METHODES D'APPARIEMENT APPLIQUEES A L'EVALUATION DES POLITIQUES DE L'EMPLOI

B. CREPON

INSEE - Division Marchés et Stratégies d'Entreprise

1. Introduction

Comment évaluer l'effet pour un individu d'un passage par un dispositif de formation ? Cette question a reçu une attention soutenue de la part des économètres (voir Heckman, Smith and Lalonde (1999) pour une revue très complète). Les méthodes proposées sont nombreuses et leur utilisation conditionne fortement les résultats (Lalonde (1986), Heckman and Hotz (1989) Heckman and Robb (1985)). Ces méthodes sont en général basées sur une comparaison entre des individus passant par un dispositif et des individus n'y passant pas. Une question récurrente dans le domaine de l'évaluation est celle des biais de sélectivité : le passage par un dispositif de formation est une décision rationnelle de la part des agents et n'est probablement pas indépendante de ce qu'aurait été leur situation s'ils n'étaient pas passés par ce dispositif. Une comparaison directe entre les deux populations est donc susceptible de conduire à des estimations biaisées de l'effet du dispositif.

Une démarche naturelle est de construire un groupe de contrôle de telle sorte que la distribution d'un ensemble de caractéristiques observables soit la même que dans le groupe des individus passant par le dispositif. Si l'intuition d'une telle démarche est claire, on ignore en général les conditions de validité de ce type d'évaluation et en particulier sa capacité à éliminer les biais de sélectivité. On dit dans ce cas qu'il y a sélectivité sur observables, situation fréquemment opposée car trop restrictive à celle dans laquelle il y a sélectivité sur observables et sur inobservables.

Dans cette contribution, on présente le cadre théorique adapté pour poser la question de l'évaluation des politiques économiques et discuter les conditions de validité de la démarche précédente. Ce cadre a été à l'origine développé par les statisticiens dans un domaine très éloigné de celui de l'évaluation des politiques économiques (Rubin (1974)). Les statisticiens ont en effet développé différents outils, permettant de définir un ensemble de concepts et de méthodes, adaptés à la question de l'évaluation des traitements dans le domaine médical. La portée de ces concepts, en particulier la notion de causalité, différente de celles développées par les économètres sur séries temporelles (Holland (1989)), et ces méthodes dépassent le cadre restreint dans lequel ils ont été développés. Ils ont récemment été utilisés dans le domaine économétrique (Imbens et Angrist (1994) et les travaux de Heckman) et ont contribué au renouveau des questions touchant l'évaluation des politiques économiques. Leur application a néanmoins été confinée jusqu'à présent à l'évaluation des programmes de formation, mais potentiellement ces méthodes devraient permettre d'aborder un champ plus vaste.

On présente d'abord le cadre causal de Rubin permettant de définir l'effet causal d'une politique et qui est bien adapté pour discuter la question des biais de sélectivité. Deux caractéristiques importantes de l'effet causal, tel que défini dans ce cadre, sont d'une part son hétérogénéité dans la population et d'autre part son inobservabilité, et donc la nécessité de formuler des hypothèses permettant d'identifier certains paramètres de sa distribution.

On se concentre ensuite sur le cas de la sélectivité sur observables, qui a reçu récemment une attention importante et on présente différents estimateurs alternatifs possibles. On présente en particulier les estimateurs par appariement proposés à l'origine par Rubin (1977) et développés récemment par Heckman, Ichimura et Todd (1998) sous le nom de "kernel matching".

On examine ensuite la situation dans laquelle il y a sélectivité sur observables et sur inobservables et on discute les hypothèses nécessaires à l'identification d'un tel modèle.

On conclut ensuite en soulignant l'utilité de chacune de ces méthodes et en soulignant aussi leurs limites. En particulier les méthodes basées sur l'hypothèse de sélectivité sur observables permettent raisonnablement d'évaluer l'effet de certains dispositifs, elles ne permettent pas néanmoins de comprendre et de mettre en évidence les différents mécanismes conduisant aux évolutions observées. Elles apparaissent ainsi peu adaptées pour guider dans l'orientation à donner aux dispositifs étudiés, faute de permettre de comprendre comment ils opèrent. En revanche les méthodes basées sur l'hypothèse de sélectivité sur observables et sur inobservables imposent en général de modéliser les comportements. Ceci constitue une restriction importante mais permet de mieux comprendre les mécanismes à l'œuvre et leur importance.

2. Le modèle causal de Rubin.

2.1. Notations

Le modèle causal présenté ici a été introduit par le statisticien D. Rubin en 1974. Il est adapté à la situation dans laquelle un *traitement* $T_i \in \{0,1\}$ peut être administré ou non à un individu i i=1,...,N. Le terme de traitement se réfère aux premiers travaux ayant permis de développer ce cadre conceptuel, et qui concernaient l'évaluation de l'efficacité des traitements dans le domaine médical. Bien qu'il ne soit pas toujours le plus approprié, en particulier dans le domaine économique, il a été conservé.

L'efficacité du traitement est mesurée à travers une variable de résultat y_i. Le modèle de Rubin introduit pour chaque individu deux variables latentes $y(0)_i$ et $y(1)_i$ correspondant aux *résultats potentiels* de l'individu selon qu'il reçoit le traitement ($T_i=1$) ou non ($T_i=0$). Elles ne sont jamais simultanément observées à la même date pour un même individu. Ainsi, pour un individu traité par exemple, $y(1)_i$ est connue et est mesurée par la variable de résultat observée de l'individu y_i tandis que $y(0)_i$ est inconnue et correspond au résultat potentiel qui aurait été réalisé si l'individu n'avait pas été traité. Pour un individu non traité, on observe au contraire $y(0)_i$ tandis que $y(1)_i$ est inconnue.

La variable de résultat observée peut donc se déduire des variables potentielles et de la variable de traitement par la relation :

$$y_i = T_i y(1)_i + (1 - T_i) y(0)_i .$$

Seul le couple (y_i, T_i) est observé pour chaque individu.

2.2. Paramètres d'intérêt

Rubin définit *l'effet causal* c_i du traitement T_i sur l'individu i par $c_i = y(1)_i - y(0)_i$. Il s'agit donc de la différence entre ce que serait la situation d'un individu s'il était traité et ce que serait sa situation s'il ne l'était pas. L'effet causal a ainsi deux caractéristiques importantes :

- il est *inobservable* puisque seule une des deux variables potentielles est observée pour chaque individu.
- il est *individuel.* Il existe donc une distribution de l'effet causal dans la population.

La distribution de l'effet causal n'est toutefois pas identifiable en raison de l'inobservabilité de l'effet causal du traitement. Néanmoins, sous certaines hypothèses sur la loi jointe du triplet $(y(0)_i, y(1)_i, T_i)$, on peut identifier certains paramètres de la distribution de l'effet causal à partir de la densité des observables (y_i, T_i). Des paramètres ayant fait l'objet d'un examen spécifique sont ainsi *l'effet causal moyen dans la population*

$$E(c_i) = E(y(1)_i - y(0)_i)$$

ou encore *l'effet causal moyen sur la population des individus traités*

$$E(c_i | T_i = 1) = E(y(1)_i - y(0)_i | T_i = 1).$$

Ces deux paramètres ne sont égaux que sous certaines hypothèses très restrictives.

D'une façon générale, quel que soit le type de restriction identifiante utilisée, les conditions nécessaires à l'identification du premier paramètre sont plus exigeantes que celles nécessaires à l'identification du second paramètre. En effet, pour ce dernier paramètre, les hypothèses ne portent que sur la loi de y(0) et T. On comprend bien que l'estimation du premier paramètre est beaucoup plus ambitieuse que celle du second. En effet, le second paramètre ne s'intéresse qu'à la population des individus traités, alors que le second s'intéresse à la population dans son ensemble. Il prétend donc mesurer l'effet de la mesure si elle était étendue à l'ensemble de la population.

3. Identification sous l'hypothèse d'indépendance entre les résultats potentiels et le traitement conditionnellement à des observables

Une condition d'identification suffisante est celle d'indépendance entre les variables de résultat latentes et la variable de traitement : $(y(0)_i, y(1)_i) \perp T_i$. Cette condition est exigeante et nécessite le contexte très particulier de la randomisation. C'est ce contexte dont on cherche à se rapprocher dans le cas des expériences contrôlées : les individus sont affectés aléatoirement au traitement. Si cette condition est satisfaite, les deux paramètres d'intérêt $E(c_i)$ et $E(c_i | T_i = 1)$ sont identifiés :

$$E(c_i) = E(y(1)_i) - E(y(0)_i) = E(y(1)_i | T_i = 1) - E(y(0)_i | T_i = 0) = E(y_i | T_i = 1) - E(y_i | T_i = 0)$$

De même

$$\begin{aligned} E(c_i | T_i = 1) &= E(y(1)_i | T_i = 1) - E(y(0)_i | T_i = 1) = E(y(1)_i | T_i = 1) - E(y(0)_i | T_i = 0) \\ &= E(y_i | T_i = 1) - E(y_i | T_i = 0) \end{aligned}$$

Dans ce cas, les deux paramètres sont identiques et peuvent être estimés simplement comme la différence des moyennes des variables de résultat sur les individus traités et les individus non traités.

On peut remarquer que l'identification du second paramètre nécessite une hypothèse moins forte. Il suffit en effet d'avoir $y(0)_i \perp T_i$.

Biais de sélectivité

Dés lors que la propriété d'indépendance précédente n'est plus satisfaite, l'estimateur naturel formé par la différence des moyennes des variables de résultat est affecté d'un biais :

$$
\begin{aligned}
E(y_i|T_i=1)-E(y_i|T_i=0) &= E(y(1)_i|T_i=1)-E(y(0)_i|T_i=0) \\
&= E(y(1)_i|T_i=1)-E(y(0)_i|T_i=1)+\left[E(y(0)_i|T_i=1)-E(y(0)_i|T_i=0)\right] \\
&= E(c_i|T_i=1)+\text{Biais}
\end{aligned}
$$

Le terme de biais trouve son origine dans le fait que ce qu'aurait été la situation moyenne des individus qui ont reçu le traitement en l'absence de traitement n'est pas la même que celle des individus n'ayant effectivement pas reçu le traitement. Il en est ainsi parce que ces deux populations ne sont pas identiques, sauf dans le cas particulier d'une expérience contrôlée, où l'échantillon des individus non traités est tiré aléatoirement dans celui des individus postulant au traitement. C'est bien dans cette situation où les deux populations ne sont pas identiques que l'on se trouve en pratique, dans l'évaluation des programmes de formation par exemple, où un échantillon de contrôle est construit mais pour lequel les densités des observables diffèrent entre individus traités et non traités.

Une condition d'identification alternative moins restrictive consiste à considérer qu'il existe un ensemble de variables observables x_i conditionnellement auquel la propriété d'indépendance entre les résultats latents et l'affectation au traitement est vérifiée : $(y(0)_i, y(1)_i) \perp T_i | x_i$. L'existence et la détermination d'un tel ensemble de variables de contrôle sont essentielles dans ce type d'analyse.

Une façon alternative de formuler cette restriction est de considérer que conditionnellement à des variables observables des individus on se situe dans le cadre d'une expérience contrôlée, c'est-à-dire avec une affectation aléatoire au traitement. La présence de cette affectation aléatoire au traitement résiduelle joue un rôle essentiel dans l'analyse.

On peut alors montrer que la densité de chacun des résultats potentiels est identifiable puisque leur densité conditionnelle aux observables est identifiable.

Proposition

$$y(0)_i, y(1)_i \perp T_i | x_i \;\Rightarrow\; l(y(0)_i),\, l(y(1)_i),\, l(y(0)_i | T_i = 1) \text{ et } l(y(1)_i | T_i = 0) \text{ identifiables}$$

De fait, sous l'hypothèse d'indépendance conditionnelle à des observables, on a :

$$l(y(0)_i|x_i) = l(y(0)_i|x_i, T_i = 0) = l(y_i|x_i, T_i = 0)$$

$$l(y(1)_i|x_i) = l(y(1)_i|x_i, T_i = 1) = l(y_i|x_i, T_i = 1)$$

Comme la densité des observables est identifiable, il en résulte que les quatre densités $l(y(0)_i)$, $l(y(1)_i)$, $l(y(0)_i|T_i = 1)$ et $l(y(1)_i|T_i = 0)$ sont identifiables.

Il est alors possible d'identifier l'ensemble des paramètres de chacune des quatre distributions. On peut en particulier identifier leur espérance et donc l'effet moyen du traitement, de même que l'effet moyen du traitement conditionnellement au traitement.

Comme précédemment, la condition d'identification pour ce dernier paramètre, est moins forte, puisqu'elle ne nécessite que l'indépendance entre le résultat potentiel en l'absence de traitement et le traitement : $y(0)_i \perp T_i|x_i$. En effet pour identifier $E(c_i|T_i = 1)$ il est suffisant d'identifier $l(y(1)_i|T_i = 1)$ ce qui ne nécessite pas d'hypothèse et $l(y(0)_i|T_i = 1)$ qui par contre requiert la propriété d'indépendance conditionnelle.

Remarquons enfin que dans un cas comme dans l'autre, la distribution des variables latente est identifiée, mais leur loi jointe ne l'est pas. Les moments d'ordre supérieur à 1 de l'effet du traitement (en particulier la variance) ne sont pas identifiés.

3.1. Indépendance conditionnellement à des observables et à des inobservables

La validité de la condition d'identification $y(0)_i \perp T_i|x_i$ peut être légitimement mise en cause par la présence d'effets fixes inobservables u_i affectant à la fois les variables de performances potentielles et la variable de traitement. Une hypothèse plus vraisemblable pourrait donc être $y(0)_i \perp T_i|x_i, u_i$, dans laquelle une partie des variables nécessaires à la propriété d'indépendance est inobservée.

Ce cadre peut se ramener au cadre précédent dès lors qu'il est possible d'éliminer l'information contenue dans l'élément inobservé sur les variables de performances par une transformation adaptée de celles-ci. En partant de l'hypothèse d'indépendance conditionnellement à des observables et à des inobservables, on peut en effet montrer le résultat suivant :

$$y(0)_i \perp T_i | x_i, u_i \text{ et } g(y(0)_i, x_i) \perp u_i | x_i \Rightarrow g(y(0)_i, x_i) \perp T_i | x_i$$

Un cas intéressant est celui dans lequel la liste des variables de conditionnement comprend les variables de performances à une date antérieure au traitement notées y_i^p, avec $y_i^p \subset \{x_i\}$ et que par différenciation il est possible d'éliminer l'effet individuel i.e. $(y(0)_i - y_i^p) \perp u_i | x_i$, à l'instar de ce qui est communément effectué dans le cadre des données de panel. On obtient alors la propriété d'indépendance conditionnellement à des observables pour les évolutions des variables de performances : $(y(0)_i - y_i^p) \perp T_i | x_i$. Ceci a conduit Heckman, Ichimura et Todd (1998) à une généralisation de l'estimateur par différence de différence, largement utilisé dans le cas de l'évaluation des politiques. La terminologie "différence de différence" correspond au fait que l'on considère la variation dans le temps de la variable de résultat, ce qui constitue une première différenciation, et qu'on la compare entre les individus traités et non traités, ce qui en constitue une seconde.

3.2. Indépendance conditionnellement au score

Avant d'aborder la question de l'estimation de l'effet du traitement, on peut mentionner une propriété importante due à Rosenbaum et Rubin (1983). La propriété d'indépendance conditionnellement à des observables implique celle d'indépendance conditionnellement au score $s(x_i) = P(T_i = 1 | x_i)$ (probabilité de traitement) :

Proposition (Rosenbaum et Rubin, 1983) :

$$y(0)_i \perp T_i | x_i \Rightarrow y(0)_i \perp T_i | s(x_i) \text{ avec } s(x_i) = P(T_i = 1 | x_i)$$

Démonstration

D'une part on a : $P(T_i = 1 | x_i, y(0)_i) = P(T_i = 1 | x_i) = s(x_i)$

Et par ailleurs :

$$\begin{aligned} P(T_i = 1 | s(x_i), y(0)_i) &= E(T_i = 1 | s(x_i), y(0)_i) = E(E(T_i = 1 | x_i, y(0)_i) | s(x_i), y(0)_i) \\ &= E(E(T_i = 1 | x_i) | s(x_i), y(0)_i) = E(s(x_i) | s(x_i), y(0)_i) = s(x_i) \end{aligned}$$

Cette propriété est importante car en général le nombre de variables de conditionnement introduites est important. Elle montre que, pour ce qui est de la propriété d'indépendance, on peut se contenter du résumé de ces variables de dimension 1 qu'est la probabilité d'être traité sachant les variables de conditionnement.

4. Méthodes d'estimation sous l'hypothèse d'indépendance conditionnelle à des observables

4.1. Estimateur par appariement

Principe de l'estimation

Le principe de l'estimation est d'utiliser les informations dont on dispose sur les individus non traités pour construire pour chaque individu traité un contrefactuel, c'est-à-dire une estimation de ce qu'aurait été sa situation s'il n'avait pas été traité.

En effet, si on considère le cas de l'effet causal du traitement sur les traités, on a :

$$\begin{aligned} E(c_i|T_i=1) &= E(y(1)_i - y(0)_i|T_i=1) = E(y_i - y(0)_i|T_i=1) \\ &= E(y_i - E(y(0)_i|x_i, T_i=1)|T_i=1) = E(y_i - E(y(0)_i|x_i, T_i=0)|T_i=1) \\ &= E(y_i - E(y_i|x_i, T_i=0)|T_i=1) \end{aligned}$$

Le problème est donc d'estimer pour chaque individu i_0 traité de caractéristique x_{i_0} la quantité : $E(y_j|x_j = x_{i_0}, T_j = 0) = g(x_{i_0})$.

L'estimateur final est alors obtenu comme la moyenne des écarts de la situation des individus traités et du contrefactuel construit :

$$\hat{E}(c_i|T_i=1) = \frac{1}{N_1}\sum_{i\in I_1}\{y_i - \hat{g}(x_i)\}$$

où I_1 est l'ensemble des individus non traités $I_1 = \{i|T_i = 1\}$ et N_1 le nombre d'individus traités.

Première solution : Matching

La méthode initialement proposée par Rubin (1977) était celle de l'appariement. Cette méthode associe à chaque individu i traité un individu non traité $\tilde{i}(i)$ dont les caractéristiques sont identiques à celles de l'individu i : $x_i = x_{\tilde{\imath}(i)}$. La quantité $y_{\tilde{\imath}(i)}$ est un estimateur de l'espérance de le résultat potentiel $y(0)_i$ conditionnellement au score de l'individu i :

$$y_{\tilde{\imath}(i)} = \hat{E}\left(y(0)_i \middle| T_i = 0, x_i\right) = \hat{E}\left(y(0)_i \middle| T_i = 1, x_i\right)$$

L'estimateur de Rubin est alors :

$$\hat{E}^R\left(c_i \middle| T_i = 1\right) = \frac{1}{N_1} \sum_{i \in I_1} y_i - y_{\tilde{\imath}(i)}$$

Problèmes rencontrés en pratique

Cette méthode simple en apparence posait toutefois en pratique de nombreux problèmes.

- **Dimensionalité de l'appariement : le rôle central joué par le score.** La propriété d'indépendance conditionnelle nécessite en général l'introduction d'un nombre important de variables de conditionnement. Il est alors nécessaire d'apparier les individus sur un grand nombre de caractéristiques. Cet appariement est difficile à réaliser en pratique, dans la mesure où on ne peut trouver systématiquement pour chaque individu traité un individu non traité ayant exactement les mêmes caractéristiques. Il est alors nécessaire de définir une distance entre les individus sur la base des valeurs prises par les variables de conditionnement et de considérer pour chaque individu traité l'individu non traité le plus proche. La propriété de Rubin et Rosenbaum est très utile puisqu'elle permet de réduire la dimension de l'appariement du nombre de variables retenues dans la liste des variables de conditionnement à la dimension du score, c'est-à-dire 1. Ainsi, il n'est pas nécessaire d'apparier les individus sur chacune des variables de conditionnement. Il suffit de les apparier sur le score, lequel constitue un résumé uni-dimensionnel de l'ensemble de ces variables. L'individu $\tilde{\imath}$ non traité, apparié avec l'individu *i* traité est alors défini par $s(x_i) = s(x_{\tilde{\imath}})$.

- **Processus d'appariement**. La qualité de l'appariement n'est jamais parfaite, dans la mesure où l'on ne dispose pas pour chaque individu traité d'un individu non traité de score identique. En outre, une question se pose. *Doit-on rejeter de l'appariement les individus non traités dès qu'ils ont été appariés une fois ?* En d'autres termes, un individu non traité peut-il " servir " à la construction du contrefactuel pour plusieurs individus traités de scores différents (appariement avec ou sans rejet) ? Si l'appariement est avec rejet, on imagine bien que la qualité va se détériorer au fur et à mesure des appariements et que le résultat sera sensible à l'ordre dans lequel on apparie les individus (par où commencer ?).

- **Convergence.** Les propriétés asymptotiques de l'estimateur de Rubin sont inconnues : l'estimateur est-il convergent ? est-il asymptotiquement normal ? qu'elle est sa vitesse de convergence ? Aucun résultat ne permet de décrire le comportement de l'estimateur lorsque le nombre d'individus traité devient grand.

Deuxième solution : Kernel matching

Les travaux effectués par Heckman, Ichimura et Todd (1997), (1998) ont permis de répondre à ces questions. L'idée principale est que la quantité $y_{\tilde{i}}$ est un estimateur non paramétrique de l'espérance du résultat potentiel $y(0)_i$ conditionnellement au score de l'individu i : $y_{\tilde{i}} = \hat{E}(y(0)_i | s(x_i))$.

La quantité $y_{\tilde{i}}$ de l'appariement par la méthode de Rubin correspond à l'estimateur par le voisin le plus proche (*nearest neighbour matching*). L'erreur quadratique de cet estimateur peut toutefois être améliorée en prenant une moyenne pondérée des observations des n voisins les plus proches i (*n nearest neighbours matching*). D'autres estimateurs non paramétriques peuvent être envisagés. Ils consistent tous en une moyenne pondérée des observations de l'échantillon de contrôle (la population des individus non traités).

Heckman et al. (1998) proposent en particulier un estimateur à noyau :

$$\hat{E}(y(0)_i | s(x_i)) = \sum_{j \in I_0} \frac{K_h(s(x_j) - s(x_i))}{\sum_{j \in I_0} K_h(s(x_j) - s(x_i))} y_j$$

où I_0 est l'ensemble des individus non traités $I_0 = \{i | T_i = 0\}$ et N_0 le nombre d'individus non traités.

$$K_h(s(x_j) - s(x_i)) = K((s(x_j) - s(x_i))/h)$$

avec K le noyau et h la fenêtre.

Chaque individu non traité participe ainsi à la construction du contrefactuel de l'individu i avec une importance qui varie selon la distance entre son score et celui de l'individu considéré. L'estimateur final de l'effet du traitement conditionnellement au fait d'être traité est ainsi :

$$\hat{E}(c_i|T_i=1)=\frac{1}{N_1}\sum_{i\in I_1}\left\{y_i-\sum_{j\in I_0}\frac{K_h\left(s(x_j)-s(x_i)\right)}{\sum_{j\in I_0}K_h\left(s(x_j)-s(x_i)\right)}y_j\right\}$$

Les auteurs montrent alors que, sous certaines hypothèses de régularité, cet estimateur est convergent asymptotiquement normal avec une vitesse de convergence de $\sqrt{n}$.

Autres estimateurs

- **Estimateurs non paramétriques par séries**. Ces estimateurs ne sont pas étudiés par Heckman et al. Le principe de cette méthode est simple. Elle consiste à approximer la quantité $E\left(y(0)_i|s(x_i)\right)$ par une somme de fonctions polynomiales du score dont le degré est fixé selon la taille de l'échantillon, i.e. $\hat{E}\left(y(0)_i|s(x_i)\right)=\sum_{k\leq d_n}P_k\left(s(x_i)\right)\hat{\theta}_k$. L'estimation du paramètre est obtenue en régressant la variable de performance sur les polynômes dans la population des individus non traités. Il reste ensuite à évaluer cette fonction en chaque point de la population des individus traités pour déterminer l'estimation de l'effet causal du traitement dans la population des traités : $\hat{E}\left(c_i|T_i=1\right)=\frac{1}{N_1}\sum_{i\in I_1}\left\{y_i-\hat{E}\left(y(0)_i|s(x_i)\right)\right\}$. Andrews (1991) donne une série de résultats permettant de préciser les propriétés de convergence de cet estimateur.

- **Réinterprétation de la condition d'indépendance : pondération des observations**. Les estimateurs par appariement, en particulier celui fondé sur les noyaux, peuvent être réinterprétés comme des estimateurs pondérant différemment les observations pour les individus traités et non traités.

  De fait, partant de la relation $E\left(c_i|T_i=1\right)=E\left(y_i-E\left(y_i|x_i,T_i=0\right)|T_i=1\right)$, la quantité $E\left(E\left(y_i|x_i,T_i=0\right)|T_i=1\right)$ se réécrit :

$$E\left(E\left(y_i|x_i,T_i=0\right)|T_i=1\right)=\int\int y f^y\left(y|x,T=0\right)dy f^x\left(x|T=1\right)dx$$

$$=\int\int\left[y\frac{f^x\left(x|T=1\right)}{f^x\left(x|T=0\right)}\right]f^y\left(y|x,T=0\right)f^x\left(x|T=0\right)dydx$$

$$=E\left(y\frac{f^x\left(x|T=1\right)}{f^x\left(x|T=0\right)}|T=0\right)$$

La moyenne des contrefactuels apparaît ainsi comme une moyenne pondérée des observations des individus non traités, les poids étant fonctions de la distance entre la densité des variables de conditionnement dans la population traitée et non traitée.

Notons que la dernière expression peut se réécrire différemment compte tenu de la relation :

$$\frac{f^x\left(x|T=1\right)}{f^x\left(x|T=0\right)}=\frac{P\left(T=1|x\right)}{P\left(T=0|x\right)}\frac{P\left(T=0\right)}{P\left(T=1\right)}=\frac{s(x)}{1-s(x)}\frac{P_0}{P_1}$$

Les pondérations sont ainsi directement reliées au score pour les individus traités et non traités. Un estimateur alternatif pourrait ainsi être :

$$E\left(c_i|T_i=1\right)=\frac{1}{N_1}\sum_{i\in I_1}\left\{y_i-\frac{P_0}{P_1}\sum_{j\in I_0}\frac{s\left(x_j\right)}{1-s\left(x_j\right)}y_j\right\}$$

L'expression de l'estimateur proposé par Heckman reflète bien cette pondération. En effet la moyenne des contrefactuels s'écrit :

$$\frac{1}{N_1}\sum_{i\in I_1}\frac{1}{N_0}\sum_{j\in I_0}y_j\frac{K\left(s\left(x_j\right)-s\left(x_i\right)\right)}{\frac{1}{N_0}\sum_{j\in I_0}K\left(s\left(x_j\right)-s\left(x_i\right)\right)}$$

soit en intervertissant les sommations :

$$\frac{1}{N_0}\sum_{j\in I_0}y_j\frac{1}{N_1}\sum_{i\in I_1}\frac{K\left(s\left(x_j\right)-s\left(x_i\right)\right)}{\frac{1}{N_0}\sum_{j\in I_0}K\left(s\left(x_j\right)-s\left(x_i\right)\right)}$$

Il s'agit donc bien d'une moyenne pondérée des observations pour les individus non traités, les poids ayant pour expression :

$$\frac{1}{N_1}\sum_{i\in I_1}\frac{K(s(x_j)-s(x_i))}{\frac{1}{N_0}\sum_{j\in I_0}K(s(x_j)-s(x_i))}\longrightarrow\frac{f^{s(x)}(s(x)|T=1)}{f^{s(x)}(s(x)|T=0)}=\frac{s(x)}{1-s(x)}\frac{P_0}{P_1}$$

4.2. Problème de support

La question du support des distributions du score conditionnellement au traitement est essentielle dans ce type d'analyse. Son importance a été soulignée par Heckman et al. (1998) qui ont montré qu'elle constitue une source forte de biais dans l'estimation de l'effet causal du traitement.

Dans les méthodes d'estimation par appariement ou par régression, il est nécessaire de pouvoir construire pour chaque individu traité un contrefactuel à partir des individus non traités, c'est-à-dire de pouvoir estimer $E(y_i|s(x_i),T_i=0)$ pour déterminer l'effet causal du traitement sur la population des individus traités. En outre, il est nécessaire d'estimer $E(y_i|s(x_i),T_i=1)$ dès qu'on s'intéresse à l'effet causal du traitement dans la population totale.

Une estimation non paramétrique de cette quantité, donc sans restriction sur la forme qu'elle prend, impose que l'on dispose pour un individu traité de score *s* d'individus non traités ayant des valeurs du score proche de *s*. Dit d'une autre manière, la densité du score pour les individus non traités ne doit pas être nulle pour les valeurs du score des individus traités considérés.

On ne peut donc construire de contrefactuel que pour les individus dont le score appartient à l'intersection des supports de la distribution du score des individus traités et des individus non traités. Ceci conduit à la conclusion que, même sous l'hypothèse d'indépendance conditionnelle à des observables, on ne peut pas systématiquement estimer $E(c_i)$ dans la mesure où $E(c_i|s(x_i))$ ne peut être estimé que pour les individus dont le score appartient au support commun de la distribution du score pour les individus traités et non traités. *L'estimateur obtenu in fine est alors un estimateur local* : $E(c_i|s(x_i)\in S_\cap,T_i=1)$ ou $E(c_i|s(x_i)\in S_\cap)$, avec $S_\cap$ le support commun défini par $S_\cap=S_T\cap S_{NT}$ avec S_T le support de la distribution du score des individus traités et S_{NT} celui des individus non traités.

Cette condition du support a une autre implication : le modèle servant à la construction du score, c'est-à-dire expliquant le traitement à partir des variables de conditionnement, ne doit pas être trop bon. Dans le cas extrême où on expliquerait parfaitement le traitement, les densités du score conditionnellement au traitement seraient toutes deux des masses de Dirac, l'une en zéro pour les individus non traités, l'autre en 1 pour les individus traités. Les supports seraient alors disjoints et aucun appariement ne serait possible.

Pour bien comprendre cette condition importante du score, il faut garder présente à l'esprit l'idée initiale de Rubin : conditionnellement à un ensemble de variables explicatives x (ou le score), on se trouve dans le cas d'une expérience contrôlée, c'est-à-dire dans laquelle on dispose d'individus traités et non traités qui sont affectés aléatoirement à chacun de ces groupes. La persistance de cette composante aléatoire de l'affectation au traitement conditionnellement à des observables est ainsi essentielle dans la procédure d'appariement.

4.3. Les étapes de l'estimation

Les différentes méthodes d'évaluation précédemment exposées se font toutes en plusieurs étapes. La première consiste à " expliquer " la variable de traitement par les caractéristiques observables x_i. Pour cela la solution la plus commode est d'estimer un modèle de type logit. Cette étape est informative dans la mesure où elle propose une description de l'affectation au traitement. Le choix des variables de conditionnement est essentiel dans cette étape et il faut conserver à l'esprit que ce qui importe n'est pas une description aussi fidèle que possible de la variable de traitement mais simplement les variables nécessaires à l'obtention de la propriété d'indépendance. Introduire un trop grand nombre de variable peut avoir des conséquences néfastes sur l'estimation à plusieurs titres. D'abord, la description de la variable de traitement étant meilleure, les supports de la distribution du score pour les individus traités et les individus non traités risquent de se dissocier davantage et les possibilités d'appariement seront alors plus restreintes. Mais surtout, introduire trop de variables de conditionnement peut conduire à biaiser les estimations. Il est possible en effet que la propriété d'indépendance soit satisfaite pour un ensemble de variables de conditionnement mais qu'elle ne soit plus vraie lorsque l'on rajoute d'autres variables de conditionnement, quand bien même celles-ci seraient significatives dans l'estimation du score.

Il n'y a pas de méthode générale pour déterminer ces variables. Une procédure raisonnable est néanmoins la suivante : éliminer tant que possible les différences permanentes entre individus en considérant des variables de résultat potentiels prises en évolution. Introduire des informations caractérisant le passé de l'individu : résultats potentiels et autres caractéristiques. Considérer une modélisation même fruste de la variable de résultat et chercher à introduire des variables de conditionnement associées aux autres sources de variabilité de la variable de résultat.

La deuxième étape consiste à déterminer le support commun des densités du score pour les individus traités et les individus non traités. Cette étape est importante et l'oublier pourrait là encore biaiser les estimations. C'est en outre une source d'imprécision des estimations importante, l'estimation du contrefactuel pour les individus n'appartenant pas au support commun étant biaisée et très imprécise.

La dernière étape est celle de l'estimation à proprement parler. Il suffit pour cela d'appliquer les formules par exemple d'Heckman pour l'estimateur par appariement fondé sur les noyaux. Les résultats de Heckman permettent en outre de garantir la convergence de l'estimateur et sa normalité asymptotique avec une vitesse de convergence de $\sqrt{n}$. L'écart-type de l'estimateur est obtenu en appliquant les méthodes du bootstrap, ce qui consiste à répliquer l'ensemble de la procédure d'estimation sur un échantillon tiré aléatoirement avec remise dans l'échantillon initial et à déterminer l'écart-type de la distribution de l'ensemble des estimateurs ainsi obtenus. Notons que l'écart-type doit aussi prendre en compte le fait que le score n'est pas connu et est de ce fait estimé. Chaque étape du bootstrap doit ainsi comprendre non seulement l'appariement sur l'échantillon tiré mais aussi l'estimation du score. L'estimation de l'écart-type peut ainsi être coûteuse en temps de calcul.

5. Le modèle de sélectivité sur inobservables

Une solution alternative pour résoudre le problème de sélectivité est de recourir à une modélisation jointe des résultats potentiels et de l'affectation au traitement. C'est la démarche qui est usuellement suivie.

$$\begin{array}{l} y(0)_i = a_0 + x_i b_0 + u_{0i} \\ y(1)_i = a_1 + x_i b_1 + u_{1i} \end{array} \quad \text{et} \quad \begin{array}{l} T_i^* = z_i \gamma + v_i \\ T_i = 1 \Leftrightarrow T_i^* > 0 \end{array}$$

Dans cette écriture on fait l'hypothèse que les éléments inobservés u_{0i}, u_{1i} et γ sont indépendants des variables explicatives x_i et z_i. On voit donc que l'évaluation qui sera faite dépend in fine des choix effectués pour la modélisation. Les paramètres b_0, b_1 et γ sont des paramètres structurels et ont donc une interprétation économique. Il est ainsi possible d'identifier l'importance des différents facteurs conduisant à la décision de participer au traitement et affectant les résultats potentiels.

L'intérêt central de cette modélisation est de pouvoir traiter l'existence d'une dépendance entre les éléments inobservés affectant le traitement et les résultats potentiels. C'est la raison pour laquelle ce modèle est appelé modèle de sélection sur inobservables.

L'estimation d'un tel modèle ne pose aucun problème lorsque l'on spécifie la loi jointe des éléments inobservés. Par exemple dans le cas d'une loi normale, on peut écrire l'espérance de la variable de résultat pour les individus traités et les individus non traités :

$$P(T_i = 1|z_i) = \Phi(z_i\gamma)$$
$$E(y_i|x_i, z_i, T_i = 0) = E(y(0)_i|x_i, z_i, T_i = 0) = a_0 + x_i b_0 - \rho_0\sigma_0\, \phi(z_i\gamma)/(1 - \Phi(z_i\gamma))$$
$$E(y_i|x_i, z_i, T_i = 1) = E(y(1)_i|x_i, z_i, T_i = 1) = a_1 + x_i b_1 + \rho_1\sigma_1\, \phi(z_i\gamma)/\Phi(z_i\gamma)$$

Néanmoins l'accent a été mis au cours des dernières années sur la possibilité d'identification de ce type de modèles sans faire d'hypothèse sur la loi jointe des perturbations. Cette généralisation conduit à une complication considérable de l'estimation et nécessite des hypothèses identificatrices fortes. La détailler permet de bien comprendre le rôle fondamental joué par la spécification de la loi jointe des effets inobservés et que le modèle de sélectivité sur inobservable n'est donc pas plus général que le modèle de sélection sur observables. Notons d'ores et déjà que ceci requiert la présence de variables affectant la décision de participation et pas les résultats potentiels, ce qui a priori n'est pas nécessaire lorsque l'on spécifie la loi des éléments inobservés, les formes fonctionnelles des termes de biais, non linéaires en général dans les variables permettant l'identification des paramètres a_k et b_k.

D'autres hypothèses sont néanmoins nécessaires pour estimer ce type de modèle. Comme précédemment on peut écrire l'espérance des résultats potentiels sachant la variable de traitement et les variables explicatives comme une fonction de la partie structurelle du modèle et d'un terme de biais. La différence est que, dans ce cas, la forme du terme de biais n'est plus spécifiée.

On peut écrire l'expression de la variable de résultat pour les individus traités et non traités sous la forme :

$$\begin{aligned} E(y_i|x_i, z_i, T_i = 0) &= a_0 + x_i b_0 + E(E(u_i|x_i, z_i, z_i\gamma + v_i < 0)|x_i, z_i, T_i = 0) \\ &= a_0 + x_i b_0 + E(E(u_i|z_i\gamma + v_i < 0)|x_i, z_i, T_i = 0) \\ &= +a_0 + x_i b_0 + K_0(P(z_i)) \end{aligned}$$
$$E(y_i|x_i, z_i, T_i = 1) = a_1 + x_i b_1 + K_1(P(z_i))$$

où $P(z_i) = P(T_i = 1|z_i)$, les fonctions K_0 et K_1 étant non spécifiées.

Les paramètres d'intérêt que sont l'effet moyen du traitement et l'effet moyen du traitement sur la population traitée peuvent se déduire des paramètres a_0, b_0, a_1, b_1, γ et des fonctions K_0 et K_1. On a en effet :

$$E(c_i) = E(a_1 + x_i b_1 - a_0 - x_i b_0)$$
$$E(c_i | T_i = 1) = E(y_i | T_i = 1) - E\left(a_0 + x_i b_0 - \frac{1 - P(z_i)}{P(z_i)} K_0(P(z_i)) \middle| T_i = 1 \right)$$

L'estimation de ce modèle est néanmoins difficile. Il est en particulier difficile d'estimer les constantes qui jouent pourtant un rôle très important dans l'estimation de l'effet causal.

Les paramètres b_0 et b_1, peuvent être estimés en utilisant la méthode proposée par Robinson (1988) consistant à régresser le résidu de la régression non paramétrique de y_i sur la probabilité d'être traité $y_i - E(y_i | P(z_i), T_i = 0)$ sur le même résidu pour les variables explicatives $x_i - E(x_i | P(z_i), T_i = 0)$. On a en effet :

$$E(y_i - E(y_i | P(z_i), T_i = 0) | x_i, z_i, T_i = 0) = E(x_i - E(x_i | P(z_i), T_i = 0)) b_0$$
$$E(y_i - E(y_i | P(z_i), T_i = 1) | x_i, z_i, T_i = 1) = E(x_i - E(x_i | P(z_i), T_i = 1)) b_1$$

L'estimation des constantes a_0 et a_1 nécessite la connaissance de valeurs des variables affectant la décision de participer au programme pour lesquelles les fonctions K_0 et K_1 sont nulles. Pour cela on peut tirer parti du fait que les résidus ne sont pas corrélés avec les variables explicatives. On a ainsi :

$$\begin{aligned} E(u_0 | x_i, z_i) &= E(u_0 | x_i, z_i, T_i = 0) P(T_i = 0 | z_i) + E(u_0 | x_i, z_i, T_i = 1) P(T_i = 1 | z_i) = 0 \\ &= K_0(P(z_i))(1 - P(z_i)) + E(u_0 | x_i, z_i, T_i = 1) P(z_i) \end{aligned}$$

soit $K_0(0) = 0$. De même on a $K_1(1) = 1$.

Les fonctions $a_0 + K_0(P(z_i))$ et $a_1 + K_1(P(z_i))$ peuvent être facilement estimées à partir des relations :

$$E(y_i - x_i b_0 | T_i = 0, P(z_i)) = a_0 + K_0(P(z_i)) = G_0(P(z_i))$$
$$E(y_i - x_i b_1 | T_i = 1, P(z_i)) = a_1 + K_1(P(z_i)) = G_1(P(z_i))$$

en considérant par exemple des estimateurs non paramétriques à noyau. Si ces fonctions peuvent être estimées aux bornes du support, c'est-à-dire en 0 et en 1, on pourra faire le partage entre les constantes et les fonctions.

On peut noter que ceci constitue une restriction supplémentaire importante : il est nécessaire qu'il y ait des individus traités ayant une probabilité 1 d'être traités. De même, il est nécessaire qu'il y ait des individus non traités ayant une probabilité nulle d'être traités. Enfin on peut noter qu'étant basées sur un estimateur non paramétrique, les constantes a_0 et a_1 sont très imprécisément estimées : l'estimation de l'effet moyen du traitement dans le modèle de sélection sur inobservable semi-paramétrique est nécessairement très imprécise.

6. Ces méthodes marchent-elles, quelles sont leurs limites et que coûtent-elles ?

Il y a ainsi plusieurs façons de traiter le problème de la sélectivité, correspondant à différentes hypothèses. Il est important en pratique de savoir laquelle est la mieux adaptée. Les travaux de Heckman et al. (1997 – 1998) apportent quelques éléments de réponse à cette question. Ils utilisent les information du Joint Training Partnership Act, programme de formation aux Etats-Unis qui a été conçu comme une expérience contrôlée. Il est donc possible d'obtenir une évaluation de l'effet de ce programme en utilisant le groupe de contrôle expérimental. En utilisant un groupe de contrôle non expérimental, il est alors possible de mettre en œuvre les différents estimateurs précédents et d'examiner ceux qui se rapprochent le plus de l'évaluation sur données expérimentales. Les enseignements en sont les suivants :

- Les méthodes d'appariement sur les niveaux ne marchent pas. Il y a persistance d'effets individuels inobservés même lorsque l'on introduit un grand nombre de caractéristiques.

- Les modèles de sélection sur inobservables semi-paramétriques marchent bien mais sont très imprécis.

- L'hypothèse de normalité dans les modèle de sélection sur inobservables conduit à des estimateurs biaisés

- Les méthodes d'appariement sur les évolutions, estimateurs par “ différence de différence ”, marchent bien, mais ceci dépend néanmoins de la richesse des variables de conditionnement.

- Les problèmes de support dans ces méthodes sont importants. Ceci souligne l'hétérogénéité de l'effet du traitement dans la population et l'impossibilité de l'extrapoler.

7. Conclusion : ces méthodes marchent-elles, quelles sont leurs limites et que coûtent-elles ?

Dans cette contribution, on a présenté les travaux récents sur l'évaluation de certaines politiques économiques, s'inspirant des travaux des statisticiens concernant l'évaluation des traitements dans le domaine médical. Une caractéristique importante du cadre causal développé par les statisticiens est de mettre l'accent sur l'hétérogénéité de l'effet du traitement dans la population et sur l'impossibilité de l'observer au niveau individuel. Ce cadre conduit ainsi naturellement à se concentrer sur les conditions d'identification de certains paramètres de la distribution correspondante. On a insisté sur l'hypothèse d'indépendance conditionnellement à des observables et montré leur généralité en les opposant aux méthodes usuelles consistant à privilégier l'hypothèse d'indépendance conditionnellement à des observables et des inobservables et à modéliser le comportement des individus.

Néanmoins, ces méthodes souffrent de limitations certaines. En premier lieu, elles n'offrent qu'une validation de l'existence d'un effet. Si elles s'affranchissent des contraintes liées à la modélisation des comportements des individus pour l'identification de l'effet de ces politiques, elles ne permettent pas non plus d'apprendre sur les comportements à l'œuvre et de mesurer leur ampleur respective. Elles sont donc peu adaptées pour guider l'orientation à donner à ces politiques et ne permettent pas en particulier d'extrapoler les résultats pour procéder à des évaluations ex ante. En outre, le choix des variables de conditionnement retenues pour procéder à l'évaluation est assez arbitraire et est susceptible d'influencer fortement les résultats. Le recours à une modélisation du comportement des individus, autant de leur décision de participer à un dispositif que celui des variables de résultat potentiel, s'il n'intervient pas explicitement au travers de l'estimation de paramètres structurels ne peut ainsi être totalement éliminé. Ces méthodes ne permettent donc de s'affranchir de la contrainte de la modélisation qu'en apparence, et sans pouvoir nécessairement en recueillir tous les bénéfices. Ces méthodes ne permettent de plus que des estimations locales en raison de la nécessité de se restreindre au support commun de la densité du score pour les individus traités et non traités. En outre les hypothèses nécessaires à l'identification de l'effet moyen du traitement dans la population sont plus fortes que celles nécessaires à l'identification de l'effet moyen du traitement sur les individus traités. Il est donc au total difficile d'extrapoler les résultats de ces estimations pour en déduire des évaluations macro-économiques.

Le champ d'application de ces méthodes reste encore restreint, puisqu'elles ne permettent que l'évaluation de dispositifs s'apparentant à un traitement unique. Néanmoins, ce domaine est en plein développement et le champ d'application de ces méthodes s'étend : Lechner (1999), Brodaty, Crépon et Fougère (1999) examinent ainsi le cas dans lequel le traitement peut prendre différentes formes (choix entre différents stages de formation par exemple). Crépon et Desplatz (2000) examinent la situation dans laquelle le traitement peut prendre des valeurs dans un ensemble continu et appliquent ces méthodes à l'évaluation des réductions de charges sur les bas salaires.

Enfin, une méthode alternative pour évaluer le type de politique entrant dans le cadre précédemment développé serait de procéder à des expérimentations. Cette façon de procéder, qui est pratiquée dans de nombreux pays (voir Fougère (1999) pour une revue), a néanmoins un coût social élevé, puisqu'elle consiste à exclure aléatoirement du traitement une partie de la population souhaitant entrer dans le dispositif et en faire le groupe de contrôle. La conclusion qui émerge des différentes études utilisant les méthodes que l'on a présentées est qu'il est possible de s'affranchir de ce coût social et donc d'évaluer les dispositifs sans procéder à des expérimentations. Ceci ne veut pas dire néanmoins que le recours à ses méthodes ne comporte pas lui aussi un coût : de fait, pour que ces méthodes soient valides et donc être utilement mises en œuvre, il est nécessaire de disposer d'un vaste ensemble d'informations sur la population entrant dans les dispositifs et celle susceptible de former le groupe de contrôle non expérimental. Si ces méthodes permettent d'éviter le coût social des expérimentations, elles comportent donc en revanche un coût statistique.

Références

Andrews D. (1991) " Asymptotic normality of series estimators for non parametric and semiparametric regression models " Econometrica 59(2) 307-345.

Brodaty T., B. Crépon and D. Fougère (1999) : " Using matching estimators to evaluate alternative youth employment programs: evidence from France 1986-1988 " forthcoming in " The Evaluation of Active Labour Market Policies in Europe " M. Lechner eds Springer Verlag

Barnow, B., G. Cain, and A. Goldberger (1980), " Issues in the analysis of selectivity bias ", in : E. Stromsdorfer and G. Farkas, eds., Evaluation Studues, vol. 5 (Sage Publications, Beverly Hills, CA) 290-317.

Crépon, B. et R. Desplatz (2000) " Evaluer l'effet sur l'emploi des réductions de charges sur les bas salaires ", mimeo

Fougère D. (2000) : " 'Exprimenter pour évaluer les politiques d'aides à l'emploi : les exemples anglo-saxons et nord européens '" mimeo.

Heckman, and Hotz (1989) " Choosing among alternative methods of estimating the impact of social programs : the case of manpower training " Journal of the American Statistical Association 84(408) :862-874.

Heckman, J. H. Ichimura, J. Smith and P. Todd (2000) " Characterizing selection bias using experimental data " Econometrica 66(5) : 1017-1098.

Heckman, J. H. Ichimura, and P. Todd (1997) “ Matching as an econometric evaluation estimator : evidence from evaluating a job training program ” Review of Economic Studies 64(4) :605 -654.

Heckman, J. H. Ichimura, and P. Todd (1998) “ Matching as an econometric evaluation estimator” Review of Economic Studies 65(2) :261-294.

Heckman J. R. Lalonde and J. Smith (1999) : “ 'The economics and econometrics of active labor market programs '" in Handbook of Labor Economics vol. III, O. Ashenfelter and D. Card eds North Holland.

Heckman, J. and R. Robb (1985) “ Alternative methods for evaluating the impact of interventions : an overview ” Journal of Econometrics, 30(1-2) :239-267.

Holland, P. (1986) “ Statistics and causal inference ”, Journal of the American Statistical Association 81(396) :945 :960.

Imbens, G. and J. Angrist (1994) “ Identification and estimation of local average treatment effect s” Econometrica, 62(4) :467-476.

Lalonde, R. (1986) “ Evaluating the econometric evaluations of training programs with experimental data ” American Economic Review 76(4) :604-620.

Lechner (1999) : Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independance Assumption" forthcoming in “ The Evaluation of Active Labour Market Policies in Europe ” M. Lechner eds Springer Verlag

Robinson, P.M. (1988) : “ Root-N-Consistent Semiparametric Regression ” Econometrica, 56 : 931-954.

Rosenbaum, P. and D. Rubin (1983), “ the central role of the propensity score in observational studies for causal effects ” Biometrika 70(1) :41-55.

Rubin, D. (1974) “ Estimating causal effects of treatments in randomized and non randomized Studies ”, Journal of Educational Psychology 66 :688-701.

Rubin D. (1977) : "Assignement to Treatment Group on the Basis of a Covariate" Journal of Educational Statistics Spring vol. 2 (1).