Chapitre 5: L'estimateur de la différence de différences (DD)

- Cet estimateur, au contraire de celui du score de propension, autorise la présence d'une hétérogénéité non observée, mais celle-ci est supposée être invariante au cours du temps.
- Sous cette hypothèse, l'emploi de données recueillies sur les participants et sur des non participants au programme, avant et après sa mise en place permet d'obtenir un estimateur sans biais de l'efficacité du programme, en éliminant cette hétérogénéité supposée *fixe* et *additive*.

- Les méthodes de randomisation et du score de propension sont des méthodes de « simple différence », qui peuvent être employées lorsque l'on ne dispose que d'une seule observation sur les individus appartenant aux échantillons test et de contrôle.
- La méthode de la double différence est appropriée lorsque l'on dispose de données de panel, c'est à dire d'au moins deux observations sur chaque individu.

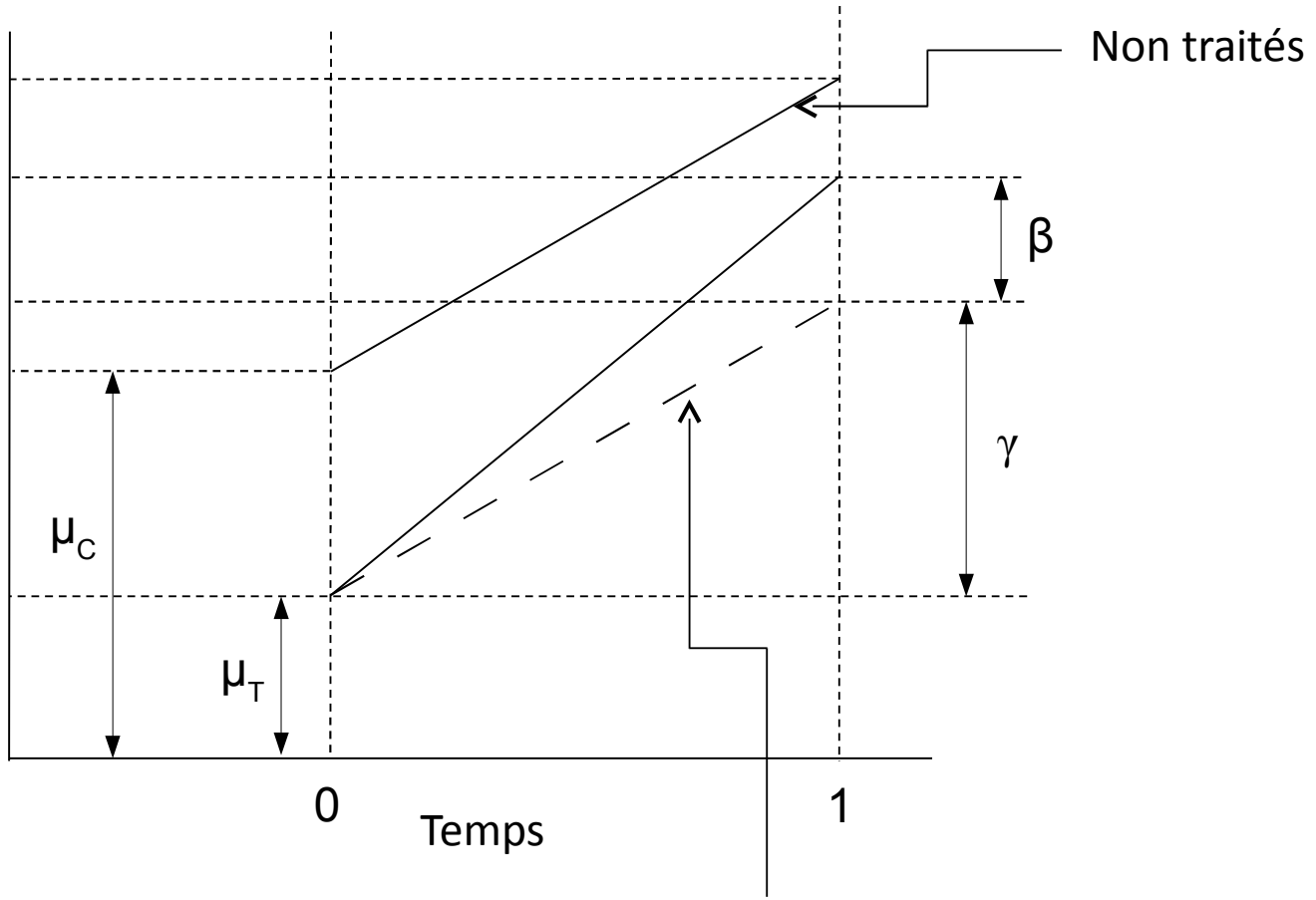
- Supposons donc que les individus sont observés à deux dates, 0 et 1. En t=0 le programme n'existe pas. En t=1 le programme est en place. On note $Y_{i,t}(0)$ la valeur de la variable d'intérêt pour les individus non traités à la date t et $Y_{i,t}(1)$ cette valeur pour les individus traités. L'estimateur de la double différence s'écrit:

$$DD = E\left(Y_{i,1}(1) - Y_{i,0}(1) | T_i = 1\right) - E\left(Y_{i,1}(0) - Y_{i,0}(0) | T_i = 0\right)$$

- Il est souvent plus commode de calculer l'estimateur de la double différence en employant une méthode de régression linéaire. Le modèle à estimer est alors de la forme:

$$Y_{it} = \alpha + \beta T_i t + \rho T_i + \gamma t + \epsilon_{it}$$

- Dans cette régression, le coefficient du terme d'interaction entre le temps, t, et la variable de traitement T_i, β, est l'estimateur de la double différence DD. Les variables t et T_i sont incluses séparément pour tenir compte d'un effet potentiel du temps qui passe (trend) et d'un effet provenant du fait d'être, ou non, inclus dans l'échantillon test.



$$\rho = \mu_c - \mu_T$$

$$Y_{it} = \alpha + \beta T_i t + \rho T_i + \gamma t + \epsilon_{it}$$

- Il est facile de montrer que les estimations des coefficients de cette régression conduisent au résultat attendu:

$$Y_{it} = \alpha + \beta T_i t + \rho T_i + \gamma t + \epsilon_{it}$$

$$E(Y_{i1} - Y_{i0} | T_i = 1) = E(Y_{i1}(1) - Y_{i0}(1) | T_i = 1)$$
$$= \alpha + \beta + \rho + \gamma - \alpha - \rho$$

et

$$E(Y_{i1} - Y_{i0} | T_i = 0) = E(Y_{i1}(0) - Y_{i0}(0) | T_i = 0)$$
$$= \alpha + \gamma - \alpha$$

- Notons que si l'on se contentait d'une simple différence, en évaluant l'écart dans la valeur de la variable Y pour les individus traités avant et après le traitement, on obtiendrait un estimateur biaisé. En effet, dans cette hypothèse ce qui est calculé est:

$$E\left(Y_{i1}-Y_{i0}\,|\,T_i=1\right)=\beta+\gamma$$

Et le biais est précisément le coefficient de t dans la régression, lequel mesure l'évolution temporelle indépendante de Y.

- De même, évaluer l'impact du programme en comparant les valeurs de Y en t=1 entre les groupes test et de contrôle conduit à calculer:

$$E(Y_{i1}|T_i=1)-E(Y_{i1}|T_i=0)=\alpha+\beta+\rho+\gamma-\alpha-\gamma$$
$$=\beta+\rho$$

où cette fois le biais est égal au terme qui marque la sélection dans le programme.

- L'emploi de l'estimateur DD repose sur l'hypothèse que la double différentiation suffit à retirer toute l'hétérogénéité non observée et potentiellement corrélée à la probabilité de recevoir le traitement:

$$Y_{it} = \alpha + \beta T_i t + \rho T_i + \gamma t + \epsilon_{it}$$

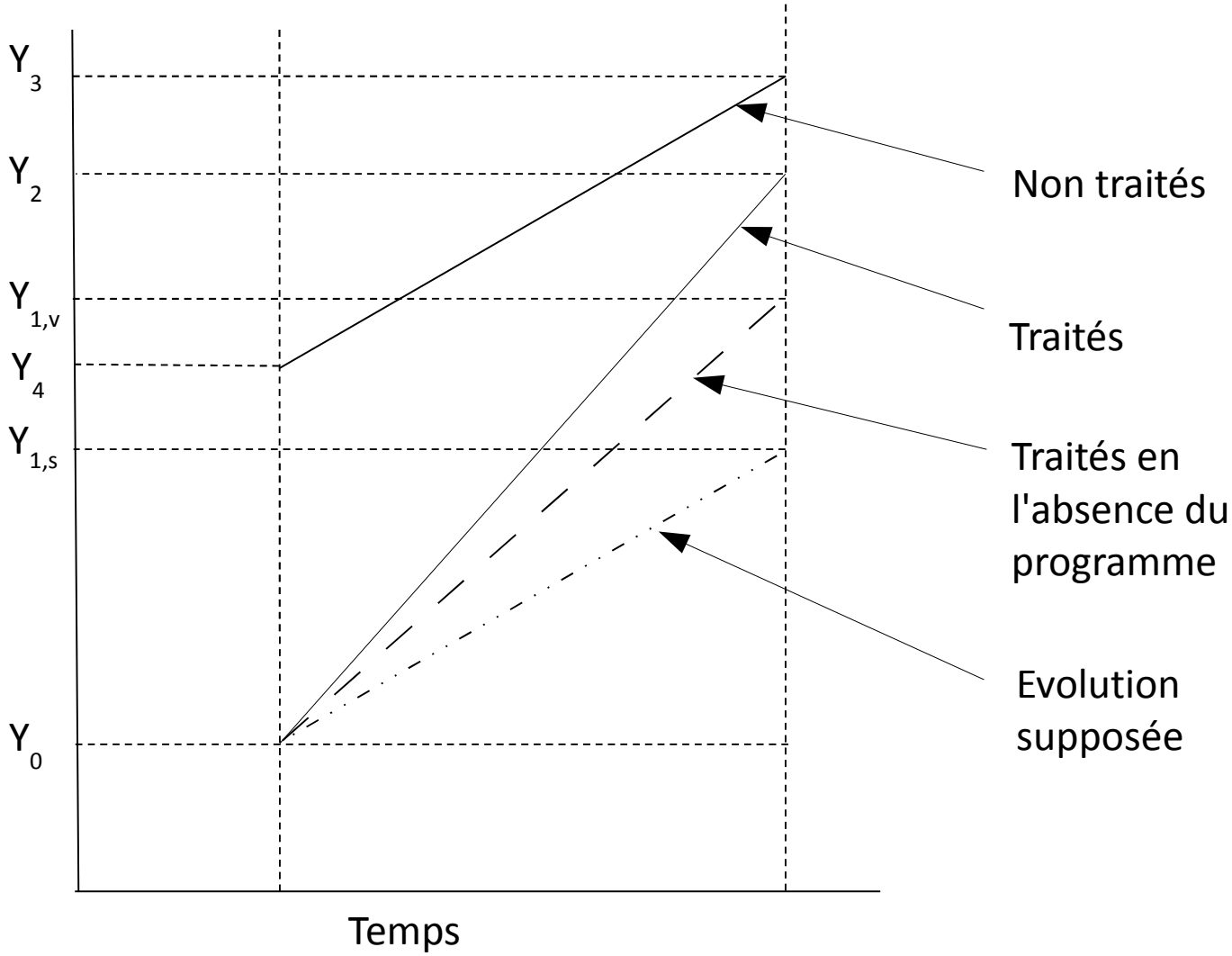
autrement dit:

$$cov(\epsilon_{it}, T_i) = 0$$
$$cov(\epsilon_{it}, t) = 0$$
$$cov(\epsilon_{it}, T_i t) = 0$$

- La dernière hypothèse signifie que les caractéristiques non observées qui affectent la participation au programme ne doivent pas varier au cours du temps en fonction de la sélection dans l'un ou l'autre groupe.

- Cette dernière hypothèse est en fait problématique, particulièrement lorsqu'il s'agit d'évaluer des programmes sociaux ou de lutte contre la pauvreté.
- Supposons par exemple que l'on cherche à évaluer l'efficacité d'un programme ciblé sur certaines régions et destiné à offrir des conditions favorables à la croissance des revenus de leurs habitants.
- Les régions ciblées peuvent l'avoir été (ou s'être autosélectionnées) parce qu'elles ont connu une baisse de leurs revenus (suite à un choc exogène) juste avant l'introduction du programme.

- Ces régions, comparées à d'autres n'ayant pas connu le choc, peuvent alors connaître un processus de rattrapage, qui va entraîner une hausse de leurs revenus plus importante et indépendante de leur participation au programme.
- Si cette hausse n'est pas prise en compte, l'estimateur DD conduira à une sur-estimation l'efficacité de la politique



L'estimateur DD repose sur l'hypothèse que l'évolution hors traitement de Y entre les deux groupes est identique: DD = $(Y_2-Y_0)-(Y_3-Y_4)$ et l'on suppose que $Y_3-Y_4 = Y_{1,s}-Y_0$, de sorte que DD = $Y_2-Y_{1,s}$ est supérieur à la véritable valeur de l'effet du traitement: $Y_2-Y_{1,v}$. Ce problème est connu dans la littérature sous le nom de « Ashenfelter Dip ».

- Ainsi, l'emploi de l'estimateur DD ne dispense de s'assurer que les échantillons test et de contrôle présentent des caractéristiques identiques avant l'application du traitement.
- Pour réduire les sources de biais potentiels, on peut par exemple envisager de combiner les méthodes PSM et DD.

- L'estimateur DD suppose que l'on dispose d'au moins deux dates d'observation.
- Sur les données de l'année de base, on applique la méthode PSM afin d'identifier les individus qui, dans l'ensemble des observations non traitées, sont le mieux appariés aux individus traités.
- Ensuite on applique la méthode DD en calculant la différence de variation moyenne de la variable d'intérêt entre l'échantillon des traités et celui des individus qui leurs sont appariés:

$$DD_{PSM} = \frac{1}{N_T} \sum_{i=1}^{N_T} \left[(Y_{i,1}^{T} - Y_{i,0}^{T}) - \sum_{j \in M(i)} \omega(j)(Y_{j,1}^{C} - Y_{j,0}^{C}) \right]$$

où M(i) désigne l'ensemble des observations appariées à i et ω(j) le poids attribué à l'observation j.

L'estimateur de la triple différence

- Il arrive souvent que l'on ne dispose pas d'observations avant la mise en place du programme testé. Dans un tel cas il est parfois possible de calculer une triple différence. Le principe est d'utiliser les observations d'un second groupe de contrôle non concerné par le programme, mais de caractéristiques proches du groupe traité.

- Par exemple, Ravallion et alli. (2005) évaluent l'impact d'un programme d'emploi public, ciblé sur les populations pauvres en Argentine. L'échantillon de base est constitué de 1500 participants observés à trois reprises entre mai-juin 1999 et mai-juin 2000. Certains participants maintiennent leur participation (les « permanents ») sur la totalité de la période, alors que d'autres le quittent (les « défaillants »). L'évaluation porte sur le revenu et consiste à comparer celui des permanents avec celui des défaillants.

- La difficulté est que les opportunités de gain en l'absence du programme des « permanents » et des « défaillants » ne sont pas forcément les mêmes, puisque la décision de quitter le programme n'est pas prise au hasard.
- Ravallion et alli. utilisent les données d'une enquête auprès des ménages réalisée au même moment pour construire un échantillon de contrôle pour le groupe des « permanents » et celui des « défaillants ». Deux vagues de l'enquête ont été réalisées à 6 mois d'intervalle, ce qui permet d'obtenir un panel d'observations sur la période.
- Le questionnaire employé pour interroger les participants au programme est le même que celui de l'enquête ménage, ce qui permet d'éviter certains biais liés à une façon différente de poser les questions.

- La triple différence est alors construite en plusieurs étapes:
  - On commence par apparier chaque participant à un non participant en utilisant une méthode de type PSM.
  - Ensuite, on apparie chaque « permanent » à un « défaillant » de nouveau en recourant à une méthode de type PSM.
  - On calcule une première double différence, A, égale à la différence des revenus en début (1) et en fin (2) de période d'observation, entre les « permanents » et les non participants qui leurs sont appariés:

$$A=\frac{1}{N_T^P}\sum_{i=1}^{N_T^P}\left[(Y_{i,2}^{T,P}-Y_{i,1}^{T,P})-\sum_{j\in M(i)}\omega(j)(Y_{i,2}^{C,P}-Y_{i,1}^{C,P})\right]$$

  - Puis une seconde double différence, B, cette fois entre les « défaillants » et les non participants qui leur sont appariés:

$$B=\frac{1}{N_T^D}\sum_{i=1}^{N_T^D}\left[(Y_{i,2}^{T,D}-Y_{i,1}^{T,D})-\sum_{j\in M(i)}\omega(j)(Y_{i,2}^{C,D}-Y_{i,1}^{C,D})\right]$$

  - L'estimateur de la triple différence, DDD, est alors: DDD = A-B

- Cet estimateur peut également être écrit sous la forme:

$$
\begin{aligned}
DDD = & \left[\left(\bar{Y}_2^{T,P} - \bar{Y}_1^{T,P} \right) - \left(\bar{Y}_2^{C,P} - \bar{Y}_1^{C,P} \right) \right] \\
& - \left[\left(\bar{Y}_2^{T,D} - \bar{Y}_1^{T,D} \right) - \left(\bar{Y}_2^{C,D} - \bar{Y}_1^{C,D} \right) \right] \\
= & \left[\left(\bar{Y}_2^{T,P} - \bar{Y}_1^{T,P} \right) - \left(\bar{Y}_2^{T,D} - \bar{Y}_1^{T,D} \right) \right] \\
& - \left[\left(\bar{Y}_2^{C,P} - \bar{Y}_1^{C,P} \right) - \left(\bar{Y}_2^{C,D} - \bar{Y}_1^{C,D} \right) \right]
\end{aligned}
$$

- Le premier terme de droite de la seconde égalité est le gain des « permanents » par rapport aux « défaillants ». Le second terme mesure le biais qui résulte de l'auto-sélection des « permanents » et des « défaillants ».

Chapitre 6: Régressions discontinues et méthode du Pipeline

Principe

- L'idée est ici d'exploiter les particularités du programme pour se rapprocher du cadre offert par la randomisation:
  - Conditions d'éligibilité au programme
  - Mise en place progressive du programme
- Il s'agit d'exploiter une règle d'allocation exogène des participants et des non participants, afin d'obtenir un échantillon de contrôle réunissant des individus semblables aux individus de l'échantillon test.

Exemples

- L'activité de la Grameen Bank est ciblée sur les ménages d'agriculteurs possédant moins d'un demi acre de terre.
- Les règles relatives au nombre maximum d'élèves par classe peuvent induire des variations exogènes dans le nombre d'élèves par classe.
- Les pensions liées au risque vieillesse sont réservées à des individus au délà d'un certain âge (minimum vieillesse par ex.)
- Les bourses d'études peuvent être réservées à des étudiants ayant obtenu une note minimum à un examen déterminé.
- Les programmes ciblés sur les établissements scolaires ou de santé peuvent être mis en place progressivement.
- Etc.

Régressions sur discontinuités

- Quel que soit le cas de figure, le principe est d'utiliser des observations sur des individus à la marge du critère d'éligibilité pour évaluer l'efficacité du programme, en supposant que les différences entre ceux qui sont de part et d'autre de la frontière, mais suffisamment proches, sont assez faibles pour ne pas biaiser l'estimation (par ex. on compare ceux qui ont eu une moyenne comprise entre 14,9 et 15 à l'examen avec ceux ayant obtenu entre 15 et 15,1).

Théorie

- Deux cas peuvent se présenter:
  - Soit le critère d'éligibilité est appliqué de façon stricte (« Sharp design). Le traitement, T, dépend alors de manière déterministe de la valeur prise par une variable de sélection S:

$$T_i = 1_{\{S_i > S^*\}}$$

  - Soit la variable de sélection, S, ne fait qu'affecter la probabilité de participation (on parle alors de « Fuzzy design »). Par exemple, un programme de distribution de bourses peut dépendre de la note aux examens et d'autres critères (revenu etc.) et de plus le fait d'être éligible n'implique pas que l'on participe. Dans ce cas, on suppose que:

$$P(T_i = 1 \mid S_i > S^*) > P(T_i = 1 \mid S_i < S^*)$$

Conditions d'identification

- Si l'on écrit la relation entre la variable d'intérêt, Y, et le traitement, T sous la forme:

$$Y_i = \alpha + \beta T_i + \psi_i$$

- L'effet du traitement est alors donné par:

$$E(Y_i | T_i = 1) - E(Y_i | T_i = 0) = \beta$$

- « Sharp design »: l'identification de l'effet du traitement repose sur l'hypothèse que la variation de la variable d'intérêt en l'absence du programme et l'effet du traitement, autour de la valeur critique de la variable S, sont continus. Ce qui se traduit par les conditions suivantes (pour toute valeur de ε arbitrairement petite):

$$E(\psi_i | S_i = S^* - \epsilon) = E(\psi_i | S_i = S^* + \epsilon)$$

et

$$E(\beta T_i | S_i = S^* + \epsilon) = E(Y_i | T_i = 1, S_i = S^*) - E(Y_i | T_i = 0, S_i = S^*)$$

- Dans ce cas, l'effet du traitement est évalué par:

$$E(Y_i | T_i = 1) - E(T_i | T_i = 0) = \lim_{\epsilon \to 0} \left[E(Y_i | S_i = S^* + \epsilon) - E(Y_i | S_i = S^* - \epsilon) \right]$$

- « Fuzzy design »: on peut écrire

$$\lim_{\epsilon \to 0}\left[E(Y_i|S_i=S^*+\epsilon)-E(Y_i|S_i=S^*-\epsilon)\right]=\lim_{\epsilon \to 0}\left[E(\beta T_i|S_i=S^*+\epsilon)-E(\beta T_i|S_i=S^*-\epsilon)\right]$$
$$-\lim_{\epsilon \to 0}\left[E(\psi_i|S_i=S^*+\epsilon)-E(\psi_i|S_i=S^*-\epsilon)\right]$$

- Si l'effet du traitement est constant au voisinage du point de discontinuité, il est alors donné par le ratio:

$$\beta=\frac{\lim_{\epsilon \to 0}\left[E(Y_i|S_i=S^*+\epsilon)-E(Y_i|S_i=S^*-\epsilon)\right]}{\lim_{\epsilon \to 0}\left[E(T_i|S_i=S^*+\epsilon)-E(T_i|S_i=S^*-\epsilon)\right]}$$

- Remarquons que l'écriture du dénominateur n'est pas triviale, puisque par hypothèse $E(T_i|S_i=S^*+\varepsilon)\neq 1$ et $E(T_i|S_i=S^*-\varepsilon)\neq 0$.

Limites

- La principale limite de ce type d'estimateur est leur caractère local (Local Average Treatment Effect): la comparaison effectuée n'est en effet valable que pour des individus « localisés » près de la frontière.
- Les résultats obtenus avec ce type d'estimation n'ont donc pas a priori une portée générale.

Mise en oeuvre

- Il faut commencer par vérifier que les variables de traitement, T, et d'intérêt, Y, présentent bien une discontinuité au point S=S*.
- Dans l'idéal il faut également s'assurer que Y ne présente pas d'autres discontinuités non expliquées, afin de se garantir contre le risque de confondre l'effet du traitement avec autre chose.
- Enfin il également préférable de vérifier que les autres déterminants de Y ne présentent pas de discontinuité aux alentours de S*.

- Ensuite il faut choisir la largeur, 2h, de la bande autour de la frontière S=S* dans laquelle seront retenues les observations employées pour calculer l'estimateur.
- Une fois cette largeur choisie, plusieurs méthodes peuvent être employées. La plus simple consiste à faire la moyenne de Y pour les observations situées dans les intervalles]S*-h,S*[et [S*,S*+h[et ensuite à faire la différence que l'on divisera ensuite par la moyenne de T sur chacun de ces deux intervalles:

$$\hat{\beta} = \frac{\frac{1}{N_h^+}\sum_{i \in T_h^+} Y_i - \frac{1}{N_h^-}\sum_{i \in T_h^-} Y_i}{\frac{1}{N_h^+}\sum_{i \in T_h^+} T_i - \frac{1}{N_h^-}\sum_{i \in T_h^-} T_i}$$

où le numérateur est égal à l'unité dans le cas d'un « sharp design ».

- En pratique cet estimateur a toutes les chances de ne pas être très robuste.
- En effet, il faut avoir beaucoup d'observations dans la bande afin d'obtenir des résultats qui ne sont pas sensibles aux valeurs prises par quelques observations.
- Il peut alors être nécessaire de choisir une bande assez large et de faire l'hypothèse que la valeur de Y ne dépend pas de la distance à la frontière, ce qui est souvent une hypothèse forte.
- Il est donc en général préférable de recourir à des méthodes plus élaborées qui reposent sur des hypothèses moins fortes.

- Par exemple, on peut effectuer des régressions locales, en se restreignant aux observations appartenant à la bande et en régressant Y sur un polynome de S-S* (estimation semi-paramétrique):

$$Y_i = y_- + b_1(S_i - S^*) + b_2(S_i - S^*)^2 + b_3(S_i - S^*)^3 + \ldots + b_k(S_i - S^*)^k + u_i \text{ pour les observations telles que } S_i < S^*$$

$$Y_i = y_+ + b_1(S_i - S^*) + b_2(S_i - S^*)^2 + b_3(S_i - S^*)^3 + \ldots + b_k(S_i - S^*)^k + u_i \text{ pour les observations telles que } S_i \geq S^*$$

- L'estimateur, dans le cas du « Sharp design » est alors simplement la différence:

$$RD = y_+ - y_-$$

- Pour le « Fuzzy design », il faut en plus estimer la probabilité de recevoir le traitement à gauche et à droite du point de discontinuité et l'estimateur RD s'écrit alors:

$$RD = \frac{y_+ - y_-}{t_+ - t_-}$$

Méthode du Pipeline

- Comme indiqué en introduction le principe est ici d'utiliser la mise en place progressive du programme pour en identifier l'impact en comparant les valeurs de la variable d'intérêt Y entre le groupe de ceux qui bénéficient de la mesure et le groupe (ou un sous ensemble de celui-ci) de ceux qui n'en bénéficient pas encore.
- Ce principe peut être combiné avec d'autres méthodes comme par exemple les doubles différences ou la méthode du score de propension.

- Par exemple, si l'on dispose des observations issues d'une enquête réalisée avant le démarrage du programme (base) et d'observations issues d'une seconde enquête réalisée une fois le programme mis en place pour certains individus, mais pas pour tous.
- L'estimateur peut alors être calculé par la double différence entre les valeurs moyennes de Y avant et après le démarrage du programme pour les observations traitées et non traitées.
- On peut aussi envisager de combiner cet estimateur avec un appariement des observations traitées et non traitées selon la méthode PSM.

- Si aucune enquête de base n'est disponible, la méthode PSM peut néanmoins être employée pour apparier les individus déjà traités avec des individus non encore traités, sur la base de leurs caractéristiques observables.