(1) Une formalisation classique de ce problème de sélection est fournie par le modèle de Roy (Heckman et Honoré, 1990). Dans sa version initiale, ce modèle tente d'expliquer les choix d'occuper telle ou telle profession en fonction des gains espérés par chacun, ceux-ci dépendant des compétences individuelles, qui varient d'une personne à l'autre.

(2) Le 20 avril 1980, suite à des mouvements de protestation à Cuba, Fidel Castro a autorisé ceux qui le souhaitaient à quitter l'île. Plus de 100 000 Cubains ont alors embarqué de la baie de Mariel, face aux côtes de Floride.

(3) Ils peuvent être différents si la mesure met du temps à produire ses pleins effets, par exemple.

(4) Disposer de plusieurs groupes de contrôle possibles peut conduire à mettre en œuvre des estimateurs de triples différences – *cf.* Givord (2010) et, pour une illustration, voir Mauroux (2014), dans ce numéro.

(5) Pour estimer l'effet de faire des études supérieures sur le revenu, Card (1993) utilise un instrument proche, la proximité dans l'enfance à une ville universitaire. Cet instrument est cependant à utiliser avec précaution : il n'est pas sûr que le choix de localisation des parents soit totalement déconnecté des possibilités futures de faire des études.

(6) On peut néanmoins obtenir des informations sur la proportion de *compliers* - voir Imbens et Wooldridge (2009), repris dans Givord (2010).

(7) De surcroît, une faible corrélation entre l'instrument et la variable de traitement suggère que la part des *compliers* est peu élevée. Même s'il était correctement configuré, cet estimateur représenterait en tout état de cause l'effet de la mesure sur une petite part de la population. Cela soulève donc également des questions sur sa validité externe.

(8) L'estimation des effets définis par (15) et (19) peut se faire composante par composante (voir Givord, 2010, pour une présentation détaillée).

(9) La mesure (29) diffère donc de la statistique de Student du test de l'hypothèse nulle d'égalité des moyennes, qui s'écrit : $ST_X^k = \frac{\overline{X}_k^1 - \overline{X}_k^0}{\sqrt{V_k^1 / N^1 + V_k^0 / N^0}}$. Cette statistique augmente mécaniquement avec la taille des échantillons. Or, si les proportions restent identiques dans les deux groupes, simplement augmenter la taille des échantillons ne permet pas de régler notre souci principal, qui est d'estimer un contrefactuel à partir du groupe de contrôle, si celui-ci présente des caractéristiques observables trop éloignées du groupe de traitement (ce que mesurent mieux *a priori* les seules différences normalisées (29)).

(10) L'effet du traitement sur toute la population est estimé en effectuant la même opération pour les individus du groupe de contrôle : on compare les revenus observés à ceux des plus proches voisins parmi le groupe des individus traités.

(11) Rappelons qu'un noyau est une fonction K non négative symétrique, intégrable, définie sur l'ensemble des réels $\Re$, telle que $\int_{-\infty}^{\infty} \mathrm{K}(u)\,du = 1$.

(12) Caliendo et Kopeinig (2008) proposent un guide pratique de mise en œuvre de l'appariement sur le score de propension.

(13) Précisément, les résultats de convergence sont établis par Hirano *et alii* (2003) pour $L = cste N^{\nu}$, avec ν tel que $0 < \nu < 1/9$ si le vrai score est infiniment différenciable.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

(14) Empiriquement, les estimateurs ne traitent pas vraiment ce problème. Un plus proche voisin d'un bénéficiaire peut se trouver en dehors du support de la distribution des observables pour les bénéficiaires. L'hypothèse linéaire permet "artificiellement" de comparer des individus pour des valeurs de covariables que l'on n'observe pas.

(15) Heckman *et alii* (1997) utilisent une méthode beaucoup plus complexe.

Bibliographie

Abadie A., Diamond A. et Hainmueller J. (2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program", *Journal of the American Statistical Association*, vol. 105, n° 490, pp. 493-505.

Abadie A. et Imbens G. (2008). "On the Failure of the Bootstrap for Matching Estimators", *Econometrica*, vol. 76, n° 6, pp. 1537-1557.

Angrist J. et Evans W. (1998). "Children and Their Parents Labor Supply: Evidence from Exogenous Variation in Family Size", *American Economic Review*, vol. 88, n° 3, pp. 450-477.

Angrist J., Imbens G. et Rubin D. (1996). "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, vol. 91, n° 434, pp. 444-455.

Angrist J. et Krueger A. (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?", *Quarterly Journal of Economics*, vol. 106, n° 4, pp. 979-1014.

Angrist J. et Lavy V. (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics*, vol. 114, n° 2, pp. 533-575.

Ashenfelter O. (1978). "Estimating the Effects of Training Programs on Earnings", *Review of Economic Studies,* vol. 60, n° 1, pp. 47-57.

Autor D.H. (2003). "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing", *Journal of Labor Economics*, vol. 21, n° 1, pp. 1-42.

Becker S. et Ichino A. (2002). "Estimation of Average Treatment Effects Based on Propensity Scores", *The Stata Journal*, vol. 2, n° 4, pp. 358-377.

Bertrand M., Duflo E. et Mullainathan S. (2002). "How Much Should we Trust Differences-in-Differences Estimates?", *The Quarterly Journal of Economics*, vol. 119, n° 1, pp. 249-275.

Black S. (1999). "Do Better Schools Matter? Parental Valuation of Elementary Education", *Quarterly Journal of Economics*, vol. 114, n° 2, pp. 587-599.

Black D., Galdo J. et Smith J.A. (2007). "Evaluating the Worker Profiling and Reemployment System Using a Regression Discontinuity Approach", *American Economic Review*, vol. 97, n° 2, pp. 104-107.

Bléhaut M. et Rathelot R. (2014). "Expérimentation contrôlée contre appariement : le cas d'un dispositif d'accompagnement de jeunes diplômés demandeurs d'emploi", *Économie et Prévision*, n° 204-205, pp. 163-181, ce numéro.

Brodaty T., Crépon B. et Fougère D. (2007). "Les méthodes micro-économétriques d'évaluation et leurs applications aux politiques actives de l'emploi", *Économie et Prévision*, n° 177, pp. 91-118.

Caliendo M. et Kopeinig S. (2008). "Some Practical Guidance for the Implementation of Propensity Score Matching", *Journal of Economic Survey,* vol. 22, n° 1, pp. 31-72.

Card D. (1990). "The Impact of the Mariel Boatlift on the Miami Labor Market", *Industrial and Labor Relations Review*, vol. 43, n° 2, pp. 245-257.

Card D. (1993). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling", Document de travail n° 4483, NBER, 35 pages.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Card D. et Krueger A. (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania", *American Economic Review*, vol. 84, n° 4, pp. 772-793.

Dehejia R. et Wahba S. (1999). "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of Statistical American Association*, vol. 94, n° 448, pp. 1053-1062.

Duflo E., Glennerster R. et Kremer M. (2008). "Using Randomization in Development Economics Research: A Toolkit", *Handbook of Development Economics*, pp.1-73.

Frölich M. (2004). "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators", *The Review of Economics and Statistics*, vol. 86, n° 1, pp. 77-90.

Givord P. (2010). "Méthodes économétriques pour l'évaluation de politiques publiques", Insee-Dese, document de travail n° G2010-08, 74 pages.

Givord P., Quantin S. et Trévien C. (2014). "A Long-Term Evaluation of the First Generation of the French Urban Enterprise Zones", Crest, document de travail n° G2014-48, 39 pages.

Givord P. et Trévien C. (2012). "Les Zones Franches Urbaines : quel effet sur l'activité économique ?", *Insee Analyses* n° 4, 4 pages.

Hahn J., Todd P. et Van der Klaauw V. (2001). "Estimation of Treatment Effects with a Quasi-Experimental Regression-Discontinuity Design", *Econometrica*, vol. 69, n° 1, pp. 201-209.

Heckman J. et Honoré B. (1990). "The Empirical Content of the Roy Model", *Econometrica*, vol. 58, n° 5, pp.1121-1149.

Heckman J., Ichimura H. et Todd P. (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, vol. 64, n° 4, pp. 605-654.

Hirano K, Imbens G. et Ridder G. (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica*, vol. 71, n° 4, pp. 1161-1189.

Huber M. (2009). "Treatment Evaluation in the Presence of Sample Selection", University of St. Gallen Department of Economics Working Paper Series, n° 2009-07, 41 pages.

Imbens G. et Kalyanaraman K. (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator", *Review of Economic Studies*, vol. 79, n° 3, pp. 933-959.

Imbens G. et Lemieux T. (2008). "Regression Discontinuity Designs: A Guide to Practice", *Journal of Econometrics*, vol. 142, n° 2, pp.615-635.

Imbens G. et Wooldridge J. (2009). "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Litterature*, vol. 47, n° 1, pp. 5-86.

Lalive R. (2008). "How Do Extended Benefits Affect Unemployment Duration? A Regression Discontinuity Approach", *Journal of Econometrics*, vol. 142, n° 2, pp. 785-806.

LaLonde R (1986). "Evaluating the Econometric Evaluation of Training Programs with Experimental Data", *American Economic Review*, vol. 76, n° 4, pp. 604-620.

Lechner M. (2002). "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies", *The Review of Economics and Statistics*, vol. 84, n° 2, pp. 205-220.

Lise J., Seitz S. et Smith J. (2004). "Equilibrium Policy Experiments and the Evaluation of Social Programs", NBER Working Paper n° 10283, pp. 385-437.

Mauroux A. (2014). "Le crédit d'impôt dédié au développement durable : une évaluation économétrique", *Économie et Prévision*, n° 204-205, pp. 89-117, ce numéro.

McClellan M., McNeil B. et Newhouse J. (1994). "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality? Analysis Using Instrumental Variables", *Journal of American Medical Association*, vol. 272, n° 11, pp. 859-866.

Miguel E. et Kremer M. (2004). "Worms: Identifying Impact on Education and Health in the Presence of Treatment Externalities", *Econometrica*, vol. 72, n° 1, pp.159-217.

Piketty T. et Valdenaire M. (2006). "L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français - Estimations à partir du panel primaire 1997 et du panel secondaire 1995". Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche, direction de l'évaluation et de la prospective, dossier n° 173, mars, 153 pages.

Rathelot R. et Sillard P. (2009). "Quel effet des zones franches urbaines sur l'activité économique et l'emploi", *Économie et Statistique,* n° 415, pp. 81-96.

Robin J.-M. (2002). *Endogénéité et variables instrumentales dans les sciences sociales*, Insee méthodes, n° 101, pp. 217-276 pages.

Rosenbaum P. et Rubin D. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika*, vol. 70, n° 1, pp. 41-55.

Rosenzweig M. et Wolpin K. (2000). "Natural "Natural Experiment" in Economics", *Journal of Economic Literature*, vol. XXXVIII, n° 4, pp. 827-874.

Thistlethwaite D.L. et Campbell D.T. (1960). "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment", *Journal of Educational Psychology*, vol. 51, n° 6, décembre, pp. 309-317.

Todd P. et Wolpin K. (2006). "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility", *American Economic Review*, vol. 96, n° 5, pp. 1384-1417.

Van der Klaauw W. (2002). "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression Discontinuity Design", *International Economic Review*, vol. 43, n° 4, pp. 1249-1487.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)