Le modèle d'analyse de sensibilité de Rosenbaum

Simon Quantin* Simon Bunel* Clémence Lenoir**

*Insee

**Direction Générale du Trésor

31 mars 2022

Analyse de sensibilité - principe

Critique usuelle : « apparier sur les caractéristiques $\mathbf{x}$ n'est pas suffisant, car il existe une caractéristique inobservée , dont il faudrait tenir compte ».

Utiliser les données pour apporter une réponse tangible : remplacer « corrélation n'implique pas causalité » par « pour disqualifier toute causalité dans la corrélation observée, le biais de sélection inobservée doit être de *telle* amplitude ».

Quantifier l'incertitude liée à une caractéristique inobservée : comme un intervalle de confiance tient compte de l'incertitude lié à l'échantillonnage.

La démarche s'appuie sur les tests d'hypothèses, et plus précisément sur la distribution des statistiques de tests.

Introduction

- S paires appariées, $i = 1,2$, une traitée, une contrôle
- **Traitement** : recourir pour la 1re fois au dispositif JEI à un âge donné $(\mathrm{Z}_{si} = 1)$ *versus* ne jamais bénéficier du dispositif $(\mathrm{Z}_{si} = 0)$
- **Variable d'intérêt** : écart d'emploi salarié ETP entre année 1 et t (R_{si})
- **Appariement sur** $\mathbf{x}_{si}$, $\mathbf{x}_{s1} = \mathbf{x}_{s2}$
- **Sensibilité à** : une caractéristique u_{si}, avec éventuellement $u_{s1} \neq u_{s2}$, non utilisée pour l'appariement.

Introduction

- Neyman (1923); Rubin (1974) : chaque entreprise i de la paire s a deux réponses potentielles, $r_{\mathrm{T}si}$ si traitée, $\mathrm{Z}_{si} = 1$, et $r_{\mathrm{C}si}$ si contrôle, $\mathrm{Z}_{si} = 0$;
- **Effet du traitement :** $\tau_{si} = r_{\mathrm{T}si} - r_{\mathrm{C}si}$ sur entreprise jamais observé
- **Réponse observée :** s'exprime en fonction de $r_{\mathrm{T}si}$ et $r_{\mathrm{C}si}$

$$\mathrm{R}_{si} = \mathrm{Z}_{si} r_{\mathrm{T}si} + (1 - \mathrm{Z}_{si}) r_{\mathrm{C}si}$$

- **Probabilité de recourir au dispositif** : π_{si} inconnues

$$\pi_{si} = \mathrm{P}(\mathrm{Z}_{si} = 1 \mid r_{\mathrm{T}si}, r_{\mathrm{C}si}, \mathbf{x}_{si}, u_{si}) = \mathrm{P}(\mathrm{Z}_{si} \mid \mathscr{F}_{si})$$

- **Une traitée au sein de chaque paire** $\left(\mathscr{Z} = \{\mathrm{Z}_{s1} + \mathrm{Z}_{s2} = 1\}\right)$
  La probabilité pour que l'entreprise 1 de la paire s soit traitée :

$$\mathrm{P}(\mathrm{Z}_{s1} = 1, \mathrm{Z}_{s2} = 0 \mid \mathscr{F}_{s1}, \mathscr{F}_{s2}, \mathscr{Z}) = \frac{\pi_{s1}}{\pi_{s1} + \pi_{s2}}$$

Note : on notera désormais $\mathrm{P}(\mathrm{Z}_{s1} = 1, \mathrm{Z}_{s2} = 0)$ pour désigner cette probabilité conditionnelle

Introduction

Considérons le test de l'hypothèse d'absence d'effet :

$$\mathrm{H}_0 : \ \tau_{si} = 0, \ \textit{versus} \ \mathrm{H}_1 : \ \tau_{si} \neq 0,$$

Sous H_0, les valeurs observées, R_{si}, sont égales aux valeurs sans traitement $r_{\mathrm{C}si}$. La différence de revenu D_s au sein d'une paire entre l'entreprise traitée et non traitée s'écrit :

$$\begin{aligned} \mathrm{D}_s &= (\mathrm{Z}_{s1} - \mathrm{Z}_{s2})(\mathrm{R}_{\mathrm{C}s1} - \mathrm{R}_{\mathrm{C}s2}) \\ &= (\mathrm{Z}_{s1} - \mathrm{Z}_{s2})(r_{\mathrm{C}s1} - r_{\mathrm{C}s2}) \\ &= \pm(r_{\mathrm{C}s1} - r_{\mathrm{C}s2}) \end{aligned}$$

selon l'entreprise traitée dans la paire $\left(\mathrm{Z}_{s1} - \mathrm{Z}_{s2} = \pm 1\right)$.

D_s peut prendre deux valeurs, dont la probabilité de réalisation dépend de :

$$\mathrm{P}(\mathrm{Z}_{s1} = 1, \mathrm{Z}_{s2} = 0) = \frac{\pi_{s1}}{\pi_{s1} + \pi_{s2}}.$$

Cas « naïf » idéal

Recourir au dispositif ne dépend pas de u, une fois pris en compte $\mathbf{x}$.
Au sein d'une même paire, la probabilité d'être traitée est la même pour les deux entreprises

$$\pi_{s1} = \pi_{s2}$$

(mais différentes d'une paire à l'autre et inconnues).

Par contre, pour toutes les paires on a :

$$\mathrm{P}(\mathrm{Z}_{s1} = 1, \mathrm{Z}_{s2} = 0) = \frac{1}{2}$$

> *Au sein de chaque paire, il y a 50 % de chances pour l'entreprise* 1 *soit la traitée (et autant pour que ce soit l'entreprise* 2*).*

D_s peut prendre deux valeurs $\pm(\mathrm{R}_{s1} - \mathrm{R}_{s2})$, chacune avec la probabilité $\frac{1}{2}$.

La distribution de la statistique de test est alors connue.

Analyse de sensibilité

Modèle d'analyse de sensibilité : u impacte la probabilité de recourir au dispositif.

Les probabilités π_{s1} et π_{s2} varient toujours d'une paire à l'autre, sont toujours inconnues, mais cette fois différentes au sein d'une paire.

Dans le modèle d'analyse de sensibilité, l'ampleur de cette divergence est bornée par Γ pour toutes les paires.

Formellement, le modèle d'analyse de sensibilité pose :

$$\frac{1}{\Gamma} \leq \frac{\pi_{s1}/(1-\pi_{s1})}{\pi_{s2}/(1-\pi_{s2})} \leq \Gamma \tag{1}$$

Par exemple, une entreprise fondée par un enseignant-chercheur serait quatre fois plus susceptible de recourir au dispositif JEI ($\Gamma = 4$).

Analyse de sensibilité

Comme les entreprises sont appariées, l'équation (1) implique :

$$\frac{1}{1+\Gamma} \leq \mathrm{P}(\mathrm{Z}_{s1}=1, \mathrm{Z}_{s2}=0) \leq \frac{\Gamma}{1+\Gamma}$$

Avec $\Gamma = 4$ on pourrait avoir, par exemple, au sein d'une paire :

- 80 % ($4/(4+1)$) de chances que l'entreprise 1 soit la traitée,
- et donc seulement 20 % pour que ce soit l'entreprise 2,
- ou l'inverse,...
- alors qu'elles sont « similaires » sur plusieurs caractéristiques !

On peut aussi avoir des écarts moins importants dans une autre paire : 60 % pour l'entreprise 1 et 30 % pour l'entreprise 2. **Mais jamais plus grands !**

Analyse de sensibilité

Dans le cadre du modèle d'analyse de sensibilité,

$$\frac{1}{1+\Gamma} \leq \mathrm{P}\big(\mathrm{Z}_{s1}=1, \mathrm{Z}_{s2}=0\big) \leq \frac{\Gamma}{1+\Gamma}$$

- **Statistique de test T** (Rosenbaum, 2007) : distribution cette fois inconnue, mais qui peut être encadrée par celles de deux statistiques de test qui correspondent aux cas *aux « bornes »* définies par Γ.

$$\mathrm{P}\big(\mathrm{T}^{\min} \geq k\big) \leq \mathrm{P}\big(\mathrm{T} \geq k\big) \leq \mathrm{P}\big(\mathrm{T}^{\max} \geq k\big), \forall k$$

- T^{min} « s'apparente » à la somme de S variables aléatoires indépendantes qui prend les valeurs $(\mathrm{R}_{s1} - \mathrm{R}_{s2})$ avec la probabilité $\frac{1}{(1+\Gamma)}$, et $-(\mathrm{R}_{s1} - \mathrm{R}_{s2})$ avec la probabilité $\frac{\Gamma}{(1+\Gamma)}$ si $\mathrm{R}_{s1} \neq \mathrm{R}_{s2}$,

- T^{max} définie en intervertissant les rôles joués par $\frac{1}{(1+\Gamma)}$ et $\frac{\Gamma}{(1+\Gamma)}$.

Résumé de la démarche

Au sein de chaque paire, on considère que l'une des deux entreprises (*pas nécessairement celle qui est effectivement traitée*) a, au plus, Γ plus de chances d'être bénéficiaire du dispositif JEI.

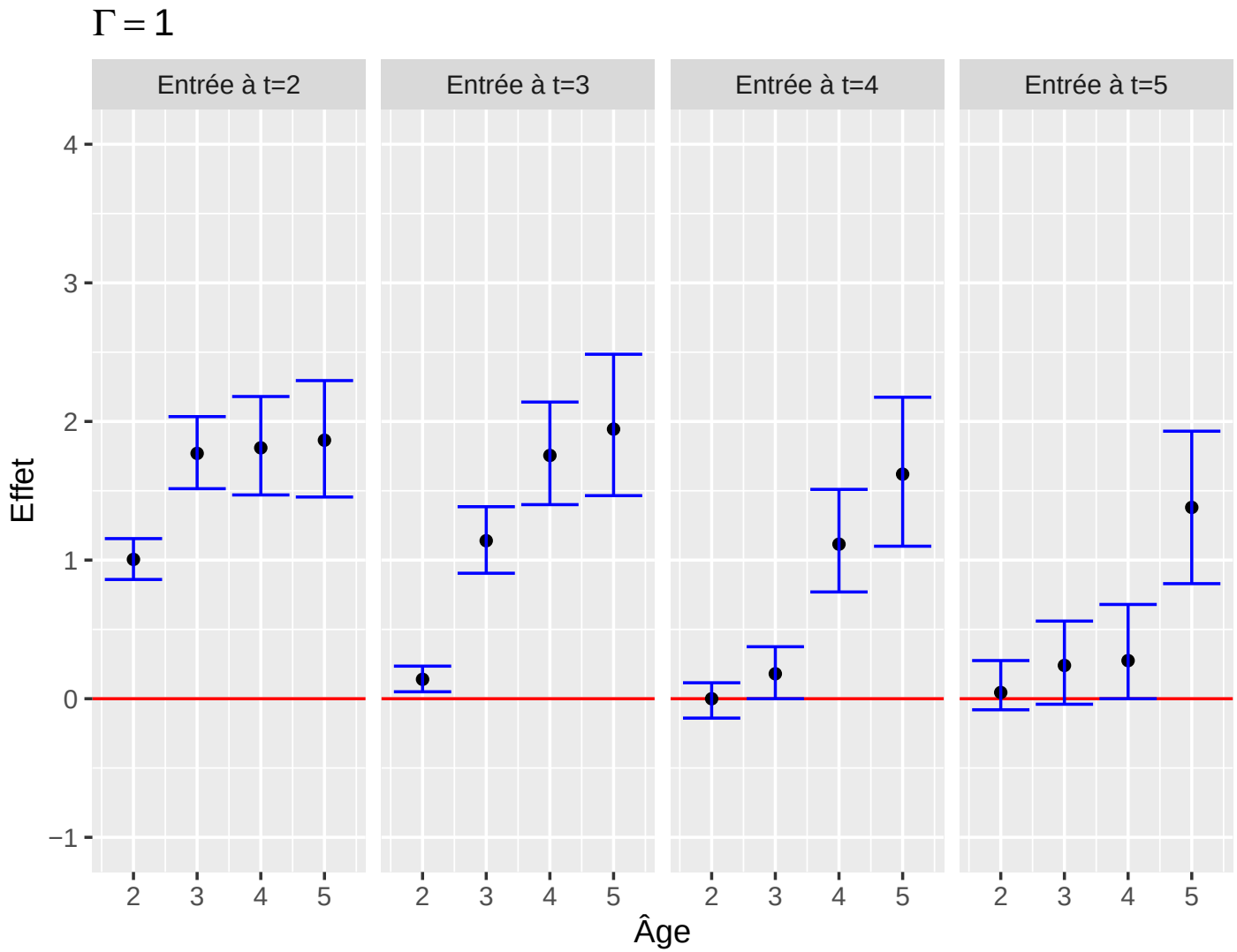
En faisant varier Γ, on peut discuter de la sensibilité des résultats à la présence d'une caractéristique inobservée (relâchement de l'hypothèse d'indépendance conditionnelle).

Par exemple, déterminer la valeur maximale de Γ, c'est-à-dire l'ampleur du biais de sélection qui doit rester après appariement pour rejeter l'hypothèse d'un effet du dispositif.

Enfin, pour $\Gamma > 1$, on peut obtenir un intervalle des valeurs possibles pour l'effet estimé et l'intervalle de confiance associé : estimateurs d'Hodges-Lehmann ($\tau_{si} = \tau_0$), extension aux *effets attribuables aux traitements* (τ_{si} quelconque).

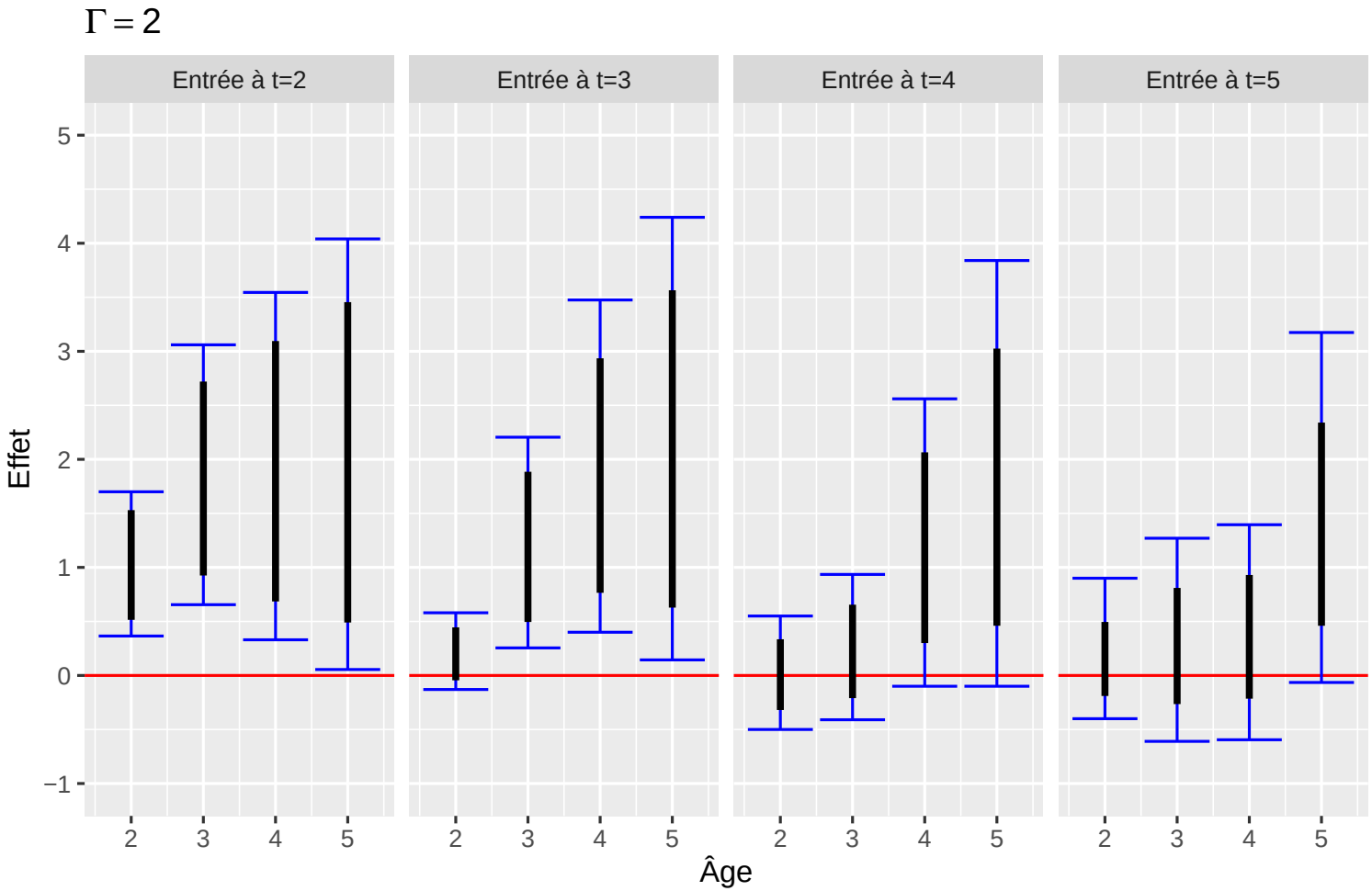
Test d'un effet additif constant : cas « naïf » idéal

Effet du dispositif JEI sur l'emploi salarié ETP en fonction de l'âge à l'entrée



Test d'un effet additif constant : Analyse de sensibilité

Effet du dispositif JEI sur l'emploi salarié ETP en fonction de l'âge à l'entrée



Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Roiniczych*, Tom X :1–51. Réimprimé en anglais dans *Statistical Science*, 1990, 5, 463-480.

Rosenbaum, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63 :456–464.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology*, 66(5) :688–701.