MASTER, ECO SUP

EVENS SALIES

ÉVALUATION des POLITIQUES PUBLIQUES

Méthodes statistiques

Cours et exercices corrigés

- Cours complet
- Évaluation des Politiques Publiques et des Programmes
- Applications avec R et Stata
- Réplications
- 50 exercices corrigés

Table des matières

1. Introduction : plan de cours

L'économètre pose des questions causales auxquelles il répond généralement en confrontant des modèles sophistiqués à des données observationnelles. Il existe malheureusement des biais d'estimation inhérents à cette approche. Par exemple,

- Un effet de sélection dans l'étude de programmes pilotes d'économie d'énergie auxquels participent surtout des ménages motivés.
- Un effet d'aubaine dans les études sur le soutien d'entreprises qui n'avaient pas besoin d'être aidées.
- Un biais de variable omise dans l'étude du « rendement » de l'éducation sans pouvoir contrôler des aptitudes non-observables des diplômé·e·s.
- Les études randomisées-contrôlées ne sont pas non-plus exemptes de biais (attrition, effet Hawthorne, validité externe faible, biais de stratification, etc.).

Ce cours est une formation de 30h aux **méthodes statistiques d'évaluation MSE**, des méthodes qui sont conçues au départ pour mesurer l'effet d'interventions, dans un cadre randomisé, mais qui sont aussi efficaces dans le cadre observationnel.

Les MSE sont aussi des méthodes de **contrôle** de l'influence de variables qui ne nous intéressent pas prioritairement, qui est un problème dans toute science !

Ce chapitre d'introduction a six sections assez courtes. Nous clarifierons d'abord le titre. Nous verrons que l'évaluation est inscrite dans la loi et qu'il existe un charte (section 1). Qui pourrait être intéressé par ce cours sera vu dans la section 2, ainsi que les compétences à l'issue de la formation (section 3) et les débouchés que vous pouvez espérer (section 4). Le plan sera ensuite présenté (section 5). La dernière section contient quelques rappels très simples de concepts indispensables avant de commencer ce cours (section 6).

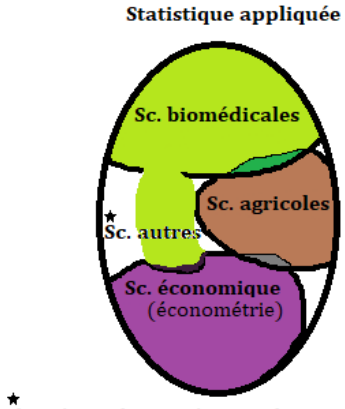
1.1. Objet du cours « Méthodes Statistiques d'Évaluation »

On entend beaucoup parler d'évaluation d'impact en économie, d'où la nécessité d'un cours de ce type ! Au passage, attention sur le terme « impact », qui correspond à des études plus poussées, qu'estimer un effet et faire des tests de robustesse.

1.1.1. Clarifier le titre

Le cours aurait pu s'appeler différemment, avec « **Économétriques** » à la place de « **Statistiques** ». J'aurais pu préciser « **Évaluations des politiques publiques** ». Le titre aurait pu contenir les mots « causal », « traitement », etc.

Pourquoi alors « **Méthodes Statistiques** » et ne pas avoir mis « **Méthodes Économétriques** » ? Pour ne pas retenir des MSE qui ne seraient utilisées qu'en économie. « Statistiques » est plus général ! Par analogie avec ce qu'écrivent Efron et Hastie (2016), l'économétrie a les faits économiques pour juger de la pertinence de ses idées. L'autre raison est liée à l'utilisation croissante de l'**apprentissage machine**, la **science des données**, qui relève plus de la Statistique.



★cliométrie, chimiométrie, anthropométrie,

L'article d'Athey et Imbens (2019), « Machine learning methods that economists should know about », va dans le sens d'une ouverture de l'économétrie à l'apprentissage machine, donc non-seulement aux algorithmes, mais aussi à l'inférence statistique.

Le cours va donc plus parler de **protocole** (statistique) que de spécification *ad hoc* (économétrie). L'économétrie devrait servir à estimer les paramètres structurels d'une ou plusieurs équations transposant une théorie économique, avec pour objectif de tester cette théorie contre une autre (O1) ou de mesurer l'effet d'une politique publique (O2). La séquence est : spécification des équations, estimation des paramètres et tests sur ces paramètres (**SET**). Le système d'équations est généralement $\boldsymbol{Y} = f(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\varepsilon})$, avec $\boldsymbol{\beta}$ les paramètres structurels à estimer, $\boldsymbol{X}$ pouvant contenir des variables de Y.

Or, la théorie n'est pas toujours claire. En a-t-on besoin pour mesurer l'effet :

- du passage à l'euro sur le PIB de la France ? Voir Gasparotti et Kullas (2019).
- d'une réforme des programmes de l'éducation nationale ?
- du service militaire sur la carrière ?
- de se syndiquer sur le salaire ?

[Réf. du papier « 20 years of the EURO: winners and loosers », Réf. supplémentaire]

Les MSE s'intéressent plutôt à O2 qui a besoin de protocoles solides. Le protocole dit *idéal*, la **randomisation**, précède la séquence SET ; Holland (1986) mettait en garde, dans sa maxime, « *no causation without manipulation* ».

C'est en train de changer : des protocoles en économétrie précèdent ou sont suivis de spécifications plus économiques que dans Angrist et Pischke (2009). Esther Duflo notamment, essaie de poursuivre O1+O2.

Sur la base de mots clés tels que « expérience aléatoire », « appariement », Bono *et alii* (2018) observent une croissance depuis 20 années.

- J'appelle la séquence protocole, spécification, estimation, test : **PSET**.

Inférence { Spécification d'un modèle statistique qui formalise cette question / Estimation de ce modèle / Tests d'hypothèses dans le modèle afin de répondre à la question de départ }

Protocole (« design ») conçu pour répondre à une question causale (cf. infra)

A propos d'« **Évaluation** », il s'agit dans ce cours d'évaluation (de l'effet) d'une intervention menée par un gouvernement, une collectivité, une entreprise.

On va surtout parler d'évaluation de politiques publiques/programmes (**EPP**) : Revenu de Solidarité Active (RSA), crédit d'impôt recherche (CIR), Programme scolaire, etc. Le terme « Évaluation » n'a pas exactement la même signification selon le niveau d'agrégation des données et la discipline économique :

- microéconométrie (données en coupe, panels, de ménages, entreprises) : *evaluating a job training program*, ou simplement *program evaluation*, ou plus compliqué, *alternative approaches to evaluation in empirical microeconomics*.
- macroéconométrie (panels de pays) : *econometric policy evaluation*.
- microéconomie (théorique, empirique) : évaluation de l'impact social d'un changement de prix (sous les courbes d'offre et de demande), avec calibrage des élasticités. Peut être proche de la microéconométrie, de la simulation

Ce cours traite surtout d'évaluation microéconométrique.

Une évaluation c'est étudier l'effet d'une intervention sur des décisions d'agents économiques plus ou moins agrégés

« Effet » est un terme un peu trop général pour apprécier le résultat d'une intervention. On distingue l'**efficacité** de l'**efficience**. D'autres critères sont la **pertinence** et la **cohérence**. Un dernier critère, est la temporalité (Morel-à-l'Huissier et Petit, 2018) :

L'évaluation peut être effectuée

ex ante (avant l'introduction d'une loi) = *a priori*

ex post (après) = *a posteriori*

in itinere (avant ou après les réformes de cette loi) = concomitante

L'efficience, c'est l'**analyse coût-bénéfice** d'interventions alternatives (voir la présentation sur le site de l'Institut des politiques publiques, IPP).

Nous éviterons de parler d'**effet causal**. Angrist, Imbens, Rubin, etc. parlent de *causal effect*, Wooldridge (2003) de *treatment effect* (**effet d'un traitement**). La **variable causale** du statisticien est la variable exogène (idéalement) pour l'économètre.

Pour chaque unité d'échantillonnage, c'est une variable qui prend au moins deux valeurs, selon que l'unité d'échantillonnage est exposée ou pas à la PP. L'exposition est déterminée par un **mécanisme d'affectation des traitements**.

Selon Angrist et Pischke (2009, 113-115), l'économètre chercherait plus à trouver des relations causales que le statisticien. La **randomisation contrôlée** est un moyen, mais elle est rare en économie ; les données sont souvent **observationnelles**.

Ce qui permet de s'approcher de la causalité c'est :

- la spécification d'un modèle structurel ;
- une méthode **quasi-expérimentale**, de sorte qu'un traitement non randomisé peut être, sous certaines hypothèses, *as if randomized*. Angrist et Pischke (2009) se réfèrent aux estimateurs à variables instrumentales (VI), quand les instruments sont objectivement exogènes !

L'économétrie théorique n'est pas à la traîne sur la recherche en inférence causale. Il y a le travail de Granger, Sims, Engle, Hendry, etc. (on teste l'H_0 de non-causalité). Ce cours adopte la méthodologie de Neyman-Rubin, le **modèle causal de Rubin** (MCR). Nous esquisserons d'autres approches dans le chapitre 2, notamment celle de Pearl.

1.1.2. Pourquoi évaluer ?

De manière générale, l'évaluation répond à une demande d'information des citoyens sur l'utilité de l'action publique, le coût de cette action (impôts).

1.1.2.1. Exemples concrets d'actions publiques (CICE, CIR, et PDV)

CICE et CIR : le CICE coûtait 20 milliards d'€/an, le CIR est passé de quelques centaines de millions à plus de sept milliards entre 2003 et 2021 (Bunel et Sicsic, 2024).

Programme de retour au travail : exemple d'évaluation d'une PP, celle du Conseil Départemental de Seine-Saint-Denis (Conseil-Départemental, 2016), qui mena un Programme d'insertion sur la période 2013-2015, poursuivi en 2015 pour 98400 foyers bénéficiaires du revenu de solidarité active (RSA) (chiffres de fin 2014). En août 2020, une personne seule sans enfant ne touchait que 564,78 €.[1]

[1]. Emilio Mestet, « L'Etat va devoir passer à la caisse », l'Humanité, 19 août 2020, p. 4.

Question : pourquoi est-il important de faire cette évaluation ?

Le RSA est versé par les départements depuis 2004 (l'État donne une compensation financière) ; dépense d'environ 11 milliards d'€ en France en 2018, pour 2,25 millions de bénéficiaires (dont 4,37 % dans le 93, en Seine-Saint-Denis).[2]

À l'échelle du 93, les dépenses de fonctionnement du programme d'insertion sont importantes : 448 millions d'€ en 2014 (425 millions d'allocations RSA + 23 millions d'€ de dispositifs d'insertion).

L'évaluation porte précisément sur le dispositif des projets de ville (PDV), dans le cadre de la loi n° 2008-1249 du 01/12/2008, de réforme des politiques d'insertion. Ces projets accompagnent une partie des bénéficiaires (4153 suivies, dont 170 plus finement, sur 22000 bénéficiaires en 2013). Un questionnaire aux bénéficiaires permit de recueillir des données socioprofessionnelles afin d'attribuer un des trois types de « parcours » :

1 Pole emploi (accompagnement du parcours professionnel)
3 Service Social départemental (associations en soutien)
2 PDV (accompagnement poussé : 1 + 3) + chargée d'insertion, psychologue
On veut mesurer l'impact de 2 relativement à 1, et relativement à 3.

L'évaluation doit aider le politique à trouver le « meilleurs » dispositif (idéalement, plus d'insertion à moindre coût !). C'est une évaluation d'impact (on mesure aussi la perception des personnes aidées et aidantes).

L'évaluation mobilisa peu de moyens supplémentaires au regard de sa pertinence : chargées d'insertion, psychologues et une étudiante stagiaire en évaluation pour l'enquête de terrain, avec le consentement des personnes accompagnées par les PDV.

1.1.2.2. L'évaluation dans la loi

À l'international (ONU, Banque Mondiale),[3] 2015 : année de l'évaluation ! Au niveau européen (évaluation des Politiques de cohésion, politiques qui figurent dans l'article 174 du Traité de Lisbonne).[4] Il y a des évaluations, comme par exemple celle de Hagen et Mohl (2008). En France, depuis décembre 2012, les gouvernements mènent une démarche d'évaluation de l'ensemble des PP : la modernisation de l'action publique (MAP). La mission d'évaluation par le parlement est dans la constitution depuis 2008 :

> *Le Parlement vote la loi . Il contrôle l'action du Gouvernement. Il évalue les politiques publiques* (Art. 24 de la constitution, loi n° 2008-724)[5]

Le but est d'équilibrer le budget des Etats, de rationaliser des PP dans le temps ⇒ meilleures réformes de cette PP.

Par exemple, l'évaluation *in itinere* du CIR plusieurs fois réformé (surtout en 2008), afin d'être plus efficient et d'éviter des **effets d'aubaine**. Bien que … la réforme de 2008 est en fait un contre-exemple !

France Stratégie (ex-Commissariat au Plan), coordonne des appels à projets (CICE, CIR, etc.).

---

[2]. Les Echos, « RSA : le nombre de foyers bénéficiaires en hausse de 7,4% en 2013 », 31 déc. 2013.
[3] https://www.unevaluation.org/, et https://unhabitat.org/2015-international-year-of-evaluation#:~:text=Year%202015%20was%20officially%20declared,Nations%20Evaluation%20Group%20(UNEG).
[4] https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:12008E174.
[5] https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000019241014/2008-07-25/#LEGIARTI000019241014.

1.1.2.3. Encadrer les E3P au quotidien

Une Charte de l'Évaluation (SFE, 2006), adoptée en 2003 (actualisée en 2006). Les évaluateur·rice·s suivent la Charte pour se mettre en conformité avec la CNIL (ex., le § « Respect des personnes »). Vous pouvez voir un extrait du genre de document que l'on signe avec la CNIL, dans le cas de l'évaluation TICELEC.[6]

Rapport Morel-à-l'Huissier-Petit du Comité d'Evaluation et de Contrôle (CEC) des PP, de l'Assemblée nationale.[7] Rapport à l'usage du citoyen voulant acquérir des connaissances sur l'évaluation de l'action publique. On peut dire que c'est un rapport d'évaluation de l'évaluation des politiques publiques. Le CEC pilote une grande partie des évaluations « transversales », pour lesquelles les MSE que nous allons voire sont mal adaptée. Par ex., en 2023, 1er juin, la Commission d'EPP relative à la mission RES. Il y a 17 autres commissions, permanente – Chacune se réunit en séance publique avec la Commission des finances et le·a ministre de la Mission budgétaire concernée.

Printemps de l'évaluation depuis 2018,[8] permet de voir comment sont menés la plupart des évaluations des parlementaires. Il s'agit d'évaluations « qualitatives » visant à vérifier que les citoyens ciblés par une PP en ont bien bénéficié. Le maintien de certaines politiques publiques inefficientes pourrait expliquer pourquoi des parlementaires des deux chambres ne prennent pas au sérieux des évaluations pourtant bien menées.

1.2 A qui s'adresse ce cours, niveau requis

À l'étudiant en économie, susceptible d'évaluer une **intervention**, dans le cadre d'un projet de recherche, en stage ou pour un mémoire de Master, une Thèse. Il est préférable d'avoir déjà suivi un cours de statistique-probabilité (au moins le niveau L2) et éventuellement un cours d'économétrie.

1.3 Compétences à l'issue de la formation

Les étudiants apprendront à :

- Poser une **question causale**. On peut dire qu'il n'y a pas d'EPP sans question de nature causale ! Nous allons voir des exemples de questions causales dans ce chapitre et les chapitres suivants. Nous verrons comment modéliser un « question causale » dans le chapitre 2. La plupart des exemples suivants du tableau 1.1 sont tirés d'évaluations connues (nous en reproduisons dans ce cours). Les autres sont hypothétiques.
- Répondre à la question causale en suivant la séquence PSET

[6] http://www.evens-salies.com/TICELEC.doc.
[7] https://www.assemblee-nationale.fr/dyn/17/organes/delegations-comites-offices/cec.
[8] https://www.assemblee-nationale.fr/dyn/16/organes/commissions-permanentes/finances/printemps-evaluation/printemps-evaluation-2024/printemps-de-l-evaluation-2024-commission-d-evaluation-des-politiques-publiques-presentation-et-agenda.

Tableau 1.1 : Études et exemples hypothétiques abordés dans ce cours

« Cause »	Variable de traitement expérimentale	« Conséquence »	Variable de résultat (*outcome*)	Unité d'observation i	Source
Instruction	$D_i =$ années d'études d'un salarié i (Bac vs L3)	Apt. professionnelles (productivité)	Salaire	Salarié qui a au moins une Licence	Angrist et Pischke (2009, 52-63, 115-129), Angrist et Krueger (1991)
Hausse du salaire minimum	$D_i = 1$ si i dans l'État où le sal. min. ↑ (0 sinon)	Offre de travail	Nombre de salariés	Restaurant « fast food »	Angrist et Pischke (2009, 227-231); Card et Krueger (1994, 2000)
Un service d'urgence de l'hôpital	$D_i = 1$ si i a été aux urgences (0 sinon)	Etat de santé	Etat de santé (déclaré par i)	Individu	Angrist et Pischke (2009, 12-13)
Conditions d'apprentissage	Effectif de la classe	Apprentissage	Notes aux examens	Elève de maternelle ou primaire	Krueger (1999) ; Angrist et Pischke (2009, 17-22, 28)
Enseignement des mathématiques	$D_i = 1$ si i suit le programme A (0 si B)	Apprentissage	Note à un test en mathématique	Elève de CM1 (« 4th grade »)	Rubin (1977)
Présentation/cadrage (« frame ») d'une politique publique en situation risquée	$D_i = 1$ si « frame 1 » (0 si « frame 2 »)	Attitude en situation risquée	Choix d'un programme	Étudiants	Tversky et Kahneman (1981) ; Kahneman (2011) ; Glimcher et Fehr (2014)
Détérioration d'un lieu de vie	Distance d'une maison i à un incinérateur	« Valeur » d'un lieu de vie	Prix de la maison i	Maison	Wooldridge (2009, 450-454)
Information sur sa consommation d'électricité	$D_i = 2$ si très équipé (1 si équipé, 0 sinon) ; TIC	Demande d'électricité	Consommation d'électricité de i	Ménage	Projet TICELEC, PACA, 2013.
Aide à l'insertion sociale	$D_i = 2$ si parcours PDV (0 ou 1 si autres parcours)	Insertion sociale	Accès à l'emploi	Bénéficiaire du RSA	Conseil-Départemental (2016)
Soutien de la production de connaissances par les entreprises	$D_i = 1$ si i a le CIR (0 sinon)	Production de connaissances	Dépenses de R&D	Entreprise de RD et/ou innovante	Bozio *et alii* (2014)
Restriction commerciale	$D_i = 1$ si *input* i exposé à PAD (0 sinon)	Demande d'*inputs*	Portefeuille d'inputs	Entreprise	Vandenbussche et Viegelahn (2015)

Tableau 1.1 (suite) – Hors sciences sociales

« Cause »	Variable de traitement expérimentale	« Conséquence »	Variable de résultat (outcome)	Unité d'observation i	Source
Deux types de mélanges thé-lait	$D_i = 1$ si le thé est versé avant le lait (0 si le contraire)	Capacité à identifier le type de mélange	Nombre de mélanges bien identifiés	Buveuse de thé de Fisher	Salsburg (2002) ; Imbens et Rubin (2015) ; Rosenbaum (2010) ; Grima (2013)
Goudron de la combustion du tabac	$D_i = 1$ si i est fumeur (0 sinon)	Etat de santé	Décès de maladie cardiaque	Individu	Rosenbaum (2010, 2-5)
Vitamine C	Dose de vitamine C	Cancer avancé	Tumeur	Individu	Rosenbaum (2010, 2-5)
Café	$D_i = 1$si i est buveur de café (0 sinon)	Cancer	Tumeur	Individu	Benkimoun (2016)
Traitement antihypertenseur	$D_i = 1$ si i est traité (0 sinon)	Insuffisance rénale	Clairance	Individu	Imbens et Rubin (2015)

Une étude peut être sommairement placée dans un tableau comme celui-ci. Le tableau distingue une « **Conséquence** » et la « **Variable de résultat** », plus facilement observable, mesurable, mais certainement réductrice. Dans le 4e exemple (**page précédente**) où « apprentissage » est un terme complexe, c'est bien l'apprentissage qui est « traité », mais comment le mesurer ? Par une production de savoirs à un examen, que l'on va mesurer par une note.

Idem en pour la distinction entre une « **Cause** » (conditions d'apprentissage) et la « **Variable de traitement expérimentale** » (effectif d'élèves en classe).

Sur cette lancée, on pourrait aussi distinguer « **Unité d'observation** » et unité d'observation expérimentale. Dans le 6e exemple, l'expérimentation de Tversky et Kahneman (1981), la population ciblée est probablement des parlementaires, des citoyens tirés au sort. La population expérimentale est constituée d'étudiant·e·s.

Concernant les **conséquences,** il existe d'autres termes, qui partent du besoin d'un agent économique : traitement (input), conséquence directe (*output*), moins directe (*impact*), indirecte (objectif de long-terme). Par exemple, une entreprise qui souhaite se différencier de ses concurrents (besoin), a recours à une aide (input & processus) pour financer une partie de sa R&D, qui lui sert à embaucher un docteur-ingénieur (output) qui fait de la R&D (outcome). Elle lance un nouveau produit (impact) et, plus-tard, devient leader de son marché (objectif de long-terme).

Les quatre dernières études du tableau 1.1 sont là pour rappeler l'encrage des MSE dans les sciences biomédicales. Rosenbaum (2010, 2-8) cite des études majeures : cancer-vitamine C, cancer-tabac.

Bien qu'étudiant·e·s en économie, intéressez-vous à d'autres disciplines où le concept de protocole expérimental vous semble plus évident.

L'article de l'effet cancérigène du café illustre l'idée de **facteur de confusion** (l'eau chaude $> 65°C$), concept central de ce cours.

Les analogies sont utiles quand on a du mal en économie à bâtir un protocole. Par exemple, comparer les effets de deux médicaments, plutôt que mesurer les effets d'un médicament, sert d'analogie à l'évaluation d'une politique relativement à une autre (par ex., PDV vs France Travail, CIR vs aides directes) que relativement à aucune, car il y souvent plusieurs politiques concomitantes qui visent la même chose et ciblent les mêmes citoyens.

1.4 Débouchés

Chargé·e d'études statistiques, chef·fe de projet, évaluateur·rice pour le compte d'un cabinet privé d'évaluation, d'une collectivité territoriale, d'une administration, d'un organisme public ou privé de recherche, national ou international.

(1) Les gouvernements, les collectivités territoriales (conseil départemental, régional, etc.) ont besoin d'évaluer des interventions sur les sujets suivants :

- Insertion professionnelle, âge de départ à la retraite,
- Logement, Hôpital public vs clinique privée,
- Investissement dans le développement durable, économies d'énergie,
- Aides aux ménages, aides aux entreprises, etc.

Des organismes comme France Stratégie, créé par le décret N° 2013-333 du 22/04/2013 (nom officiel : Commissariat Général à la Stratégie et à la Prospective, ex-Commissariat général au plan). FS, qui fait connaître les EPP des laboratoires universitaires, a besoin d'évaluateurs, ou de personnes qui comprennent les méthodes d'EPP.

Agence Parlementaire de l'Évaluation, peut-être un jour …

Des laboratoires d'économie font beaucoup d'évaluations. Par exemple, l'Institut des Politiques Publiques (IPP), http://www.ipp.eu, développé dans le cadre d'un partenariat entre Paris School of Economics et le CREST (laboratoire d'économétrie de l'INSEE). Les thématiques sont surtout les politiques sociales, les politiques de l'éducation, de la santé, les retraites, le logement et l'aménagement du territoire. L'OFCE, de plus en plus.

(2) Les entreprises, qui ont besoin d'évaluer des « projets »

- Introduire une nouvelle campagne publicitaire, A/B testing
- Améliorer le service après-vente
- *Re-pricing*

Elles font appel à des cabinets de conseil, aux collectivités. Par exemple, le cabinet microeconomix, absorbé par Deloitte Economic Advisory. Il y a des cabinets d'experts en évaluation, ou plus précisément, en évaluation prospective des PP, qui mènent une approche qualitative aussi centrée sur des aspects juridiques pour l'État et les collectivités territoriales.[9] Les thèmes sont : cohésion sociale et développement durable des territoires, développement de l'innovation, gouvernance des politiques et programmes publics (les CPER, par exemple). Les cabinets sont Itinere Conseil (http://www.itinere-conseil.com), Quadrant Conseil (https://www.quadrant-conseil.fr), Epices (https://www.epices-net.fr), G.A.C. Group (https://group-gac.com/). Le cabinet Veneficus (https://veneficus.nl) aux Pays-Bas, qui a déjà utilisé la méthode de la double différence.

(3) La Société Française de l'Evaluation liste des offres de stages et d'emplois : https://www.sfe-asso.fr/les-offres-demploi-et-de-stage/.

[9] Feu Arnaud de Champris du Cabinet E.C.s. www.cabinet-ecs.org., fermé depuis 2017.

1.5 Déroulement de la formation [et annonce du plan]

Il y a 30h de cours (20h de magistral, 10h de TD + TP + suivi des étudiant·e·s). C'est un cours appliqué (des données, des estimations, etc.) et méthodologique, qui fait intervenir des concepts présents depuis une trentaine d'années dans la littérature (résultats potentiels, effet du traitement sur les traités, les non-traités, etc.).

Nous programmerons avec STATA et R des estimateurs pour des exemples concrets d'évaluations. Nous n'utiliserons pratiquement pas Excel. Peut-être du Python, etc.

Le cours est sur Moodle et sur mon GitHub.

1.5.1 L'évaluation … des étudiants

Trois travaux possibles (seul·e ou à deux) :

- Résumer un des articles de la Liste suivante, en quelques pages (choisissez un article court). Servez-vous du cours.

- Prendre n'importe quelle base de données qui vous intéresse et appliquez une des méthodes du cours sur cette base. Présentez votre travail sur un document de quelques pages. Exemple.

Quel que soit votre choix parmi ces deux options, il faudrait me dire quels sont les « conséquences » étudiées ? La « cause » examinée ? Le traitement expérimental ? Les résultats potentiels ? Les variables de confusion possibles, observées ou pas ? Quel protocole vous semble le plus judicieux (EXpérience idéale) ? La randomisation éviterait quels biais ? Serait-elle éthique ? Ne rendrait-elle pas l'expérimentation artificielle ?

- Prendre un des programmes fait avec STATA et le refaire R en ou vice versa. Peut-être en Python ? Il peut s'agir d'une commande. Comparer alors les options et les avantages/inconvénients de chaque langage.

1.5.2 Conseil de lecture, formations

Le cours inclut une bibliographie à la fin de chaque chapitre. Sentez-vous libres d'aller consulter l'un des documents référencés. C'est un bon moyen pour prendre de l'assurance.

Une courte introduction littéraire à l'EPP existe dans CAE (2013), et également dans les chapitres 2 et 3 de Wasmer (2010). Les autres ouvrages que je peux vous conseiller sont en anglais et sont plutôt épais. Il y a le manuel de Angrist et Pischke (2014) très accessible, qui comporte des rappels des concepts de statistiques nécessaires.

Le manuel d'Angrist et Pischke (2009) est plus difficile. Malheureusement, il ne comporte pas d'exercices d'entraînement. Enfin, il y a le chapitre 21 de Wooldridge (2010), qui introduit des notations un peu différentes de la version de 2003 de l'ouvrage.

Dans le cadre d'une formation professionnelle à destination de salariés du secteur public, il y a le Master « Evaluation des politiques publiques : recherche et mesure de la performance » de Science Po.[10]

Il y a aussi un paquet de sites :

- Andrew Heiss : https://www.andrewheiss.com/, que m'a conseillé James Kennedy (M2 de l'an dernier)
- …

[10] https://www.sciencespo.fr/executive-education/evaluation-des-politiques-publiques-recherche-et-mesure-de-la-performance

1.5.3 Les chapitres du cours

Le **chapitre 2** débute avec le concept de question causale et explique pourquoi corrélation n'est pas causalité. Nous classerons différents protocoles (expérimentation de laboratoire, naturelles, etc.) permettant de répondre à la question causale, qui diffèrent par la capacité de l'évaluateur à contrôler les facteurs de confusion. Ces facteurs peuvent être à l'origine de biais, dont le « paradoxe de Simpson ». Nous verrons l'approche de la causalité de ce cours, le modèle causal de Rubin (MCR) avec ses concepts de résultats potentiels, observés, d'effet causal individuel, et le problème fondamental de l'évaluation, dont le contournement repose sur une hypothèse de « stabilité » (SUTVA).

Cette approche de la causalité permet d'établir des tests, comme nous le verrons dans le **chapitre 3**, consacré aux mécanismes d'affectation des individus dans les groupes de traitements, aux protocoles dans le cas parfaitement contrôlé et randomisé (*randomized control trial*). Nous présenterons aussi le mécanisme pleinement aléatoire dans lequel on choisit le nombre d'individus tests et témoins plutôt que de laisser faire le hasard. Nous verrons le test exact de Fisher, pour de petits échantillons, son équivalence avec le test de Wald (estimateur de Neyman) et l'ANOVA.

Les expérimentations dans lesquelles le statisticien a peu (ou pas) de contrôle sur le MAT, ce qui est le cas dans la plupart des évaluations en économie, font partie des études observationnelles (EO). Nous introduirons les EO dans le **chapitre 4**. Tout comme les expérimentations randomisées, les EO peuvent être décrites dans le MCR. Nous listerons des avantages (moins onéreuses, plus éthiques, etc.) et des inconvénients (cas extrême où tous les individus sont traités, biais de sélection, non-recouvrement, attrition, etc.). Nous introduirons le concept de biais de sélection (BS) dans le cadre du MCR.

Le **chapitre 5** sera consacré à la stratification, une méthode de contrôle de facteurs de confusion observables (biais manifeste), dans le cas d'expérimentations contrôlées ou pas, à condition que ces facteurs soient peu nombreux. Nous montrerons comment la stratification produit un estimateur sans biais équivalent au cas randomisé. Ce chapitre sera aussi l'occasion d'introduire le concept d'effets moyens sur les traités ou sur les non-traités.

Le **chapitre 6** présente les méthodes d'appariement exact et inexact pour le contrôle de l'influence de facteurs de confusion observables en grand nombre. Nous verrons l'importance de la supposition d'indépendance conditionnelle (CIA) pour ces méthodes et celle du score de propension estimé avec un modèle logit. Nous verrons également une méthode d'appariement par pondération inspirée de l'estimateur d'Horvitz-Thompson. Ce chapitre porte aussi sur l'estimation de l'effet de traitements continus, ou doses (le traitement prend plus que deux états), qui repose sur le score de propension dit généralisé.

Le lien entre appariement et régression sera étudié dans le **chapitre 7**. La régression est un autre moyen d'estimer les résultats potentiels contrefactuel, mais de manière non-consistante. Nous verrons ensuite le lien entre CIA et l'hypothèse d'exogénéité en moyenne $E(\boldsymbol{\epsilon}|\boldsymbol{X})$ de l'économétrie, lorsque la population est scindée en deux groupes de traitements. La régression peut s'interpréter comme un switching regression model à la Heckman. Après avoir introduit la méthode des fonctions de contrôle du biais de sélection, nous présenterons la méthode de régression en discontinuité lorsque CIA ne tient pas.

Les chapitres restants seront consacrés aux cas où certains facteurs de confusion ne sont pas observables, de sorte que CIA n'est plus très utile. Nous verrons la spécification en double et triple-différences dans le **chapitre 8**, utilisée pour sa simplicité dans les cas d'un panel ou pseudo-panel. Nous verrons aussi la méthode du contrôle synthétique, qui généralise l'approche en différences. Nous n'aurons pas le temps d'étudier la méthode des études d'événements, très utile pour le cas de dates variables d'exposition à un traitement, avec *staggered adoption* (les individus qui sont traités, le restent).

Le **chapitre 9** sera consacré aux estimateurs à variables instrumentales. Après un bref rappel de ce type d'estimateurs pour résoudre des biais classiques (variable omise, simultanéité et erreur de mesure), nous verrons l'estimateur LATE (*local averate treatment effect*) dans le MCR. Nous conclurons ce chapitre par une présentation de modèle de Heckman de 1979, pour estimer un effet causal dans les situations extrême où l'on n'observe la variable de résultat que pour les individus tests.

1.6 Exercices sur le chapitre 1

1) À l'oral :

a) Poser une question causale et proposer, par analogie avec le tableau 1.1, le traitement, la variable de traitement expérimentale, ce qui est traité, le résultat et l'individu.

b) Quelle évaluation vous semblerait-il utile de mener aujourd'hui ? Quel protocole envisageriez-vous ?

c) La variable sur laquelle on mesure l'effet d'une PP (la variable de résultat) n'est pas toujours celle qu'on aimerait mesurer. Pouvez-vous illustrer cette affirmation avec un exemple de PP dans le domaine de l'éducation, du travail ou autre ?

2) Les exercices suivants sont des tests de connaissances de base pour bien suivre le cours. Ces exercices portent sur des méthodes statistiques que vous avez vues en L1.

a) **Programme d'aide de retour à l'emploi de personnes sans travail**. Le programme vient de s'achever. On note les deux événements suivants : « ne pas retrouver de travail » dans les six mois, « retrouver du travail » dans les six mois. Y est une variable aléatoire de Bernoulli ; c'est une fonction de ces évènements vers $\{0,1\}$; $Y(\text{"ne pas retrouver du travail"}) = 0$, $Y(\text{"retrouver du travail"}) = 1$. On tire un échantillon aléatoire de n individus ayant suivi le programme ; on suppose que les variables $\{Y_1, \dots, Y_n\}$ suivent toutes la même loi que Y : $Y_i \sim i.i.d.$, $\Pr(Y_i = 1) \equiv p$.

- Quelle est l'espérance de Y_i ?, sa variance ?
- La statistique qui nous intéresse est $\frac{1}{n}\sum Y_i$, que l'on note $\bar{Y}$. Que mesure-t-elle ?
- Quelle est l'espérance de $\bar{Y}$?, sa variance $V(\bar{Y})$? Comment appelle-t-on $\sqrt{\bar{Y}}$?
- Quelle est la loi exacte suivie par $\sum Y_i$? Ecrire cette loi, $\Pr(\sum Y_i = k)$.
- Quel est la valeur à 10^{-3} près de $\Pr(\sum Y_i = 100)$, si $n = 100$ et $p = 0{,}99$.
- Quelle est la loi asymptotique de $Z \equiv \frac{\bar{Y} - E(\bar{Y})}{\sqrt{V(\bar{Y})}}$?
- On souhaite tester l'hypothèse nulle H_0 : le programme n'a pas d'effet ($p = 1 - p = 1/2$). Pour cela, on utilise notre échantillon ($n = 100$) qui révèle que 80% des personnes aidées ont retrouvé un travail. Construire un intervalle au niveau de confiance de 95% pour p basé sur la loi normale. Rejetez-vous H_0 ?

b) **Zéro corrélation vs indépendance**. Soient Y et X deux variables aléatoires indépendantes en distribution. Sont-elles corrélées ?

c) **Espérance conditionnelle**. Dans le tableau des fréquences suivant, nous avons reporté le nombre de jours nécessaires à 100 individus pour retrouver un travail. Le nombre de jours est en colonne, le genre en ligne.

		Y		
		100	110	120
X	« H »	10	10	30
	« F »	20	20	10

- Calculer le nombre moyen de jours pour les hommes (la moyenne conditionnelle $m_{Y|X="H"}$), pour les femmes (la moyenne conditionnelle $m_{Y|X="F"}$), puis le nombre moyen de jours pour l'ensemble de l'échantillon (la moyenne totale m_Y).
- Montrer que $m_{Y|"H"} \times f_{"H"} + m_{Y|"F"} \times f_{"F"} = m_Y$.
- Comment s'appelle cette loi ?

d) **Régression**. On considère une population et le modèle de régression linéaire simple $Y = \beta_1 + \beta_2 X + \epsilon$, avec Y, X et ϵ des variables aléatoires.
- Rappelez la formule des Moindres Carrés (MC) pour $\boldsymbol{\beta_2^{MC}}$ et $\boldsymbol{\beta_1^{MC}}$.
- La variable $\boldsymbol{X}$ est Bernoulli (ne prend que les deux valeurs, 0 et 1) et l'on note $\mathbf{Pr}(\boldsymbol{X} = \mathbf{1}) \equiv \boldsymbol{p}$. Montrer que $\boldsymbol{\beta_2^{MC}} = \boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X} = \mathbf{1}) - \boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X} = \mathbf{0})$ et $\boldsymbol{\beta_2^{MC}} = \boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X} = \mathbf{0})$.
- En posant $\boldsymbol{X}("\boldsymbol{H}")=0$ et $\boldsymbol{X}("\boldsymbol{F}") = \mathbf{1}$, et $\boldsymbol{Y}$ tiré de l'exercice précédent, déduire $\boldsymbol{\beta_2^{MC}}$.

Correction des exercices du chapitre 1

1) Oral.

2)

a) $E(Y_i) = 1 \times p + 0 \times (1-p) = p, V(Y_i) = 1^2 \times p + 0^2 \times (1-p) - p^2 = p(1-p)$.
La fraction d'individus aidés qui retrouve un emploi ; le pourcentage : $100\bar{Y}\%$.
$E(\bar{Y}) = p, V(\bar{Y}) = \frac{p(1-p)}{n}$. L'erreur-type.
Binomiale. $\Pr(\sum Y_i = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$.
0,366.
D'après le Théorème Central Limite : $Z = \sqrt{n}\frac{\bar{Y}-p}{\sqrt{p(1-p)}} \sim N(0,1)$.

$\Pr(|Z| \leq 1{,}96) = 0{,}95$
$\Leftrightarrow \Pr\left(\left|(\bar{Y} - E(\bar{Y}))/\sqrt{V(\bar{Y})}\right| \leq 1{,}96\right) = 0{,}95$
$\Leftrightarrow \Pr\left(\left|(\bar{Y} - p)/\sqrt{p(1-p)/n}\right| \leq 1{,}96\right) = 0{,}95$
$\Leftrightarrow \Pr\left(|(\bar{Y} - p)| \leq 1{,}96\sqrt{p(1-p)/n}\right) = 0{,}95.$
$\bar{Y} - 1{,}96\sqrt{p(1-p)/n} \leq p \leq \bar{Y} + 1{,}96\sqrt{p(1-p)/n}.$
$\Leftrightarrow 0{,}702 \leq p \leq 0{,}898$. Oui, on rejette H_0.

b) Y et X sont indépendantes en distribution, donc la distribution jointe $j(y,x) = m(y)m(x)$, où $m(y) \equiv \int j(y,x)dx$ et $m(x) \equiv \int j(y,x)dy$.
$Cov(Y,X) = \iint (y - E(Y))(x - E(X))j(y,x)dydx = \iint (y - E(Y))(x - E(X))m(y)m(x)dydx$.
(**Théorème de Fubini**) $\iint (y - E(Y))(x - E(X))m(y)m(x)dydx = \int (y - E(Y))m(y)dy \int (x - E(X))m(x)dx$. Par conséquent, $Cov(Y,X) = 0 \times 0 = 0$.

c) Tableau des fréquences relatives.

	100	110	120	
Homme	0,1	0,1	0,3	0,5
Femme	0,2	0,2	0,1	0,5
	0,3	0,3	0,4	1

Pour les hommes : $m_{Y|"H"} = 100 \times 0{,}1 \div 0{,}5 + 110 \times 0{,}1 \div 0{,}5 + 120 \times 0{,}3 \div 0{,}5 = 114$.
Pour les femmes : $m_{Y|"F"} = 100 \times 0{,}2 \div 0{,}5 + 110 \times 0{,}2 \div 0{,}5 + 120 \times 0{,}1 \div 0{,}5 = 108$.
Pour la population : $m_Y = 100 \times 0{,}3 + 110 \times 0{,}3 + 120 \times 0{,}4 = 111$.
$f_{"H"} = 0{,}5$, $f_{"F"} = 0{,}5$, $m_{Y|"H"} = 114$, $m_{Y|"F"} = 108$, et $m_Y = 111$.
$0{,}5 \times 114 + 0{,}5 \times 108 = 111$ est bien égal à m_Y.
La **loi des espérances itérées** (LEI).

d) $\beta_2^{MC} = Cov(Y,X)/Var(X)$, $\beta_1^{MC} = E(Y) - E(X)\beta_2^{MC}$.
Par définition, $Cov(Y,X) = E(XY) - E(X)E(Y)$ et $V(X) = E(X^2) - E^2(X)$.
Or, d'après la LEI, $E(Y) = E\big(E(Y|X)\big)$ et $E(XY) = E\big(XE(Y|X)\big)$.
Dans le cas où $X \in \{0;1\}$, avec $\Pr(X = 1) \equiv p$, on a
$E(X) = 1p + 0(1-p) = p$, et $V(X) = p(1-p)$, deux résultats connus quand X Bernoulli. Par ailleurs, $E(Y) = E\big(E(Y|X)\big) = E(Y|1)p + E(Y|0)(1-p)$, et
$E(XY) = E\big(XE(Y|X)\big) = 1E(Y|1)p + 0E(Y|0)(1-p) = E(Y|1)p$. Donc,
$Cov(Y,X) = E(Y|1)p - p[E(Y|1)p + E(Y|0)(1-p)] = E(Y|1)p(1-p) + E(Y|0)p(1-p)$. Par conséquent, $\beta_2^{MC} = E(Y|1) - E(Y|0)$. Dont on peut facilement déduire $\beta_1^{MC} = E(Y|0)$.

2. Méthodologie de l'évaluation

Ce chapitre pose le concept de question causale pour des données en coupe (section 2.1). C'est dans cette section que nous définirons corrélation, causalité et contrôle des facteurs, à partir de l'analyse du « paradoxe » de Simpson et d'un modèle simulé. Le chapitre introduit le modèle causal de Rubin (MCR) et l'approche contrefactuelle en termes de résultats potentiels (section 2.2). Nous évoquerons différents types d'expérimentations (EX) courantes en économie (section 2.3). La section 2.4 comporte quelques exercices et la section 2.5 une note historique sur le MCR.

2.1 Une question causale

Il y a deux types de questions causales pour l'évaluation d'une PP, d'une intervention, etc., chacune renvoyant à un type de données (Granger, 1986). Illustrons ce point à partir d'un exemple hypothétique en économie de l'environnement :

> *Cross section causality* : pourquoi est-ce que le pays *A* pollue moins que le pays *B* ? Ou, pourquoi A est placé à gauche de la distribution de la variable mesurant la pollution (les GES, par exemple) ?
>
> *Temporal causality* : pourquoi est-ce que la pollution de *A* a baissé cette année ? Ou, pourquoi est-ce que les paramètres de la distribution de GES ont changé ?

Le premier type de question causale ne fait pas explicitement référence au temps qui passe. Des données en coupe devraient suffire à y répondre. C'est plutôt ce type de questions que nous avons dans ce cours.

Le MCR permet au départ de formaliser le premier type de questions. On suppose qu'il existe une cause variable et un individu placé dans différents états **contrefactuels** (un par « valeur » de la cause). La cause est manipulable. Par exemple, une régulation environnementale (la cause) est introduite par un gouvernement. La question causale est plus explicite : est-ce que le pays *A* pollue moins avec la régulation environnementale que sans ?

Ce type de causalité peut paraître limité. En effet, dans n'importe quelle EX, du temps s'écoulerait entre la date de traitement et celle de réalisation du résultat. Rubin reconnaît ce point dans la méthodologie qu'il introduit en 1974, mais il suppose un laps de temps suffisamment « petit » pour que l'individu ne change pas entre les dates. Dans le premier exemple du Tableau 1.1, les travailleurs sont interrogés bien après qu'ils ont fini leurs études. Combiné à une spécification en double différences, sous l'hypothèse de tendances communes, que nous verrons dans le chapitre 8, le MCR peut encore s'utiliser.

Le deuxième type de question causale considère explicitement le passage du temps. Les modèles avec séries temporelles (autorégressifs, vectoriels à correction d'erreur, données de panel, sont appropriés). Le concept de **causalité à la Granger** s'appliquerait dans ce cas.

La question est de savoir si on a bien une causalité, ou une corrélation. Pour Holland (1986), la « causalité » à la Granger ne permet pas de répondre à des questions causales, mais de corrélation/**association**. Cependant, les deux concepts de causalité (le MCR et la causalité à la Granger) coïncident sous certaines condition.[11]

[11] Dans le cadre d'un modèle dynamique, Lechner (2011) donne les conditions sous lesquelles la non-causalité à la Granger implique l'absence d'effet causal à la Rubin et *vice versa*.

Encadré 2.1 : Corrélation n'est pas causalité : la table de vérité statistique

C = « causalité », U = « corrélation », v = « vrai », f = « faux ».

Proposition					
1	U	v	C	v	v
2	U	f	C	v	f
3	U	v	C	f	v
4	U	f	C	f	v

La troisième proposition est le fameux aphorisme « corrélation n'est pas causalité » :

« U est v » et « C est f » est v

En effet, d'après les propositions 1 et 3, les seules pour lesquelles on a corrélation (« U est v »), on peut avoir causalité (« C est v ») ou pas (« C est f »). Par conséquent, la corrélation n'est pas une **condition suffisante** pour la causalité.

Cependant la corrélation est une **condition nécessaire** de la causalité ! En effet, d'après la proposition 2, il est logiquement impossible d'avoir de la causalité (« C est v ») sans corrélation (« U est f »).

Deux conséquences logiques, de C vers U : la causalité est une condition suffisante mais pas nécessaire de la corrélation (ce sont toujours les propositions 1 et 3, mais lues de droite à gauche).

On abrège ces propositions en utilisant le signe d'implication « ⇒ ». Les deux propositions équivalentes suivantes sont vraies :

$$\not{U} \Rightarrow \not{C} \Leftrightarrow C \Rightarrow U$$

la corrélation est une condition nécessaire ⇔ la causalité est une condition suffisante

Ces deux propositions sont logiquement connectées. L'une est la **contraposée** de l'autre.

2.1.1 La corrélation n'est pas une condition suffisante de la causalité

Ou, *association does not imply causation*. C'est un **aphorisme**. Dans le langage de la logique, cet aphorisme énonce que la proposition suivante est vraie :

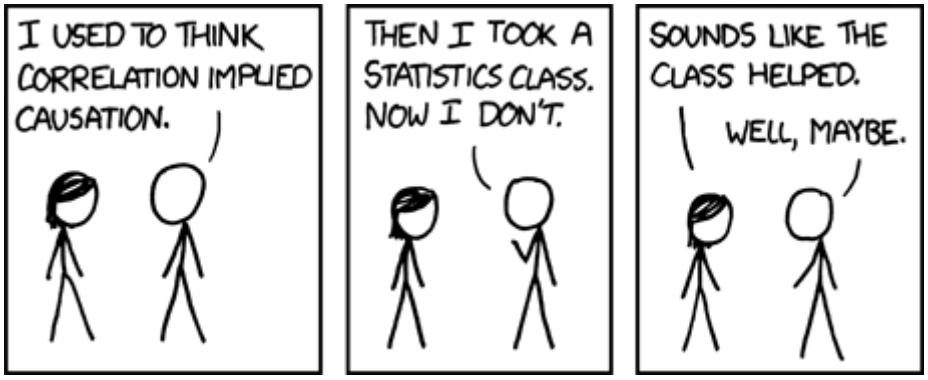
« deux variables sont corrélées » et « aucune n'est causale »

Dans la plupart des études, on dépasse rarement le stade de la corrélation. Prétendre avoir trouvé une relation causale est un peu péremptoire. L'aphorisme précédent est logiquement équivalent à dire que la proposition suivante est fausse :

« une variable cause l'autre » et « elles ne sont pas corrélées »

Autrement dit, causalité ⇒ corrélation. On peut établir une table de vérité (encadré 2.1) afin d'y voir clair. Retenons que puisqu'on ne peut avoir causalité sans corrélation, la corrélation entre deux variables est une condition nécessaire (mais pas suffisante) de l'existence d'une causalité entre elles, et la causalité est une condition suffisante de la corrélation.

Graphique 2.1 : « I used to think... »



Source : http://xkcd.com/552/.
Lecture : l'individu chevelu pense qu'un cours de Stat a éclairé son ami sur la relation corrélation-causalité : l'ami ne pense plus que corrélation implique causalité. Mais il n'est pas sûr que ce soit le simple fait d'avoir assisté au cours qui l'a fait changer d'avis.

2.1.2 Le paradoxe de Yule-Simpson

Parmi les « paradoxes » célèbres en statistique montrant un lien entre corrélation et causalité, le plus connu est très probablement le **paradoxe de Yule-Simpson**. Ce paradoxe est repris à différentes pages du bouquin de Pearl *et alii* (2016).[12] Il est cité dans des revues et bouquins de statistique en science sociale. On le trouve aussi sur la chaîne YouTube de David Louapre,[13] Wikipedia, etc. Enfin, si vous tapez « paradoxe de Simpson magazine tangente », vous trouverez des entrées dans Tangente, le magazine.

Ce paradoxe était à l'origine visible dans des tableaux de contingence (Nelson, 2004, 161) : une corrélation entre deux variables existe dans différentes sous-populations repérées par un facteur (âge, genre, etc.), mais cette corrélation disparaît, voire change de signe, au niveau de la population, autrement dit quand on agrège les sous-populations.

[12] Moins connu, le paradoxe de Lord (Lord, 1967) repris dans ce même ouvrage et Imbens et Rubin (2015).
[13] Lien : https://www.youtube.com/watch?v=vs_Zzf_vL2I.

Nous allons voir trois illustrations du paradoxe de Yule-Simpson. La première vient de Tangente. La seconde, que nous pouvons qualifier d'artificielle (nous l'avons créée pour ce cours), est une adaptation d'un exemple de l'ouvrage de Pearl, Glymour et Jewell (2016), qui nous permettra de formaliser ce paradoxe. Puis, nous verrons une illustration économétrique.

Exemple 1 (« Un clivage démocrates-républicains à nuancer »).[14] En 1964, le président des Etats-Unis, Lyndon Baines Johnson, démocrate, fit voter par la Chambre des représentants le *Civil Rights Act*, qui interdisait la ségrégation raciale en vigueur dans les Etats du sud. Les dirigeants de ces Etats étaient opposés à cette loi, contrairement à ceux du nord. Les résultats des votes furent les suivants :

Tableau 2.1 : Un clivage démocrates-républicains à nuancer

	Démocrates	**Républicains**	
Nord	145/154 soit 94 %	138/162 soit 84 %	90%
Sud	7/94 soit 7 %	0/10 soit 0 %	7%
	61%	80%	

Source : Bibliothèque Tangente, No. 62, p. 65, B. Hauchecorne.

Les démocrates affichèrent partout un pourcentage d'approbation supérieur. Mais au niveau agrégé (Nord + Sud), les anti-ségrégations sont plutôt républicains. Cet exemple donne un avant-gout du casse-tête que représente le « paradoxe » de Yule-Simpson. En notant $Y(\text{"pour"}) = 1$, $Y(\text{"contre"}) = 0$, R « Républicain », D « Démocrate », on a trouvé $E(Y|R) > E(Y|D)$, mais $E(Y|R,N) < E(Y|D,N)$ et $E(Y|R,S) < E(Y|D,S)$.

Exemple 2 (« Les économistes ne sont pas plus au chômage que les juristes »). En 2022-2023, les effectifs d'étudiant·e·s inscrit·e·s en Master de Droit et Economie/AES furent respectivement 77000 et 66500 (Baudry, 2022). Le tableau ci-dessous donne, pour un échantillon hypothétique de taille 700, la proportion qui trouve un travail en Île-de-France à la sortie du M2, selon la zone de résidence.[15]

Tableau 2.2 : Diplômé·e·s trouvant un travail dans les six mois en Île-de-France

	Master 2 Économie		Master 2 Droit	
Paris	$\mathbf{\frac{81}{87}}$	93 %	$\mathbf{\frac{234}{270}}$	87 %
Banlieue	$\mathbf{\frac{192}{263}}$	73 %	$\mathbf{\frac{55}{80}}$	69 %

1) Combien d'étudiants de l'échantillon habitent Paris et la Banlieue ?
2) Quelle discipline a le plus fort taux d'insertion par zone géographique ?
3) Complétez les cases vides de la ligne « Total »
4) Même question qu'au 2), mais au niveau agrégé (Paris + Banlieue)
5) Quel est la valeur de l'effet « Économie » ?

[14] Hauchecorne, B., 2018. Les concepts généraux de la microéconomie. Bibliothèque Tangente n° 62, Mathématiques et économie, p. 85

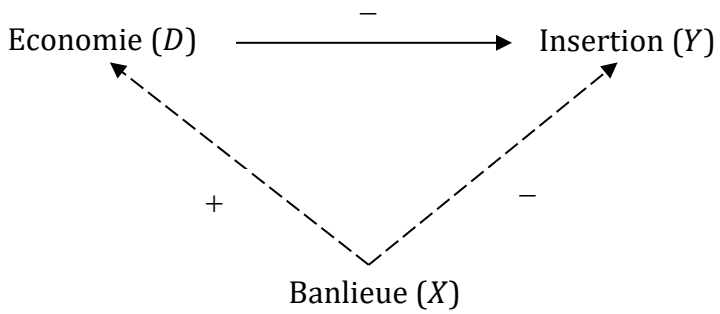
[15]. Le total de 350 dans chaque discipline du tableau 2.2 reflète mal la répartition 53 % vs 47 %.

Explications : le taux d'insertion des étudiant·e·s de banlieue est plus faible dans les deux disciplines ($\mathbf{73 < 93}$, $\mathbf{69 < 87}$). Ces étudiant·e·s peuvent être discriminé·e·s car ils ont plus de transport - surtout pour travailler à Paris -, et à cause de de leur lycée d'origine. Or, ils sont plus nombreux à faire Économie ($\mathbf{263 > 80}$) que les étudiants de Paris qui, eux, ont plutôt fait du Droit ($\mathbf{270 > 87}$). Ainsi, être en banlieue est un facteur commun de faire plutôt Économie, et d'avoir moins de chance de trouver un travail dans chaque discipline. C'est un **facteur de confusion** (on donnera une définition dans la sous-section 2.1.3). Au niveau agrégé, l'Économie, qui comporte surtout des étudiants de banlieue, donne moins de travail à la sortie du M2.

La variable de résultat, l'insertion, est un **collisionneur** (*collider*) entre la variable de traitement (mention du Master) et de confusion (lieu de résidence). On peut représenter cette situation à l'aide d'un **graphe acyclique dirigé** (*directed acyclic graph*, DAG). Le graphe est dirigé car les flèches ont un sens. Il est acyclique car il ne boucle pas. On verra des DAG dans ce cours. Ils sont utiles à l'inférence causale, comme le montre Judea Pearl depuis deux décennies au moins (Pearl, Glymour et Jewell, 2016).

Le DAG montre qu'il y a une corrélation non-souhaitable entre le lieu de résidence et la mention du Master. La corrélation entre le lieu de résidence et l'insertion est intéressante mais n'est pas notre question causale.

Graphique 2.1 : Graphe acyclique dirigé de l'exemple 2



Alors, faut-il agréger ou pas ? Et si oui, de quelle manière ? On a vu comment agréger les résultats par origine géographique en Economie, 93 % et 73 %, pour obtenir 78 %. *Idem* pour le Droit (on agrège 87 % et 69 % pour obtenir 83 %). Dans le cas de l'économie, en calculant $\frac{81+192}{87+263}$, et pour le Droit $\frac{234+55}{270+80}$. Ces deux ratios sont respectivement égaux à :

$$\frac{81}{350}+\frac{192}{350}=\frac{81}{87}\frac{87}{350}+\frac{192}{263}\frac{263}{350}=\ 0{,}93\frac{87}{350}+0{,}73\frac{263}{350},$$

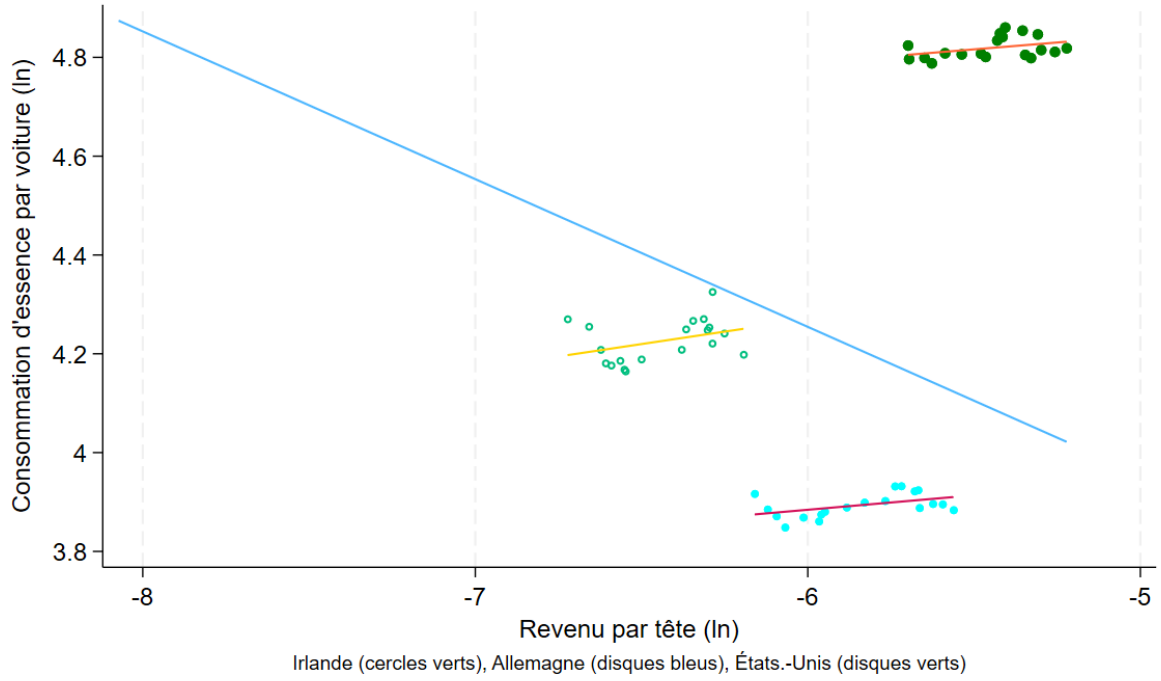
$$\frac{234}{350}+\frac{55}{350}=\frac{234}{270}\frac{270}{350}+\frac{55}{80}\frac{80}{350}=\ 0{,}87\frac{270}{350}+0{,}69\frac{80}{350}.$$

Certes, 73 % est plus petit que 93 % mais a plus de poids $\left(\frac{263}{350} \gg \frac{87}{350}\right)$. Cette 'surpondération' biaise négativement l'effet « Economie ». Il n'y aurait pas ce problème si les effectifs en Economie et en Droit et les lieux de résidence étaient indépendants (par exemple, si au lieu de 87, 263, 270 et 80 nous avions 70, 280, 70 et 280). L'exercice 2.4.2 reprend cette explication formellement, mesure le biais et propose une autre agrégation.

La situation caractérisée par le paradoxe de Simpson-Yule est courante dans les études économétriques. Lee (2016, 25-26) est, à notre connaissance, le seul économètre à y faire explicitement référence. Les économètres parlent plutôt de biais de variable omise. Le facteur de confusion du statisticien est la **variable omise** (**VO**) de l'économètre. Rappelons qu'une VO détermine la variable dépendante et est corrélée avec une variable explicative (endogène, du coup). Elle est observable ou pas, mais pas observée !

Illustrons le paradoxe à partir des données de l'article de Baltagi et Griffin (1983) portant sur les déterminants de la demande d'essence dans un panel de 18 pays de l'OCDE sur la période 1960–1978. Dans l'article, la dépense d'essence par tête est reliée au revenu réel par tête, au prix réel de l'essence et au nombre d'automobiles par tête. On retire les deux dernières variables, qui deviennent des variables omises !

Graphique 2.2 : Relation consommation d'essence-revenu



[charger le programme **`simpsonparadox_exemple3.do`**]

2.1.3 Le contrôle des facteurs

Comment se débarrasser de l'influence d'un facteur X sur l'effet « causal » $D \rightarrow Y$? Il y a-t-il une corrélation indirecte (qui passe par X) entre D et Y ? Wasmer (2010, p. 31) parle de **causalité externe**. Avant de résumer très brièvement plusieurs approches du contrôle d'un facteur tel que X, définissons-le.

Définition 2.1 : un **facteur de confusion** X est une variable corrélée à la fois avec le traitement D (la variable « causale » de premier intérêt) et la variable de résultat (Y).[16] On dit que X confond la relation entre D et Y.

Dans l'exemple 2, où D représente la mention de l'étudiant·e, Y représente l'insertion sur le marché du travail, et X le lieu de résidence, alors X peut confondre la relation entre D et Y, car le lieu de résidence influence à la fois la probabilité de choisir une filière et celle de trouver un travail.

1) Se débarrasser de X

On retire X. En physique, on peut vérifier la loi de la chute des corps dans une cloche (expérimentation menée à l'École polytechnique fédérale de Lausanne), ou sur la lune (mission Apollo XV) ; Galilée avait raison. On a retiré l'air (X).

En médecine, pour évaluer un traitement contre l'hypertension, on peut empêcher l'exposition de patients hospitalisés à certains facteurs. On peut contrôler l'alimentation par ex., en supprimant le sel ajouté (X).

Quand on ne peut pas retirer un facteur, comment fait-on ?

2) Bloquer X, apparier les individus

On évalue pour chaque valeur de X. L'exemple 2 relève de cette approche. On bloque chaque valeur de $X \in \{1(\text{"Banlieue"}); 0(\text{"Paris"})\}$. On bloque $X = 0$, et pour les valeurs du traitement $D \in \{1(\text{"Economie"}); 0(\text{"Droit"})\}$, on évalue en prenant la différence $E(Y|D = 1, X = 0) - E(Y|D = 0, X = 0)$. On bloque $X = 1$, puis on agrège les résultats.

Quand les facteurs sont des caractéristiques individuelles, comme dans cet exemple, contrôler conduit à retenir un sous-ensemble d'individus tests ($D = 1$) et témoins ($D = 0$) **comparables**. Ce sont les individus qui sont appariés, par un blocage de leurs caractéristiques. Par exemple, les Parisiens de 24 ans vivant chez leurs parents, etc.

Les MSE ont commencé à adopter cette approche au début des années 1990 (Imbens et Rubin, 2015), qui est courante en épidémiologie et dans les sciences biomédicales. Elle s'inspire des **expérimentations contrôlées** dans l'industrie (Box *et alii*, 1978) et dans l'agriculture (Mercer et Hall, 1911).

3) Manipulation de D

Dans ces sciences, le contrôle s'effectue souvant par **manipulation** de D. La manipulation résout le problème de régression vers l'infini (je ne me pose plus la question de savoir si X a une cause, etc.). D précède Y (on n'abandonne pas le principe de causalité), et la manipulation de D assure l'indépendance de D et X. Cette indépendance est produite par du hasard contrôlé (Klein, 2018). C'est pour cela que l'on parle d'expérimentation randomisée contrôlée (randomized control trial).

Le nombre de réplications doit être grand.

[16] Lê Nguyên Hoang de la chaîne YouTube Science4All a une vidéo sur les facteurs de confusion : https://www.youtube.com/watch?v=0NbyYOcIwAY.

4) Supposer une forme fonctionnelle en X

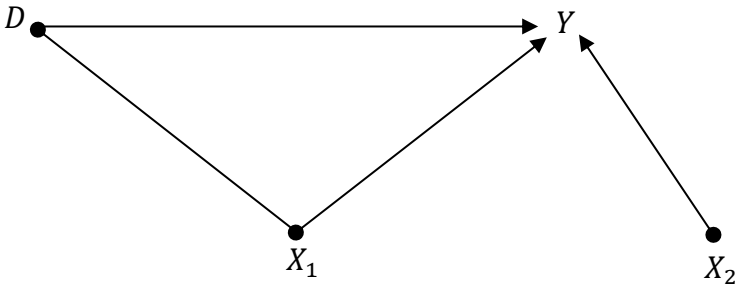
L'exemple 3 de la sous-section 2.1.2 relève de l'approche économétrique classique. On a estimé une équation de demande de carburant, comportant des variables explicatives dans un modèle linéaire. Pour faire le lien avec les écritures précédentes, supposons que la variable explicative de premier intérêt ne soit pas continue (le revenu), mais dichotomique. L'approche suppose que $E(Y|D=d, X=x)$ suit une forme fonctionnelle $f(d,x)$, par exemple $\beta_1 + \beta_2 d + \beta_3 x$. Le coefficient β_2 mesure l'effet du traitement.

Toutes ces approches sont équivalentes, sous des hypothèses. Ces approches peuvent être combinées.

5) Faut-il contrôler les X qui ne confondent pas l'effet causal ?

Dans les études, il y a plusieurs variables de confusion, mais aussi d'autres variables. D'où la question : faut-il tout contrôler tous les X de $\boldsymbol{X}$? Considérons par exemple le schéma suivant. Nous allons voir que ce n'est pas grave de ne pas contrôler une variable qui n'affecte que la variable de résultat.

Graphique 2.3 : Graphe acyclique dirigé (Y, D, X_1, X_2)



[charger le programme **`control_short.do`**]

Enfin, cette sous-section n'aborde pas le cas très particulier de **variables cointégrées**. Dans ce cas, toutes les variables de la relation de cointégration bougent ! On distingue le court-termes (non-causalité de Granger) du long-termes (exogénéité faible).

sous-section en construction

2.2 Le modèle causal de Rubin[17]

Le **modèle causal de Donald B. Rubin** (MCR par la suite), qui tire son nom de l'article de Holland (1986),[18] permet de formaliser une question causale sous la forme algébrique. Il permet également de formaliser le contrôle expérimental des facteurs, de discerner causalité de corrélation et du coup, de résoudre des paradoxes comme celui de Simpson, Lord.

Les différents types d'EX que nous présentons dans la section 2.3 sont compatibles avec le MCR. Par exemple, une randomisation du traitement peut être simplement formalisée dans ce modèle. Enfin, le MCR peut aussi être combiné à un modèle de régression linéaire. Pour Wooldridge (2010), le ***counterfactual framework***, qui définit un effet causal dans le MCR, a été inventé par Rubin (1974), puis adopté par Heckman.[19]

2.2.1 Traitement, résultats potentiel et contrefactuel

On a vu dans le chapitre 1 que, un traitement dans une EX d'économie, c'est par exemple être exposé ou pas (un état) à une PP, une réforme de cette PP, un programme pilote, etc. (cf. les exemples du Tableau 1.1). Ce traitement s'applique à un individu à une période. C'est le **traitement actif**. En face, il y a toujours au moins un autre traitement ou pas de traitement du tout, le *status quo*. Dans ces deux cas, le traitement alternatif est appelé **traitement de contrôle**. Par exemple,

(*i*) « politique économique 1 » face à « politique économique 2 »
(*ii*) « politique économique 1 » face à « pas de politique économique du tout »

Ce traitement, s'il était appliqué à une individu, produirait une valeur d'une variable de résultat, Y, un **résultat potentiel** (RP) de cet individu. Le MCR formalise les deux RP, et pas seulement le résultat qui sera observé et que nous définirons plus loin. Imaginer quel serait le résultat si l'on exposait un individu à un traitement ou à l'autre, c'est faire une expérience de pensée dans laquelle tous les états du traitement et des RP existent. C'est le monde des « what if ? » (Imbens et Rubin, 2015, 7).

$$\text{Un traitement } D_i, \text{ un RP } Y_i(D_i), \text{ pour } i = 1, \dots, N \text{ individus}$$

On peut aussi écrire le RP avec D_i en indice de Y.

$$Y_i(D_i) \stackrel{\text{def}}{=} Y_{D_i}.^{20}$$

Le MCR suppose que le traitement dans la question causale (section 2.1) est une variable **manipulable**, i.e. dont l'état, la valeur, à laquelle l'individu est exposé, puisse varier indépendamment. Deux exemples :

- La question « Est-ce que les chômeurs jeunes décrochent plus facilement un travail que les chômeurs vieux » n'est pas une question causale dans le MCR car l'âge n'est pas manipulable … Ce n'est pas tout à fait vrai ; voir **ci-dessous**.
- Le lieu de résidence n'est pas manipulable.

[17] Il existe une page en anglais bien plus fournie : https://en.wikipedia.org/wiki/Rubin_causal_model.
[18] Voir par exemple Lee (2016, 25-27) sur ce point.
[19] Pour Dunning (2012, 107), il s'agit du modèle des résultats potentiels de Neyman, ou du **modèle de Neyman-Holland-Rubin**. D'après Heckman et quelques autres auteurs, une approche similaire à celle de Rubin a été développée indépendamment par Quandt (1972) ; il s'agirait alors plutôt du **modèle de Neyman-Quandt-Rubin**, dans l'ordre alphabétique des noms, qui correspond d'ailleurs aussi à l'ordre chronologique des inventions, 1923, 1972 et 1974. Tous reconnaissent la paternité de Neyman.
[20] Cette application est surjective ; voir Holland (1986).

L'âge, le lieu de résidence, … sont ce que Holland (1986) appelle des caractéristiques, des **attributs**, des individus. Des expérimentations trouvent des astuces pour manipuler ces caractéristiques, de sorte qu'elles prennent le statut de variable de traitement.

Dans le premier exemple, supposons que l'on manipule l'âge ou le lieu de résidence écrits dans un CV ; nous pourrions même considérer le genre comme variable de traitement. Voire, construire un CV hypothétique dans lequel nous ne toucherions à rien, sauf à une caractéristique de notre choix, qui serait alors notre traitement.

La condition de traitement expérimental manipulable n'est pas si contraignante !

Non seulement le traitement doit être manipulable, mais le RP contrefactuel doit être **clairement défini**.

Pour comprendre ce point, considérons l'exemple économique suivant : « Avec le prêt garanti par l'État (le traitement), je n'ai pas fait faillite en 2020 ». Dans ce cas, le RP contrefactuel (la survie de mon entreprise si je n'avais pas eu le prêt) est évident : j'aurais fait faillite ou pas.

Prenons un autre exemple, avec :

- Individu i : une entreprise qui fait de la recherche durant les années t et $t+1$.
- Variable traitée expérimentale : l'accroissement des dépenses de R&D que l'on note $Y_{i,t+1}$ pour i et que l'on mesure par les effectifs supplémentaires à la R&D (output), ou le supplément de R&D (outcome), entre t et $t+1$.
- Un traitement pour i qui prend deux valeurs : d (i a une subvention à la R&D en t) et d' (i n'a pas de subvention à la R&D en t).

La question causale est de savoir si avoir recours à une subvention entraîne plus de recherche que ne pas avoir recours. C'est une des questions posées par Marino *et alii* (2016) dans leur étude sur données françaises.

[Voir la feuille **`marinoetalii_2016.pdf`**]

Il y-a un problème de définition du traitement de contrôle dans notre exemple, contrairement aux traitements dans l'article de Marino et *alii*. En effet, le traitement « n'a pas de subvention à la R&D » n'est pas clair. Veut-il dire que i n'a pas du tout d'aide ? L'entreprise peut ne pas avoir de subvention mais des avantages fiscaux (le CIR, par exemple). Pour s'en sortir il faudrait définir, comme dans cet article, les traitements suivants : $d =$ « subvention à la R&D, pas de CIR » et $d' =$ « subvention à la R&D et CIR », $d'' =$ « aucun des deux ».

2.2.2 Résultat observé et l'équation de Rubin

Le plus souvent, on a deux traitements modélisés par une dummy $D_i \in \{0; 1\}$. Précisons qu'il n'y a qu'un D_i pour i, qui est le traitement auquel i est exposé. Nous verrons l'utilité du concept de **traitement potentiel** dans le chapitre 9.

$Y_i(0)$ et $Y_i(1)$ sont deux quantités possibles dans le monde des « what if ? » au sens où si i était exposé à $D_i = 0$ on pourrait observer $Y_i(0)$ (idem pour $D_i = 1$ et $Y_i(1)$). Le **résultat observé** (RO) pour un individu i, Y_i, est le RP qui correspond au traitement qu'il reçoit :

$$Y_i^{obs} = \begin{cases} Y_i(1) & \text{si } D_i = 1 \\ Y_i(0) & \text{si } D_i = 0 \end{cases}$$

Y_i^{obs} est généralement écrit Y_i, qui est ce que Rubin (1974) appelle la vraie (*true*) valeur, mesurée un peu plus tard que le moment du traitement. Un seul traitement

s'applique à la fois. Donc, l'un des deux RP au plus sera observé (Pearl, Glymour et Jewell, 2016). La manière d'écrire cette relation entre RO et RP tient en une équation :

$$Y_i^{obs} = Y_i(1)D_i + Y_i(0)(1 - D_i).$$

Q. C'est l'équation du RO. Ecrire l'équation du résultat manquant, Y^{mis}.

?

Ce RP non-observé s'appelle **RP contrefactuel**. Puisque c'est seulement après l'exposition à un traitement que l'on sait quel RP est contrefactuel, on peut dire qu'il y a deux RP, mais qu'un RP contrefactuel.

2.2.3 Effet causal individuel et problème fondamental de l'évaluation

On compare les RP en définissant un effet causal dans le MCR ainsi :

Définition 2.2 : l'**effet (causal) individuel (ECI)** du traitement actif sur i, en un point du temps (une période t courte bien définie), est :

$$Y_i(1) - Y_i(0)$$

Cette quantité, ECI_i, peut être définie quel que soit le RP qui sera observé (donc quel que soit D_i), car c'est un calcul sur les RP (avant traitement). Elle ne dépend pas non plus du mécanisme suivant lequel l'individu i reçoit tel ou tel traitement.

Nous avons souligné « un point du temps » pour préciser que l'individu i exposé aux deux traitements n'a pas le temps de changer. Dans la réalité, ses caractéristiques changent (Imbens et Rubin, 2015, p. 8). Ajoutons trois points :

- $Y_i(1) - Y_i(0)$ est une constante ou une variable aléatoire selon le contexte.
- D'autres fonctions (*causal estimands*) de $Y_i(1)$ et $Y_i(0)$ peuvent être choisies, comme $Y_i(1)/Y_i(0)$; voir Imbens et Rubin (2015, pp. 18-19).
- Le troisième point et le plus important. Après traitement, si un seul traitement s'applique (1 traitement ↔ 1 RP), alors il est impossible d'observer les valeurs $Y(1)$ et $Y(0)$ simultanément. Il est donc impossible d'observer à la période t l'ECI. Holland (1986) appelle ce problème *the fundamental problem of causal inference*. Dans ce cours nous parlerons de **problème fondamental de l'évaluation**. Avec $2N$ individus, au moins N RP ne seront pas observables sur $2N$ RP possibles ; Imbens et Rubin (2015, p. 14).[21]

Dans le monde des « what if ? » (avant l'exposition au traitement), ECI_i est défini pour chaque individu i à la même période. Dans une expérimentation de pensée, on peut toujours imaginer cela possible (sous-section 2.3.1). En effet,

« [L]e problème fondamental en économie est que l'on peut rarement faire deux fois la même expérience à conditions inchangées sauf dans les rares cas comme le champs de l'économie expérimentale, restreints à des problèmes microéconomiques précis [...] » ; Wasmer (2010).

[21] On a écrit « au moins » car il peut y avoir des RO manquants pour certains individus.

2.2.4 Effet causal moyen

Une PP s'applique rarement à un seul individu. Dans le monde des « what if ?» l'**effet causal moyen** (ECM) est un objet plus intéressant :

$$N^{-1}\sum_i (Y_i(1) - Y_i(0)).$$

Notons $\overline{Y(d)} := N^{-1}\sum_i Y_i(d)$, d égal à 0 ou 1. Nous pouvons réécrire l'ECM comme $\overline{Y(1)} - \overline{Y(0)}$. Cette différence suggère l'utilisation d'un estimateur qui contourner le PFIC puisque je ne peux pas calculer l'ECM. Je pourrais contourner le problème en ne considérant que les $Y_i(1)$ des N_1 individus exposés au traitement actif, et les $Y_i(0)$ des N_0 exposés à l'autre traitement, i.e., calculer ${N_1}^{-1}\sum_{i:D_i=1} Y_i(1) - {N_0}^{-1}\sum_{i:D_i=0} Y_i(0)$. D'après la définition de Y_i^{obs}, cette différence peut être obtenue avec l'**estimateur** suivant :

$${N_1}^{-1}\sum_{i:D_i=1} Y_i - {N_0}^{-1}\sum_{i:D_i=0} Y_i.$$

Définissons $\bar{Y}_d := {N_d}^{-1}\sum_{i:D_i=d} Y_i$, avec $d \in \{0,1\}$. Cet estimateur est appelé **effet moyen du traitement** (EMT) : $\bar{Y}_1 - \bar{Y}_0$. Cet estimateur, une différence de moyennes, est naturel et calculable, mais rarement utilisé seul car susceptible d'être doublement biaisé. (le PFIC porte sur $N_1^{-1}\sum_{i:D_i=1} Y_i(0)$ et $N_0^{-1}\sum_{i:D_i=0} Y_i(1)$).

On lui préfère l'**effet moyen du traitement sur les traités** (EMTT), un estimateur de l'**ECM sur les traités** (ECMT) :

$$\begin{aligned} N_1^{-1}\sum_{i:D_i=1}(Y_i(1) - Y_i(0)) &= N_1^{-1}\sum_{i:D_i=1} Y_i(1) - N_1^{-1}\sum_{i:D_i=1} Y_i(0) \\ &= \bar{Y}_1 - N_1^{-1}\sum_{i:D_i=1} Y_i(0). \end{aligned}$$

Le PFIC ne porte que sur $N_1^{-1}\sum_{i:D_i=1} Y_i(0)$.

L'exemple fictif suivant va nous aider à comprendre ces formules. Supposons que les colonnes $Y(0)$ et $Y(1)$ contiennent les taux de croissance en % des dépenses de R&D de huit entreprises quand celles-ci ont une subvention, avec ou sans Crédit d'impôt recherche (CIR). Supposons que seulement les quatre premières ont choisi le CIR ($D = 1$). On suppose qu'en moyenne, ajouter du CIR à une subvention produit plus de croissance de la R&D. Remplissez les colonnes 4 et 6.

Tableau 2.3 : ECM et différence des moyennes

i	Y(0)	Y(1)	ECI_i	D_i	Y_i
1	14	13		1	
2	0	6		1	
3	1	4		1	
4	2	5		1	
5	3	6		0	
6	1	6		0	
7	10	8		0	
8	9	8		0	
Total	40	56		*	
Moyenne	5	7		*	

La moyenne des ECI, c.-à-d. ECM, vaut 2. On aurait pu faire directement $\overline{Y(1)} - \overline{Y(0)} = 7 - 5$. L'EMT vaut $7 - 5{,}75 = 1{,}25 < 2$. Les entreprises qui ont recours au CIR sont-elles moins performantes ? Ce qui est sûr, c'est que l'EMT est un peu biaisé dans cet exemple. Illustrons ce résultat pour l'EMT sur de vraies données.

[charger le programme **`eci.do`**]

2.2.5 Stabilité des individus (SUTVA)

Dans cette sous-section, nous introduisons une hypothèse très importante du MCR. Elle porte sur les relations entre individus.[22]

En théorie, le mélange d'individus accroît le nombre de RP. Supposons deux individus, i et j. On peut penser que la valeur du traitement D_i que reçoit l'individu i agit sur la réponse de j à la valeur de son propre traitement D_j.

La plupart des études écartent cette situation et font une hypothèse connue sous l'acronyme **SUTVA** (*Stable Unit Treatment Value Assumption*). Une manière plus formelle de l'écrire – dans le cas de deux individus $i = 1,2$ – est la suivante : $Y_1(D_1, D_2) = Y_1(D_1)$. Et, *idem* pour l'individu 2 (il suffit d'intervertir 1 et 2 partout).

Pour illustrer cette situation, on peut donner l'exemple des externalités induites par une PP, dès lors que celle-ci est majoritairement adoptée.

Encadré 2.2 : SUTVA et effet d'équilibre général du crédit d'impôt recherche (CIR)

Le modèle causal de Rubin originel suppose qu'il n'y a pas d'interaction entre les agents. Appliqué à la PP du CIR, cela veut dire que les dépenses de R&D d'une entreprise, que cette dernière ait recours ou pas au CIR, ne dépendent pas des décisions de souscription des autres entreprises.

Le taux de CIR appliqué au volume des dépenses de R&D, qui valait 5 % en 2004, a doublé en 2006 pour atteindre 30 % en 2008. Le nombre de déclarants de dépenses de R&D éligibles au CIR a augmenté de 51 % entre 2004 et 2007 et 38 % ne serait-ce qu'entre 2007 et 2008 (MENESR, 2010), pour atteindre le nombre d'environ 20000 en 2013. La générosité et le nombre de bénéficiaires du CIR se sont suffisamment accrus pour que nous nous posions la question de l'existence d'un effet d'équilibre général du dispositif.

Supposons que cet effet passe par le prix moyen des produits des entreprises qui font de la R&D et bénéficient du CIR, et qu'il varie avec le nombre de ces entreprises. L'approche théorique développée par Heckman *et alii* (1998), pour le secteur de l'éducation, suggère qu'avec le nombre croissant de bénéficiaires, le rapport entre prix – ou marge – de ces derniers et celui des non-bénéficiaires pourrait avoir diminué, rendant le dispositif moins attractif à long-terme (les effets de la concurrence).

[22] SUTVA comme justification de l'équation de Rubin (Rosenbaum, 2010) ; SUTVA violée (Morgan et Winship, 2007).

2.3 Des types d'expérimentations possibles[23]

Toute étude reporte les résultats d'une expérimentation (EX), terme usuel en statistique. Concernant l'EPP, la définition d'une EX fait référence à l'idée d'intervention.

Définition 2.3 : le terme « expérimentation » (EX) décrit toute intervention dont le résultat n'est pas connu à l'avance avec certitude. Cette définition s'inspire de cette de celle de Shadish *et alii* (2002, 12), pour qui une EX est :

« A study in which an intervention is deliberately introduced to observe its effects »

Cette définition de l'expérimentation est plutôt synonyme de « protocole randomisé » (Morgan et Winship, 2007, 6), ou contrôlé dans une certaine mesure par l'évaluateur (les EX mal randomisées, voire pas du tout, sont des EX à part entière).

Nous n'avons pas toujours le choix du type d'expérimentation. Cependant, chaque type est compatible avec le MCR vu dans la section précédente.

Les critères « Traitement », « Traité » et « Individu », que nous avons introduits dans le **chapitre 1**, permettent de décrire sommairement les études.

D'autres critères peuvent être considérés pour décrire le type d'EX d'une EPP, même si ces critères ne permettent pas toujours de bien distinguer les différents types :

- La manière de contrôler les facteurs (par randomisation, etc.)
- L'individu statistique, et plus généralement, la nature des agents économiques : réels ou des étudiants, dans un laboratoire
- Nature du bien : réel ou virtuel (dans un ordinateur)
- Règle du jeu (s'il y en a une)
- L'enjeu

Pris ensembles, ces critères constituent ce que Harrison et List (2004) appellent le contexte de l'expérimentation ; voir aussi List (2009).
Nous pouvons retenir cinq types d'EX selon ces critères :[24]

EX de pensée, EX de Laboratoire, EX Sociale, EX de Terrain et EX Naturelle

2.3.1 Expérimentation de pensée

Nous pouvons affirmer qu'au stade de (ou même avant) la mise en place du protocole, une évaluation commence toujours par une EX de pensée (EXP) !

On peut dire aussi expérimentation mentale. On l'appelle **expérience de pensée** alors qu'en anglais on dit bien *thought experiment*.

L'EXP vient certainement de la physique. Pour Klein (2022), « on pense des mondes contrefactuels pour mieux comprendre le monde factuel. »

Dans une EXP vous pouvez imaginer l'ensemble des RP et les chances qu'ils se produisent. Alors que dans le réel, le RP contrefactuel n'est pas observable (Shadish, Cook et Campbell, 2002), vous pouvez mesurer la réaction d'un individu au traitement auquel

---

[23]. Les types d'EX se sont enrichis en fonction des avancées de la recherche en statistique (économétrie, psychométrie, sciences biomédicales, etc.). Par exemple, on peut dater la randomisation aux travaux de Fisher dans les années 1900 (Salsburg, 2002).

[24]. L'Institut des Politiques Publiques (IPP) fait une classification des expérimentations ; http://www.ipp.eu.

vous l'avez exposé, mais pas ce qui se serait passé avec l'autre traitement (le PFIC vu dans la **section 2.2.3**), sur la même période.

Une EXP permet d'imaginer le protocole parfait, sans contrainte de ressources (Angrist et Pischke, 2009, 4-5), manipuler la variable causale et les facteurs à sa guise. La question causale est aussi précise que l'on veut. Pour ces auteurs, l'EX idéale (EXI) est le plus souvent l'EXP, car l'EXI n'est souvent réalisable qu'en pensée. En physique, c'est Einstein qui s'imagine à la fois dans le train et sur le quai. En mathématique, c'est le paradoxe de Zénon d'Élée rapporté par Aristote. En statistique, c'est la buveuse de thé de Fisher. Si on n'arrive pas à imaginer une EXP qui marche, ce n'est pas la peine de faire l'EX en vrai !

Les types d'EX qui suivent font référence à des EX réelles.

2.3.2 L'expérimentation de laboratoire

L'expérimentation de laboratoire (EXL) permet de recréer une situation économique, comme une enchère, le financement d'un bien public, etc.
C'est un outil puissant, surtout quand :

- Le comportement des agents économiques concernés est difficile à observer dans le monde réel (passager clandestin, entente, fraude, effet d'aubaine, etc.).
- Un marché n'a pas encore été mis en pratique. Une enchère par ex., mettant en évidence la fameuse malédiction du vainqueur (Harrison et List, 2004, 1022).

L'EXL permet un protocole randomisé. L'EX fictive de la buveuse de thé pourrait se faire en laboratoire très facilement.
L'EXL comporte une règle « du jeu » (le participants peuvent-ils interagir, etc. ?)
L'enjeu est faible, même si les joueurs peuvent être « rétribués ».
Limites de cette méthode d'analyse :

- Difficulté à généraliser les résultats d'une EXL à la population
  - C'est vrai de toutes les EX ! On dit dans ce cas que l'EX a peu de validité externe. Dans l'EX de l'*Asian desease problem*, les participants sont des étudiants.
- Environnement un peu stérile
  - Les participants savent qu'ils sont dans une EXL. Il n'y a pas d'enjeu, alors que dans la vraie vie, ils seraient confrontés aux conséquences de leurs décisions.

Les trois types de protocoles qui suivent débouchent sur des quasi-EX, des études observationnelles (EO), généralement, toutes les études dans lesquelles l'expérimentateur ne contrôle pas parfaitement le MAT. Et surtout, les études *ex post*, la majorité des études en économie.[25]

- On oppose les évaluations *ex post* à celles *ex ante*, qui comportent deux étapes : la construction d'un modèle théorique, calibré. Puis, l'utilisation de ce modèle pour simuler les RP (plus de détails sur le site de l'IPP) ; simulation de politiques en place ou contrefactuelles, permettant d'évaluer une future réforme. Une étape clé est la simulation de la législation (ex. : simuler l'impôt sur les sociétés). Les données législatives sont généralement en accès libre, sur Etalab, par exemple.

On peut ajouter un autre exemple. Le laboratoire d'action contre la pauvreté (J-PAL), fondé en 2003 au MIT par Abhijit Banerjee, Esther Duflo et Sendhil Mullainathan, a fondé sa réputation sur l'utilisation exclusive d'expérimentations contrôlées pour mesurer les effets des programmes sociaux et de lutte contre la pauvreté.

[25]. L'étude *exp post* est à l'économie ce que l'étude rétrospective est à l'épidémiologie.

J-PAL Europe, basé à PSE-École d'Économie de Paris, regroupe des chercheurs européens faisant des EPP dans de nombreux pays du monde.

[Aller faire un tour à https://www.povertyactionlab.org/fr, puis sur le site de l'IPP où se trouvent des tas d'expérimentations]

2.3.3 L'expérimentation de terrain

L'expérimentation de terrain (EXT) peut être vue comme une variante moins contrôlée d'une EXL. Il y a de l'attrition, du lobbying pour bénéficier du traitement (comme dans TICELEC). L'article de Harrison et List (2004) est complet sur ce type d'EX que nous n'allons pas plus développer. Notons seulement que Google mène ce type d'expérimentations. Le saviez-vous ? Il y a par exemple les « Geo Experiments », dont on peut trouver un article précurseur sur le site de Google, « Measuring Ad effectiveness using Geo expriments » (Vaver et Koehler, 2011) .
L'individu ne sait pas forcément

- qu'il participe à une expérience,
- qu'il est randomisé,
- que son comportement est scruté.

Il ne perçoit pas le fait d'être exposé à un traitement comme quelque chose qui n'est pas naturel.

Il faut relativiser l'importance des EXL et EXT en science sociale.

Par exemple, en matière de chômage, quelques requêtes sur Jstor au 20/05/2020 révèlent que « unemployment » a 306823 résultats, (« unemployment » and « propensity score »), 915, quant à (« unemployment » and « randomized experiment »), 475.

En confinant les requêtes aux publications de 2010 à 2020, nous obtenons 55448, 603 et 248 réponses (0,45 %). La part des études randomisées, en évoquant la randomisation, et en lien avec le problème du chômage, est significativement plus importante sur les 10 dernières années (elle a triplé),[26] mais reste marginale.

Si on remplace « unemployment » par « employment », on obtient 155190, 1365 et 571 (0,36 %) entrées sur la période 2010-2020.

2.3.4 L'expérimentation naturelle

Expérimentation naturelle (EXN) ou EX quasi expérimentale (Angrist et Pischke, 2009, 21)
Exemple de commerce international :[27]

L'UE, vers qui la Chine exporte 60 % de ses panneaux et composants solaires, s'est préoccupée du prix bas de ces derniers (dumping). En 2013, elle déclenche (provisoirement) une politique antidumping (PAD) sur les cellules photovoltaïques et les panneaux. La PAD « oblige » des importateurs de ces composants en provenance de Chine à modifier leur portefeuille d'inputs (en substituant des inputs non-taxés à ceux de Chine taxés) où sortir du marché.

Les entreprises qui changent ou pas de fournisseur le font pour une raison en partie exogène, la PAD, qui est un choc réglementaire. En forçant ces importateurs à changer de fournisseur, la PAD permet à l'expérimentateur de comparer ensuite les performances des entreprises qui modifient leur portefeuille avec celles qui ne le modifient pas. Et donc, de répondre à des questions causales dans de meilleures conditions de contrôle de

[26]. (248/55448)/(475/306823) ≅ 2,9.
[27]. Chen, Y. (2015). EU-China solar panels trade dispute: settlement and challenges to the EU. European Institute for Asian Studies, 7 pp., pose bien le problème.

certains facteurs, comme par exemple, la question du lien entre portefeuille d'inputs et performance, ou changement de fournisseur et performance.

[Donner l'exemple du Poster du mémoire de Rania dans l'hôtellerie,
et Zubizarreta et alii (2014)]

L'expérimentation « Asian desease problem » de Tversky et Kahneman (1981) est un exemple historique d'EXL menée il y a 40 ans auprès d'étudiants de deux universités, Stanford et British Columbia. La situation est celle d'une épidémie qui va s'abattre sur une population de 600 personnes. Différents programmes de sauvetage peuvent être mis en place. Quel programme les participants vont-ils choisir ? La question causale est de savoir si l'attitude vis-à-vis du risque dépend de la manière de décrire la situation risquée (on pourrait parler d'incertitude radicale, mais ce n'est pas le propos de l'étude de l'époque). Dans son ouvrage « Pensée lente, pensée rapide », Kahneman (2011, 368-369) rappelle son expérimentation. Il y a également un résumé dans Glimcher et Fehr (2014, 459-460).

Les auteurs ont sélectionné deux universités afin d'avoir assez d'hétérogénéité dans l'échantillon pour contrôler l'influence de facteurs tels que la catégorie sociale des familles des étudiants, l'appartenance ethnique, etc. Dans chacune des deux universités, les deux présentations (cadres, « frame ») suivantes ont été faites :

Programme	Frame 1 (« saved »)		Programme	Frame 2 (« die »)	
A	200	109 (72%)	C	400	34 (22%)
B	$600\left(\frac{1}{3}\right)$; $0\left(\frac{2}{3}\right)$	43 (28%)	D	$0\left(\frac{1}{3}\right)$; $600\left(\frac{2}{3}\right)$	121 (78%)
	$N_1 = 152$			$N_2 = 155$	

152 étudiants ont dû faire un choix entre les programmes A et B dans le Frame 1, et 155 entre les programmes C et D dans le Frame 2. Le Frame 1 présente les effets des programmes A et B en termes de vies sauvées, tandis que le Frame 2 les programmes C et D en termes de vies perdues.

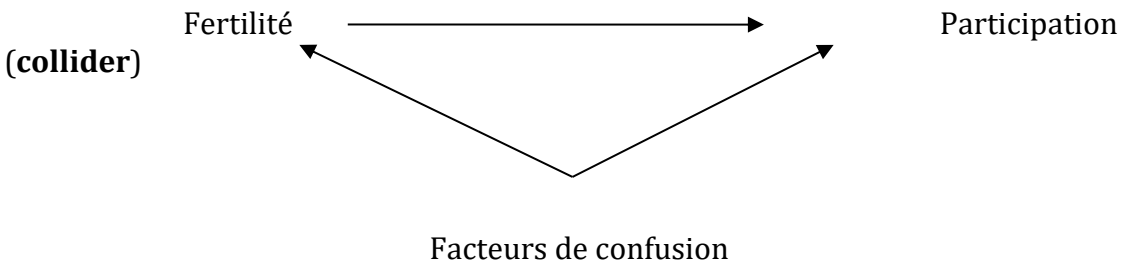
Utilisons nos critères pour décrire cette expérience de laboratoire :

- Il s'agit d'un problème de santé publique
- Traitements (causes) : cadres (« frames »), un cadre inclut 2 programmes
- Traitement expérimental : la conséquence de chaque programme décrit
- Résultat (conséquence) : attitude vis-à-vis du risque
- Résultat expérimental : choix d'un programme
- Unités d'observation/individus : étudiants
- Nature du bien : vies humaines fictives
- Enjeu : nul

Que révèlent les réponses des participants dans chaque Frame ? Les réponses dans le Frame 1 révèlent une aversion au risque des individus (risquophobes). L'individu préfère un gain certain (200) à son équivalent espéré. Le Frame 2 révèle au contraire un attrait pour la prise de risque (risquophiles) au sens où l'individu préfère des morts probables à certaines (400). Or, $600 \times (1/3) + 0 \times (2/3) = 200$ survivants et $0 \times (1/3) + 600 \times (2/3) = 400$ morts sont la même chose dans une population de 600 personnes. Le facteur responsable de cette incohérence des préférences est appelé **« framing effect » (effet de cadrage)**. Il n'y a pas d'enjeu, car l'étudiant ne risque rien, pourtant, on assiste à un « retournement » des préférences.

Cette EXL est restée célèbre. Daniel Kahneman a reçu en 2002, avec Vernon Smith, le prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel.

L'expérimentation de Zubizarreta *et alii* (2014) sur la participation des femmes au marché du travail montre l'intérêt d'isoler une EXN. Le schéma de cette EX est simple :



La fertilité est relativement simple à formaliser comme variable causale ; il s'agit d'une grossesse d'un ou plusieurs enfants (des jumeaux par exemple) qui arrive à son terme. La participation au marché du travail, ce qui est traité, l'est également ; la jeune maman travaille ou pas après avoir donné naissance. La variable de résultat expérimentale sera précisément les heures travaillées. Entre les deux, il y a des facteurs de confusion (les autres décisions dans la vie) tels que la relation avec le père, le niveau d'éducation (éducation et carrière sont interconnectés ; le niveau d'éducation détermine notamment les plans de carrière). La santé compte aussi.

Dans cette EXN la nature intervient sur le mécanisme de détermination du nombre d'enfants, en occurrence sur le fait d'avoir un ou deux enfants (des jumeaux). Par conséquent, le nombre d'enfants est aléatoire. Plus précisément, c'est conditionnellement au fait d'être enceinte que les événements « un enfant » ou « des jumeaux » (deux enfants) sont aléatoires. L'interaction avec les facteurs de confusion peut se montrer en partant de la probabilité d'avoir un certain nombre d'enfants, un garçon par exemple :

On peut partir de l'hypothèse que $\Pr(G|\text{enceinte}) = \Pr(F|\text{enceinte}) = 0{,}5$. Alors, $\Pr(G) = \Pr(G, \text{enceinte}) + \Pr(G,\sout{\text{enceinte}}) = \Pr(G|\text{enceinte}) \times \Pr(\text{enceinte})$, qui vaut donc $0{,}5 \times \Pr(\text{enceinte})$.

Pr(enceinte) dépend des facteurs de confusion. Pour cela, les auteurs sélectionnent une population précise de femmes. Ils excluent l'adoption, qui serait un signe de revenus plutôt élevés, et considèrent des femmes ayant le même nombre d'enfants, par exemple 1, avant le traitement (la famille qui a des jumeaux passe de 1 à 3 enfants, celles qui n'attend qu'un enfant, de 1 à 2). Les auteurs étudient la distribution des sexes biologiques dans les fratries. La famille qui a un deuxième enfant, et seulement cette famille, a généralement ce dernier de sexe différent, car on sait que ce type de famille ne cherche pas à avoir plus de deux enfants en général (il est plus probable de passer de G à G+F ou F à F+G ; c'est ce que montrent les données, et c'est cohérent avec la répartition femmes-hommes dans les sociétés humaines), et les familles ayant trop enfants dont des jumeaux ne font plus d'enfants.

Les auteurs trouvent qu'un enfant supplémentaire réduit les heures travaillées de 5 % par enfant supplémentaire.

Exemple du lien entre fertilité et participation au marché du travail. Il s'agit de l'expérimentation de Zubizarreta, Small et Rosenbaum (2014) qui vise à réponse à la question causale suivante : « est-ce qu'avoir des enfants affecte la participation des femmes au marché du travail ? » (voir l'encadré à la **page précédente**).

Contrairement à l'EXL, l'EXN n'est pas une construction artificielle. Les MAT dans chacun de ces exemples, le chercheur (Tversky et Kahneman, 1981) ou la nature (Zubizarreta, Small et Rosenbaum, 2014), sont +- randomisés. C'est le cas du deuxième exemple, à condition quand même de tenir compte d'autres facteurs. Car, dans cet exemple, la probabilité d'avoir un enfant pour un couple dépend de celle que la femme soit enceinte, qui elle-même dépend d'un tas de facteurs (de confusion).

Dans l'EXN, le MAT est un MAT non contrôlé (MATnC), au sens où le chercheur ne le contrôle pas. C'est la « nature » qui le fait. Néanmoins, Zubizarreta, Small et Rosenbaum (2014) considèrent cette intervention de la « nature » comme un protocole à part entière.

Ce type de MAT est utile quand les bénéficiaires d'un dispositif sont sélectionnés à partir de critères difficilement manipulables ; la manipulation ne serait pas éthique dans l'évaluation de Zubizarreta, Small et Rosenbaum (2014) ; en effet, il n'est pas éthique de manipuler le nombre d'enfants à la naissance ! Nous reviendrons sur l'avantage des études observationnelles pour éviter ce type de question sur l'éthique (cf. **chapitre 4**).

Pour faire le lien avec le type d'EX suivant, disons que l'EXN est une EXS dans laquelle c'est un « agent » extérieur, la nature, qui manipule les traitements, car tout y est réel.

2.3.5 L'expérimentation sociale

L'EX sociale (EXS), sous-entendu, dans la société. Un exemple récent d'EXS qui a demandé pas mal de moyens, est celle faite pour l'évaluation du crédit d'impôt compétitivité emploi (CICE).

Dans une EXS, tout est réel : les individus (entreprises, ménages), l'environnement dans lequel ces individus agissent, etc. L'évaluatrice a peu de contrôle, voire aucun, sur le MAT, qui n'est de surcroit pas randomisé.

On peut généralement construire un « groupe de contrôle » *ex post*, à condition que les individus observables n'aient pas tous été exposés au traitement actif.

> Il n'y a pas de groupe de contrôle quand la PP mise en place s'applique à tous les individus. Cette situation conduit à un paradoxe, appelé paradoxe de Lord, que nous ne développerons pas dans ce cours (Lord, 1967).

Notons enfin que l'EXS est sans doute le protocole le plus courant pour l'EPP. Et ce n'est pas parce que les modèles économétriques et les protocoles expérimentaux de type randomisés divergent (Lalonde, 1986), que ces derniers sont meilleurs. Il y a un mouvement pour imposer les protocoles randomisés, représenté par des auteurs dont Esther Duflo et ses co-auteurs, dont Bruno Crépon en France. En face, il y a James Heckman, Angus Deaton. La controverse dépasse le cadre de **ce cours**.

2.4 Exercices sur le chapitre 2

2.4.1) Poser une question causale de votre choix, et imaginer l'expérimentation mentale que vous voudriez mener pour l'évaluer.

2.4.2) Paradoxe de Yule-Simpson (tableau 2.2). Notons $Y(\text{"trouve un travail"}) = 1$, $D(\text{"Master d'économie"}) = 1$ et $X(\text{"habite en banlieue}) = 1$. La valeur des variables Y, D, X pour les autres événements contraires est zéro : $Y(\text{"ne trouve pas un travail"}) = 0$, $D(\text{"Master de Droit"}) = 0$ et $X(\text{"habite Paris}) = 0$. L'effet « Economie » est $E(Y|D = 1) - E(Y|D = 0)$. En utilisant la LEI, on arrive à comprendre formellement le problème de dépendance à l'origine du « paradoxe ».

a) En vous appuyant sur la décomposition chiffrée de l'exemple, écrire $E(Y|D = 1)$.

b) Si le choix d'une filière et du lieu d'habitation étaient indépendants, Pr $(X = 0|D = 1)$ serait égal à Pr $(X = 0)$ et Pr $(X = 1|D = 1)$ à Pr $(X = 1)$. Réécrire l'effet « économie » dans ce cas, de manière formelle.

c) Qu'est-ce que ça donne avec les chiffres du tableau.

d) Quelle est la valeur du biais ?

2.4.3) Que veut dire « résultat potentiel contrefactuel » dans l'évaluation d'une réforme de l'assurance chômage ?

Remarque : si vous le voulez, vous pouvez utiliser dans votre réponse les notations $Y_i(0)$ et $Y_i(1)$

2.4.4) Avec la réforme du collège entreprise par Mme Najat Vallaud-Belkacem en 2016, la deuxième langue étrangère (LV2) était enseignée dès la 5e au lieu de la 4e, pour une durée hebdomadaire réduite (2h30 au lieu de 3h). Cette réforme est une occasion de tester l'effet causal d'une réduction de l'enseignement d'une LV2 sur l'apprentissage des langues. On note i et j deux élèves qui rentrent respectivement en 5e et 4e en septembre 2016. La variable de résultat expérimentale pour un élève $k \in \{i, j\}$ est la moyenne annuelle de ses notes de LV2 en 4e. La moyenne observée pour k en juin de l'année t est $Y_{k,t}$, et $Y_{k,t}(1)$ est cette moyenne si l'élève a bénéficié de la réforme et $Y_{k,t}(0)$ sinon.

a) À quel résultat potentiel est égal $Y_{i,2017}$? Même question pour $Y_{j,2016}$.
b) Si en septembre 2016, j est en 4e pour la 2e fois, quels sont <u>ses</u> deux résultats potentiels observables ?
c) La quantité $Y_{i,2017}(1) - Y_{i,2017}(0)$ est-elle observable ?
d) $Y_{i,2017} - Y_{j,2016}$ et $Y_{j,2017} - Y_{j,2016}$ sont-ils des estimateurs biaisés (on maintient b) ? Pourquoi ?

2.4.5) Pour chacune des affirmations ci-dessous, dites si elle vous semble fausse

- ☐ L'expérimentation idéale est telle que le MAT est aléatoire
- ☐ La variable de résultat doit toujours être observée avant une évaluation ex-post
- ☐ Les expérimentations de laboratoire sont des études observationnelles.

2.5 Notes

Nous attribuons le MCR à Donald B. Rubin (https://g.co/kgs/zjyCXH) et Jerzy Neyman (https://g.co/kgs/hdB2zZ) un peu plus tôt. D'après James Heckman (https://g.co/kgs/TSLdPp), l'inventeur d'un estimateur en deux étapes pour corriger le biais de sélection, et récompensé pour son travail sur l'EPP, nous devrions attribuer l'approche contrefactuelle du MCR à Haavelmo et Quandt ; voir Heckman (2005). Les développements récents sur le lien entre le MCR et la causalité à la Granger sont dus à Halbert Lynn White Jr. (https://g.co/kgs/nb8qaR), le père des erreurs-types robustes (l'option `robust` de Stata !), décédé en 2012. [Lechner, 2011]

On a pas abordé le concept de corrélation fallacieuse (spurious correlation), développée par Granger et Newbold (1974), et qui remonte au moins à Yule (1925). Il y en aurait trop à dire car, ces concepts de causalité et de corrélation posent aussi la question de ce que l'on entend par exogénéité, avec le fameux article de Engle *et alii* (1983), entre autres, qui mériteraient d'être connus de tous les chercheurs qui doivent faire un tri entre les différents déterminants d'une variable. La Banque de Suède en la mémoire d'Alfred Nobel a, en 2021, récompensé les méthodes d'évaluation, en remettant sont prix à D. Card, J. Angrist et G. Imbens, après avoir, en 2019, récompensé E. Duflo, pour une application de ces méthodes pour sortir des ménages de la précarité.

Au principe de causalité, qui guide la science, se superpose le principe de corrélation, paradigme du *Big data*. Le *data scientist* produit des corrélations à l'aide d'algorithmes. L'inférence causale n'est pas sa priorité. Il part d'un échantillon, en général justifié pour des raisons de coût de sondage et de stockage (argument classique en statistique). Ce coût étant beaucoup moins élevé qu'avant, le *Big data* part d'une masse de données aussi large que possible, comme si on avait affaire à la population ! On étudie une corrélation à l'aide d'algorithmes et de mathématique, plus que de statistique, quitte à comprendre plus tard s'il n'y a pas une cause qui a conduit à la corrélation observée ; Peliks (2014). La différence avec l'économétrie se situe sans doute dans le fait que l'économètre construit des estimations en relation avec la théorie économique, et considère qu'identifier une relation causale reste le but de l'économétrie, que cette relation soit la transposition d'un modèle tiré de la théorie économique ou pas. Par ailleurs, les deux s'intéressent à la prédiction (la prévision dans le cadre de séries temporelles).

Correction des exercices du chapitre 2

2.4.1)
(réponse orale des étudiants en classe)

2.4.2)
a) $E(Y|D=1,X=0)\Pr(X=0|D=1)+E(Y|D=1,X=1)\Pr(X=1|D=1)$.
b) $[E(Y|D=1,X=0)-E(Y|D=0,X=0)]\Pr(X=0)+$
$[E(Y|D=1,X=1)-E(Y|D=0,X=1)]\Pr(X=1)$.
c) $(93-87)\frac{357}{700}+(73-69)\frac{343}{700}=5{,}02$.
d) Le biais vaut donc $(-5)-(5{,}02)=-10{,}02$.

2.4.3)
C'est le RP que l'on n'observe pas. Pour un chômeur après réforme, c'est son insertion sur le marché du travail s'il n'y avait pas eu la réforme. Pour un chômeur avant réforme, c'est son insertion s'il y avait la réforme.

2.4.4)
a) Seul $Y_{i,2017}(1)$ sera observable en juin 2017, donc $Y_{i,2017}=Y_{i,2017}(1)$. Et, $Y_{j,2016}=Y_{j,2016}(0)$, car j n'a pas pu, en juin 2016, bénéficier de la réforme introduite en septembre 2016.
b) Bien qu'en juin 2016, j avait déjà une LV2, il ne s'agissait pas de la LV2 réformée, introduite en septembre 2016 (question a). Donc $Y_{j,2016}=Y_{j,2016}(0)$, mais $Y_{j,2017}=Y_{j,2017}(1)$, il faut attendre juin de l'année suivante.
c) L'ECI de i, $Y_{i,2017}(1)-Y_{i,2017}(0)$ n'est pas observable (PFIC), car $Y_{i,2017}=Y_{i,2017}(1)$.
d) $Y_{i,2017}-Y_{j,2016}=Y_{i,2017}(1)-Y_{j,2016}(0)$ est d'abord intéressant car on compare i qui a eu LV2 réformée en 5e avec j qui a eu LV2 non-réformé en 4e. Même si c'est un estimateur biaisé, car les individu ne sont pas comparables en termes d'âge notamment, c'est le mieux qu'on ait car $i,2017$ est nécessaire si je veux évaluer la réforme sur des élèves de 5e (l'ECMT).
$Y_{j,2017}-Y_{j,2016}=Y_{j,2017}(1)-Y_{j,2016}(0)$ a l'avantage de comparer j avec lui-même. Mais, une année s'est écoulé (j avec réforme en juin 2017 avec j sans réforme en juin 2016), mais surtout car en juin 2016, j était déjà en 4e, donc avait déjà eu LV2 puisqu'avant la réforme, la LV2 était dès la 4e. Celle-ci est le mieux si je veux évaluer la réforme sur les élèves de 4e. Mais $j,2016$ est le meilleurs groupe de contrôle possible.

2.4.5)
1 fausse car elle reviendrait à traiter des individus qui en n'ont pas besoin, et ne pas traiter des individus qui en ont envie. Cette question soulève des jugements de valeur ; par ex., pour quelqu'un à la rue, il faut lui proposer un toit, qu'il en ait envie ou pas. C'est l'une des raisons pour lesquelles il faut distinguer **affectation** et **participation**. Autre ex., en médecine, on donne le meilleur médicament connu à des enfants malades.
2 la réponse n'est pas tranchée
3 fausse par définition, car dans une EXL, l'expérimentateur contrôle le MAT, alors que dans une EO, il n'intervient pas.

3. Sélection aléatoire des individus, inférence

Ce chapitre introduit le concept de mécanisme d'affectation des traitements (MAT) dans les groupes de traitement (section3.1). Le mécanisme peut être aléatoire ou pas, contrôlé ou pas. Après avoir expliqué simplement ce qu'est un MAT confondu (section 3.12, nous nous arrêterons sur le cas du MAT aléatoire (randomisé) contrôlée (MATAC) dans lequel les individus sont sélectionnés aléatoirement dans les groupes de traitement tout en contrôlant leur nombre dans chaque groupe. Nous verrons les vertus de la randomisation (section 3.3) qui permet à l'EMT d'estimer l'ECM sans biais (il ne s'agit pas ici du biais de sélection qui sera vu dans le chapitre suivant, mais du biais d'estimation). Enfin, nous verrons des estimateurs et statistiques de tests pour MATAC (section 3.4), le test exact de Fisher, le test de Neyman, et l'ANOVA, que nous appliquerons aux données d'évaluation d'un programme éducatif américain des années 1970, The Electric Company. Ce sera l'occasion d'introduire la méthode d'appariement par paires.

3.1. Mécanisme d'affectation des traitements

Les traitements auxquels sont exposés les individus sont déterminées par un **mécanisme d'affectation des traitements** (MAT). Le MAT fait partie du protocole expérimental, au même titre que le choix de la variable de traitement expérimentale, du résultat expérimental (tableau 1.1). Une définition simple du MAT dans le cas de deux traitements est la suivante :

Définition 3.1 : le MAT est le mécanisme qui détermine pour chaque unité i la valeur de son traitement D_i, donc le RP qui sera observé, $Y_i = Y_i(D_i)$ et le RP contrefactuel $Y_i(1 - D_i)$.

On représente ce mécanisme par la probabilité conditionnelle que les traitements $D_1, \dots, D_N$ prennent respectivement les valeurs $d_1, \dots, d_N$, avec $d_i \in \{0; 1\}$. Le modèle est assez général pour tenir compte de l'affectation des traitements en fonction des RP – ce qui ouvre la voie aux anticipations –, et faire intervenir des déterminants observables, $\boldsymbol{X}$ (variables de confusion, de contrôle, etc.) ; Imbens et Rubin (2015, 34) :

$$\Pr(D_1 = d_1, \dots, D_N = d_N | \boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X}) \equiv f(d_1, \dots, d_N, \boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X}). \qquad (3.1)$$

Envisageons un cas plus simple en pratique : les traitements sont affectés aléatoirement en jetant N fois une pièce (pile, l'individu 1 a le traitement actif et face, il a le traitement passif, *idem* pour l'individu 2, etc.). Alors f ne dépend de rien d'autre que des N lancers :

$$f(d_1, \dots, d_N) = (1/2)^N.$$

Avec 12 individus, par exemple, chaque affectation a une probabilité de 0,025 % environ.

Remarques :

- Un MAT qui ne dépend pas des RP, $\Pr(d_1, \dots, d_N | \boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X}) = \Pr(d_1, \dots, d_N | \boldsymbol{X})$, est un MAT non-confondu (Imbens et Rubin, 2015, 38). Par exemple, $\Pr(D_i = 1 | X = x) = 20/(20 + x)$.
- On a zappé une étape préalable importante : l'échantillonnage. On supposera qu'il est aléatoire afin de ne pas se poser le question de la représentativité.
- $Y(1)$ et $Y(0)$ ne sont pas nécessairement des v.a., contrairement à ce qui est suggéré dans (3.1), mais des quantités fixes.

Se pose la question de savoir qui contrôle le MAT. Trois cas :

1 L'expérimentateur ? comme dans l'évaluation de Rubin (1977). Par ex., le programme pilote TICELEC (cf. tableau 1.1) ?
2 Une institution (le Parlement) ? La nature ? comme pour la réduction de la demande de nuitées dans le secteur de l'hôtellerie pendant la Covid.
3 Les individus eux-mêmes ? Dans ce cas, on parle de sélection non-aléatoire des individus (Imbens et Rubin, 2015, 32, 41) ; *self-selection into treatment*.

Ce n'est que dans le premier cas que nous parlerons de contrôle ! Si un individu exposé à un traitement n'aurait pas pu être exposé à l'autre, il n'y a pas de contrôle !

Dans ce chapitre on s'intéresse au **MAT aléatoire contrôlé** (MATAC), f est connue ! Le MATAC est le protocole des EXL, les EXT. Par ex., dans l'EX de Bertrand et Mullainathan (2004), des noms hypothétiques, que les employeurs associent à tel ou tel continent d'origine, furent randomisés. En revanche, dans une EXS, comme l'évaluation d'une aide à la R&D, le MAT par lequel une entreprise obtient une aide n'est ni aléatoire ni contrôlé (f est inconnue) ; le cas 3 ci-dessus.

3.2. Problème du MAT confondu

Cette section montre le problème d'un **MAT confondu** (le MAT dépendant des RP) dans un exemple fictif. L'exemple est adapté de Rubin (2004) et Imbens et Rubin (2015, 14-15). On reprend les données du tableau 2.3 où l'ECM, qui vaut 2, reflète l'idée que le soutien à l'innovation est plutôt une bonne PP. Dans le tableau, on imagine ce que pourraient être les taux de variation de la R&D, avec subvention et Crédit d'impôt recherche (CIR), $Y(1)$, ou seulement avec subvention, $Y(0)$.[28]

Tableau 3.1 : Effet du CIR sur les dépenses de R&D (Taux de variation en 2014, %)

i	$Y(0)$	$Y(1)$	ECI_i	D_i	Y_i
1	14	13	−1	0	14
2	0	6	6	1	6
3	1	4	3	1	4
4	2	5	3	1	5
5	3	6	3	1	6
6	1	6	5	1	6
7	10	8	−2	0	10
8	9	8	−1	0	9
Total	40	56	16		60
Moyenne	5	7	2 (ECM)		7,5

Supposons que le MESRI a suffisamment d'information sur les entreprises pour n'autoriser le CIR qu'à celle dont $Y_i(1) > Y_i(0)$. L'évaluateur n'observe que D_i et Y_i.

a) Refaites le tableau avec la colonne D_i qui reflète ce MAT, et la colonne Y_i
b) Peut-on calculer les ECI_i ? Et l'ECM ?
c) Calculez l'**effet moyen du traitement** (EMT) par la différence des moyennes $\bar{y}_1 - \bar{y}_0$?

$$\bar{y}_1 - \bar{y}_0 = \frac{1}{5}(6+4+5+6+6) - \frac{1}{3}(14+10+9) = 5{,}4 - 11 = -5{,}6.$$

[28] Les données reflètent les taux de variation en 2014 des 8 premières entreprises françaises du top 2500 mondial : SANOFI (1,2), Peugeot (15), et Alcatel-Lucent (−5,2), Valeo (3,7), qui arrive en 9e position, etc. Les données sont tirées de l'EU R&D Investment Scoreboard (European Commission, 2016).

d) Le MESRI a-t-il pris un bonne décision ?
Oui en terme d'accélération de la R&D (elle est maximum), mais pas en termes d'évaluation (l'EMT est négativement biaisé).

> Si on veut évaluer l'effet d'une PP, on ne doit pas affecter les individus sur les valeurs observées ou prédites des résultats potentiels (RP) sous différents traitements, parce que ça induit fatalement un biais.

Commentaires

L'EMT < 0 s'explique par la distribution des RP. Avec l'affectation du MESRI, le taux de variation de la R&D des entreprises tests est plus élevé avec que sans CIR, mais ces taux sont plus faibles que ceux des entreprises témoins ($\bar{y}_1 - \bar{y}_0 = -5{,}6$). Néanmoins, le RO moyen obtenu avec l'affectation du MESRI ($\bar{Y} = 7{,}5\ \%$) est légèrement supérieur à celui de l'affectation où toutes les entreprises ont le CIR, $\overline{Y(1)} = 7\ \%$.

Supposons maintenant que le MESRI soutienne les entreprises connues pour avoir des taux de croissance de la R&D élevés ($I_1 \equiv \{1; 7; 8\}$). Comme on vient de le voir, ce sont aussi celles qui, lorsqu'elles ont le CIR, réduisent légèrement leur R&D. Le taux de croissance moyen de la R&D baisse ($\bar{Y} = 4{,}5\ \%$). En revanche, l'EMT vaut plus de quatre fois l'ECM : $\bar{y}_1 - \bar{y}_0 = (13 + 8 + 8)/3 - (0 + 1 + 2 + 3 + 1)/5 = 8{,}27$.

On voit que le problème est plus statistique (le biais de l'EMT) qu'économique (la croissance de la R&D). Cet exemple hypothétique pose trois questions : (i) L'ECM est-il la bonne mesure d'efficacité ? (ii) Quel est le bon vecteur d'affectation sachant qu'une affectation bonne pour l'économie ne le sera pas forcément pour évaluer l'efficacité d'un dispositif ? Ces deux questions sont liées. Et (iii) qu'apporterait une évaluation randomisée ?

L'ECM compare deux états contrefactuels : toutes les entreprises sont traitées *vs* aucune. La sélection est non-confondue (elle ne dépend pas de $Y_i(1) - Y_i(0)$), certes, mais il y a deux problèmes. L'EMT nécessite un groupe de contrôle. De plus, certaines entreprises n'ont pas besoin de CIR (effet d'aubaine), le budget de la France est contraint.

En laissant le choix de recours aux entreprises, la sélection est toujours non-aléatoire mais devient confondue. La bonne mesure d'efficacité est l'ECMT. La confusion du MAT est un problème statistique que l'on atténuera en ramenant d'autres variables. De possibles effets d'aubaine existent, mais il n'y a pas plus d'entreprises concernées que dans la situation correspondant à l'ECM où on aiderait toutes les entreprises.

L'évaluation randomisée a des vertus que nous allons voir, et dont nous parlerons souvent dans ce cours. Mais elle soulève la question de savoir comment utiliser ce protocole pour estimer l'ECMT …

3.3. Vertus du MAT aléatoire contrôlé

Dans cette section, nous rappelons les concepts d'**hypothèse nulle**, **alternative**, de **seuil de significativité**, et introduisons celui de **p-valeur exacte de Fisher**. Ce dernier a imaginé un test dans une EXP célèbre, celle de la buveuse de thé, que l'on trouve dans Rosenbaum (2010), Grima (2013), Salsburg (2002), etc. Dans cette EX, le MAT est aléatoire et contrôlé.

3.3.1. A l'origine, l'EX de Fisher de la buveuse de thé

« À la fin des années 1920, à Cambridge, en Angleterre, un groupe de professeurs, leurs épouses et leurs invités prennent le thé dehors, profitant d'un après-midi agréable.

Après avoir bu une première gorgée, une dame, une tasse à la main, déclare avoir noté que la saveur du thé est différente selon que le thé est servi avant ou après le lait.

Poliment, bien sûr, une personne fait quelques remarques sur la difficulté d'admettre ce fait et cela donne lieu à une discussion au cours de laquelle on brandit toutes sortes d'arguments tirés du monde de la physique et la chimie : la composition du produit résultant est la même si l'on verse d'abord le thé ou le lait [physique], les particules dissoutes sont à la fin toutes les mêmes [chimie], la différence de température ne joue pas, etc. Il semble impossible de distinguer une tasse d'une autre …

*Une des personnes présentes, un homme d'environ 40 ans nommé Ronald Aylmer Fisher [de la statistique de Fisher, la p-valeur et autres contributions], propose de dissiper tous les doutes possibles au moyen d'une procédure 'révolutionnaire' : effectuer une […] dégustation » [dans le cadre d'un **P** aléatoire contrôlé]*

Nous n'avons aucune information sur la buveuse, à part qu'elle dit ne jamais se tromper. » Grima (2013).

a) Imaginons une expérimentation avec une tasse.

Il y a deux résultats possibles selon que le lait est versé avant ou après (on tire une pièce pour décider). Pile je mets le lait avant, face c'est le contraire. C'est pile. Soit la buveuse devine, soit elle ne devine pas.

Fisher pose l'hypothèse nulle (H_0) que le traitement n'a pas d'influence ; i.e., la buveuse ne sait pas faire la différence (elle tire ses réponses au hasard) face à l'hypothèse alternative (H_1) qu'elle sait faire la différence. H_0 s'appelle **hypothèse exacte de Fisher**. Sous H0, la buveuse aurait la même réponse, même si nous avions mis le lait après !

En donnant sa réponse au hasard, la buveuse aurait donc une chance sur deux de deviner … par hasard. En termes statistiques, la probabilité qu'elle devine par hasard (rejet de H0 à tort) est donc de 50 % (sous H_0) ! C'est trop, ce protocole n'est pas satisfaisant. C'est une ***p-value* exacte**. L'idée géniale de Fisher fut de proposer le protocole suivant : préparer 8 tasses de thé (4 de chaque mélange), et les disposer aléatoirement (une permutation) devant la buveuse.

Cette idée peut se généraliser à des tas d'autres EX !

Figure 3.1 : une combinaison huit tasses numérotées de « 1 » à « 8 »

Position des tasses	1	2	3	4	5	6	7	8
Ingrédient versé en premier	Lait	Thé	Lait	Thé	Thé	Thé	Lait	Lait

b) Quel est le nombre de combinaisons des huit tasses avec quatre de chaque mélange ?

$70 = C_8^4$. Il y a 98,6% de chance (`1-1/comb(8,4)` dans Stata), sous H_0, que la buveuse ne trouve pas (exactement 69 combinaisons qu'elle ne sait pas différencier). Pour comprendre la *p-value*, supposons qu'elle donne la même réponse quelle que soit la combinaison de tasses devant elle ; ses réponses sont par exemple celles ci-dessous, quelle que soit la combinaison qui lui est présentée :

Figure 3.2 : la réponse de la buveuse

Position de la Tasse	1	2	3	4	5	6	7	8
Ingrédient versé en premier	Lait	Thé	Lait	Thé	Thé	Thé	Lait	Lait
Réponse	Lait	Lait	Lait	Lait	Thé	Thé	Thé	Thé

Si elle identifie correctement les 8 tasses, c'est par accident, car il y a 1 chance sur 70 (1,4 %) que la combinaison qui lui est présentée soit celle-ci. C'est la probabilité de rejet de H_0 à tort (la p-valeur exacte).

3.3.2. Le MATAC en pratique, l'affectation des traitements ?

Reprenons le tableau 3.1 et supposons un MATAC consistant à exposer aléatoirement 5 entreprises au traitement actif (3 au traitement de contrôle). Il y a $C_8^5 = 56$ vecteurs d'affectations possibles. La probabilité d'un vecteur est 1/56 (1,78 %). Bien que non-aléatoire, le vecteur d'affectation du tableau, (0, 1, 1, 1, 1, 1, 0, 0), est nécessairement l'un des 56 vecteurs de la feuille Excel (le 21e) ; voir **`mat56.pdf`**.

Le fait de contrôler le MAT nous permet de retenir des affectations aléatoires où un nombre raisonnable d'entreprises sont traitées. Car, il y a en fait 2^8 vecteurs d'affectation possibles, dans le cas d'un **MAT de Bernoulli**. Ce MAT n'est pas intéressant pour des échantillons de taille finie, car les deux vecteurs d'affectation suivants sont possibles :

- (1, 1, 1, 1, 1, 1, 1, 1) : pas d'individu témoin, ou
- (0, 0, 0, 0, 0, 0, 0, 0) : pas d'individu test.

En contraignant le MAT de Bernoulli à avoir N_1 individus parmi N exposés au traitement actif, on a un MAT qu'Imbens et Rubin (2015, 25) appellent **MAT pleinement randomisé** (*Completely randomized experiment*).

Illustrons comment obtenir un vecteur d'affectations des traitements avec **Excel** grâce à la fonction **ALEA.ENTRE.BORNE(a; b)** où $[a; b]$ est la plage de valeurs dans laquelle Excel va tirer un nombre aléatoire entier distribué uniformément ; fichier **`mat-ac.png`**.

3.3.3. MATAC et biais de l'EMT

L'objet de cette sous-section est de voir le lien entre MATAC et biais de l'EMT, dans une approche fréquentiste (en moyenne, la différence des moyennes est-elle sans biais ?).

Petit rappel d'estimation sans biais (approche fréquentiste)

Supposons un échantillon aléatoire de deux individus piochés sans remise dans une population de trois individus numérotés 1, 2 et 3 ($\mathcal{P} \stackrel{\text{def}}{=} \{1; 2; 3\}$). L'ordre de tirage ne compte pas (les deux individus sont tirés ensemble). On suppose pour simplifier que l'univers $\Omega = \mathcal{P}$: la variable aléatoire est simplement $Y(\{i\}, i \in \mathcal{P}) \subset \mathbb{R}$. Les probabilités associées, $p(i) = 1/3$. L'espérance de Y est $(1 + 2 + 3)/3 = 2 \stackrel{\text{def}}{=} \bar{Y}$.

Il y a $C_3^2 = 3$ échantillons aléatoires équiprobables : $\{1\ ;\ 2\} \stackrel{\text{def}}{=} S_1$; $\{1\ ;\ 3\} \stackrel{\text{def}}{=} S_2$ et $\{2\ ;\ 3\} \stackrel{\text{def}}{=} S_3$. La moyenne d'échantillonnage $\sum_{k=1,2,3} \hat{Y}_k \Pr(S_k)$, avec $\Pr(S_k) = 1/3$, est sans biais. En effet, $\hat{Y}_1 = 3/2$, $\hat{Y}_2 = 4/2$ et $\hat{Y}_3 = 5/2$, et la moyenne d'échantillonnage :

$$\frac{3}{2} \times \frac{1}{3} + \frac{4}{2} \times \frac{1}{3} + \frac{5}{2} \times \frac{1}{3} = \frac{12}{6} = 2\,,$$

ce que l'on note généralement par $E\hat{Y} = \bar{Y}$.

Nous voulons savoir si l'EMT est un estimateur sans biais de l'ECM quand le MAT est pleinement randomisé. Autrement dit, si nous pouvions randomiser et répéter l'EX autant de fois que l'on veuille, EMT serait-il égale à ECM, en moyenne ?

Notons qu'à la différence de l'encadré, on veut savoir si une différence de moyennes, non pas une moyenne, est sans biais, comme pour un test d'égalité de moyennes. En revanche, comme dans l'encadré où $\hat{Y}_1 \neq \bar{Y}$ (on a qu'un moyenne d'échantillonnage, qu'un tirage), vous n'aurez qu'un seul vecteur d'affectations, et donc qu'un EMT, de sorte que EMT ≠ ECM en général.

Dans l'exercice 3.5.5 vous devez démontrer ce résultat ; en notant l'ECM $\overline{Y(1) - Y(0)} \equiv \tau$ et l'estimateur EMT par $\bar{Y}_1 - \bar{Y}_0 \equiv \hat{\tau}_{\text{dif}}$, vous devez démontrer que $E_D(\hat{\tau}_{\text{dif}}) = \tau$ (l'indice D souligne que les moyennes $\bar{Y}_d = N_d^{-1} \sum_{i:D_i=d} Y_i$, $d \in \{0\,;\,1\}$ sont aléatoires à cause des variables D_i, et non à cause de $Y_i(1)$ et $Y_i(0)$ qui sont des quantités fixes dans l'exercice).

$E_D(\hat{\tau}_{\text{dif}}) = \tau$ peut se démontrer rapidement quand on suppose les RP aléatoires (Angrist et Pischke, 2009) et qu'on a un **MAT ignorable** (*ignorability*), i.e. : la randomisation de D_i entraîne $\big(Y_i(1), Y_i(0)\big) \perp D_i$. La seule distribution de probabilité utilisée pour l'inférence est celle créée par l'évaluateur (Rosenbaum, 2010, 23).

Le test exact de Fisher s'appuie sur ces idées. C'est une approche non-paramétrique de l'estimation de l'ECM, reposant sur l'**imputation** des RP contrefactuels. Fisher substitue les RO aux RP manquants sous l'hypothèse nulle d'absence d'effet. En comparaison, il y a l'approche de Neyman de construction d'un intervalle de confiance, sans imputation.

3.4. Tests de causalité pour MAT pleinement aléatoire

Nous allons voir deux types de tests. Le premier découle des travaux de Fisher et le second, plus classique, des travaux de Neyman. Le test de Fisher est faisable pour des populations de petite taille. L'hypothèse testée est celle d'absence d'effet causal individuel, $H_0 : Y_i(1) - Y_i(0) = 0, \forall i$, ou plus généralement que l'effet causal individuel est exactement égal à une valeur donnée τ :

$$H_0 : Y_i(1) - Y_i(0) = \tau \in \mathbb{R}, \forall i.$$

Nous allons construire ce test pour une évaluation concrète dans le domaine éducatif, qui figure dans Gelman et Hill (2007, 174-176) et Rubin (2005).

3.4.1. Test Exact de Fisher : application aux données de The Electric Company

Une EX dans le domaine éducatif fut conduite aux États-Unis dans les années 1970, sur l'effet d'un show télévisé ludique, « The Electric Company » (traitement actif), visant à améliorer la lecture et le niveau de grammaire d'élèves de primaire. Une douzaine d'écoles furent sélectionnées dans chacune des deux villes suivantes : Fresno (dans l'ouest de l'Etat de Californie) et Youngstown dans le Nord Est de l'Ohio. Dans chaque école, et chaque niveau, les deux classes ayant les moins bonnes moyennes en lecture furent tirées, dont une seule sélectionnée aléatoirement dans le groupe de traitement.

C'est un **tirage en grappes** dans un protocole de **comparaison par paires** (*Paired Comparisons Design*). C'est un protocole ancien. On parle de *matched* (*related*) *samples*, mais aussi *matched pair samples* (Nelson, 2004). Une illustration du tirage en grappes, avec deux écoles, **`ecoles.png`**. Nelson (2004, 105-106) précise que « pairs can also be obtained by using single units twice ».

Ce protocole fut adopté dans les quatre niveaux scolaires (*grades*), du CP au CM1. L'échantillon comporte 192 classes (96 tests, 96 témoins) ; Gelman et Hill (2007), étudient cette EX. (Nelson, 2004). Le fichier de données s'appelle

electric.company.csv. Il comporte 96 observations, 46 pour Fresno, 50 pour Youngstown. L'effectif des classes varie par ville et niveau.

Nous allons travailler sur un sous-ensemble de six classes afin de pouvoir visualiser les calculs intermédiaires. Le jeu des données retenues est le suivant (**Figure 3.3**).

Figure 3.3 : The Electric Company : RP avant imputation

	unit	treatment	y	y0	y1
1	1	0	55.0	55.0	NA
2	2	0	72.0	72.0	NA
3	3	0	72.7	72.7	NA
4	4	1	70.0	NA	70.0
5	5	1	66.0	NA	66.0
6	6	1	78.9	NA	78.9

Comme on peut le voir sur la figure, il y a trois classes test et témoins. On retrouve – pratiquement – ces RP dans la base complète (Fresno, 2 écoles de CP, 1 de CE1). On ne tient pas compte des paires pour l'exercice. Le test exact de Fisher comporte six étapes :

1) **S**pécifier une hypothèse nulle H_0 sur la taille (***effect size***) de l'ECI pour chaque i
2) **S**pécifier une statistique de test
3) **E**stimer la statistique et préciser les valeurs qui seront plus grandes
4) Imputer les résultats manquants (les RP contrefactuels) à partir des RO, sous H_0
5) Pour chaque nouvelle affectation, **E**stimer la statistique (sous H_0)
6) Le **T**est consiste à déterminer la **p-valeur** (empirique), c'est-à-dire la probabilité des affectations pour lesquelles les statistiques sont supérieures ou égales à celles de l'échantillon des RO

Application

1) L'hypothèse de Fisher est généralement celle d'absence d'ECI : H0 : $Y_i(1) = Y_i(0) + \tau, \forall\, i$, avec τ fixé. Il s'agit d'une hypothèse simple (***sharp***)[29]
2) Une statistique de test courante est la différence des moyennes, $\bar{Y}_1 - \bar{Y}_0$, l'EMT
3) La valeur observée, $\bar{y}_1 - \bar{y}_0$ est 5,067 que l'on arrondit à 5,1. Les valeurs supérieures sont évidentes : $\bar{Y}_1 - \bar{Y}_0 > 5{,}1$
4) L'imputation des valeurs sous H_0 donne le tableau suivant :

Figure 3.4 : The Electric Company : RP contrefactuels après imputation sous H_0

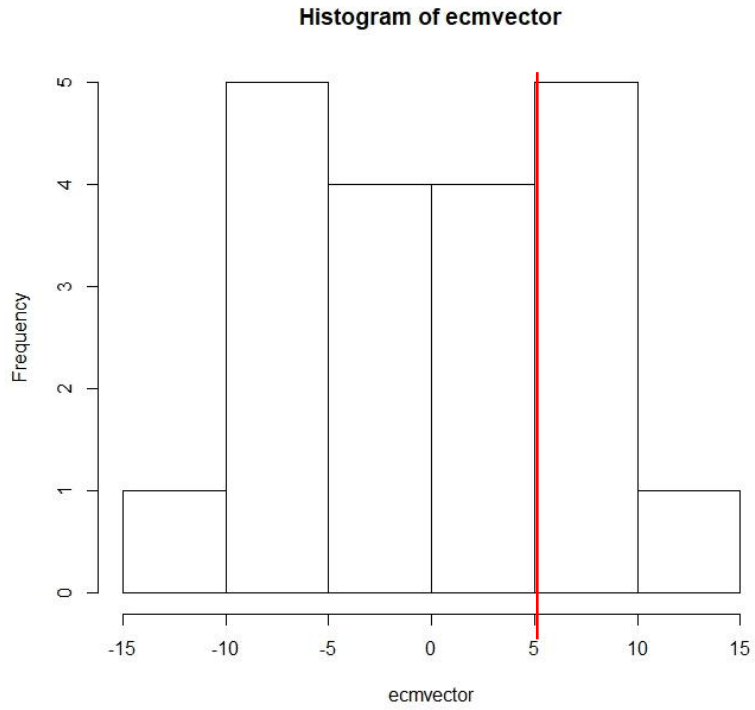
	unit	treatment	y	y0	y1
1	1	0 →	55.0	55.0	55.0
2	2	0 →	72.0	72.0	72.0
3	3	0 →	72.7	72.7	72.7
4	4	1 ←	70.0	70.0	70.0
5	5	1 ←	66.0	66.0	66.0
6	6	1 ←	78.9	78.9	78.9

[29] Pour Paul Rosenbaum (communication privée), H_0 est « sharp » si sous H_0 je peux connaître $Y_i(0)$ et $Y_i(1)$ à partir de Y_i, et donc déduire $Y_i(1) - Y_i(0)\ \forall i$. Par exemple, si $H_0 : Y_i(1) = Y_i(0)$, les RP sont tous les deux égaux à Y_i. Si $H_0 : Y_i(1) = Y_i(0) + \tau$, τ fixé à l'avance, alors si i: $D_i = 1$, $Y_i(1) = Y_i$ et $Y_i(0) = Y_i - \tau$, et si i: $D_i = 0$, $Y_i(0) = Y_i$ et $Y_i(1) = Y_i + \tau$.

5) Il y a $C_6^3 = 20$ vecteurs d'affectation possibles avec trois classes tests sur six classes. Pour chacune, on recalcule $\bar{y}_1 - \bar{y}_0$, qui est la différence de moyennes qui aurait été observée pour chaque vecteur de traitement (sous H_0). Dans le cas présent il est possible de montrer tout ça à partir d'un tableau. Visualisons le tableau dans le fichier **`mat20.pdf`**.

6) Il s'agit maintenant de prendre une décision. Sous H_0, la probabilité d'observer une différence de moyenne égale à celle de notre échantillon, 5,1, ou plus grande, est de 6/20, qui vaut exactement 0,3, c'est la p-valeur. On peut localiser la statistique estimée dans la **distribution de randomisation** de la statistique de test (une différence de moyennes). Le show n'a pas d'effet.

Graphique 3.5 : Distribution de randomisation (The Electric Company)
$N_1 = N_0 = 3$, Nb. vecteurs $C_6^3 = 20$.



Nous pouvons reprendre cet exercice en remplaçant H_0 par l'hypothèse plus générale d'un **effet additif**, $H_0: Y_i(1) = Y_i(0) + \tau$, avec $\tau \in \{-5\ ;\ +5\}$ (exercice 3.5.1).

3.4.2. Test de Neyman

On sait (exercice 3.5.5) que l'EMT $\bar{Y}_1 - \bar{Y}_0$ est un estimateur sans biais de l'ECM. À partir de l'EMT, on peut bâtir un intervalle de confiance. Contrairement au test exact de Fisher, il nous faut une estimation sans biais de la variance de l'EMT. D'après le théorème 6.3 d'Imbens et Rubin (2015, 92), sous la condition d'un effet additif, on a l'estimateur de la variance suivant :

$$\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}, \tag{3.2}$$

où $s_d^2 = \frac{1}{N_d - 1} \sum_{i:D_i=d} (Y_i - \bar{Y}_d)^2$, pour $d = 0{,}1$, qui dépend des variances de Y_i dans chaque groupe de traitement. L'estimateur (3.2) est une approximation de la variance de l'EMT. Il y a un 3e terme à estimer qu'il faudrait soustraire à (3.2) :

$$\frac{2}{N-1} \sum_i (\mathrm{ECI}_i - \mathrm{ECM})^2.$$

On voit que si l'effet est additif, alors $\text{ECI}_i - \text{ECM} = 0$, le 3e terme s'annule. L'estimateur est sans biais. L'estimateur (3.2) est alors la formule d'un test d'égalité des moyennes de populations indépendantes, avec variances inégales et inconnues, mais sans correction complexe des degrés de liberté que l'on trouve pour ce type de test en général (Newbold *et alii*, 2007, 378-379). Imbens et Kolesar (2016) font le point sur ce type de correction.

La construction d'un intervalle de confiance (IC) repose sur un seuil de significativité théorique α, contrairement au test exact de Fisher. L'IC au seuil α est :

$$\bar{Y}_1 - \bar{Y}_0 \pm z_{\alpha/2} \times \left(\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}\right)^{1/2},$$

où $z_{\alpha/2}$ est tel que $\Pr\left(Z \leq z_{\alpha/2}\right) = 1 - \alpha/2$, avec $Z \sim N(0,1)$. Pour un test au seuil $\alpha = 5$ %, alors $\alpha/2 = 2{,}5$ % et $z_{\alpha/2} = 1{,}96$ d'après la table de la fonction de répartition de Z.

Les résultats ci-dessous détaillent les calculs à la main pour l'obtention de l'IC de Neyman, qui sont faciles à implémenter sous **R** ou **Stata** : **neymantest_pairwise.do**

i	$Y(0)$	$Y(1)$	D	Y	$Y - \bar{Y}_d$	$(Y - \bar{Y}_d)^2$
1	55,0		0	55,0	−11,6	133,79
2	72,0		0	72,0	5,4	29,52
3	72,7		0	72,7	6,1	37,62
4		70,0	1	70,0	−1,6	2,67
5		66,0	1	66,0	−5,6	31,73
6		78,9	1	78,9	7,3	52,80

$\bar{Y}_1$	71,6
$\bar{Y}_0$	66,6
EMT	5,1
$\sum_{i:D_i=1}(Y_i - \bar{Y}_1)$	87,2
$\sum_{i:D_i=1}(Y_i - \bar{Y}_1)$	200,9
S_1^2	43,6
S_0^2	100,5
$\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$	48,022
Ecart-type	6,93
Borne inf	**-8,51**
Borne sup	**18,64**

L'intervalle de confiance à la Neyman $(-8{,}51\ ;\ 18{,}64)$ nous fait aboutir à la même conclusion qu'avec la *p-value* du test exact de Fisher : on ne rejette pas.

Ce n'est pas cohérent d'utiliser la valeur critique $z_{0{,}025}$ avec N petit. L'IC serait plus approprié pour encadrer l'effet sur le groupe témoin, $N = 34$) (code **fisherexacttest_electriccompany.R**), mais pour Imbens et Rubin (2015, 96), il est difficile de savoir quelle est la loi suit suivie par la valeur critique si on ne connaît pas la distribution jointe des RP : $j(y(0), y(1))$.

Neyman a montré que pour une région de confiance de 95%, l'IC est plus étendu que nécessaire, quand le vrai effet causal est additif et N grand. On peut donc le réduire un peu en prenant un seuil de 10 % ($z_{0{,}05} = 1{,}65$) ; **qnorm(0.05)** dans **R** $(-1{,}64)$. L'idée est que si le 3e terme n'est pas nul, la variance estimée est supérieure à sa valeur théorique. On resserre L'IC en prenant un seuil de 10 %.

Revenons à la base complète et tenons compte du protocole randomisé par paires décrit au début de la sous-section 3.4.1 (Imbens et Rubin, 2015, 219-239).

On a 96 écoles = paires ($N = 192$ classes en tout). Notons G_i l'école à laquelle appartient la classe i. Les formules pour écrire l'EMT dans chaque école, et au niveau agrégé, sont, respectivement :

$\hat{\tau}_{dif}^{pai}(j) = \sum_{i:G_i=j}(2D_i - 1)Y_i$, avec $j = 1, \dots, N/2$ (96) écoles, et
$\hat{\tau}_{dif} = \frac{1}{N/2}\sum_j \hat{\tau}_{dif}^{pai}(j)$.

Nous trouvons $\hat{\tau}_{dif} = 5{,}65$, qui est exactement la moyenne $\bar{Y}_1 - \bar{Y}_0$ puisque dans un appariement par paires, la moyenne des différences est la différence des moyennes. Concernant la variance, c'est plus compliqué. L'estimateur proposé par Imbens et Rubin (2015, 226-227) est :

$$\frac{1}{\frac{N}{2}\left(\frac{N}{2}-1\right)}\sum_{j=1}^{N/2}\left(\hat{\tau}_{dif}^{pai}(j) - \hat{\tau}_{dif}\right)^2.$$

Nous trouvons pour l'estimation, 1,10. Cet estimateur suppose que l'ECI_i est constant, et additif, pour chaque paire et d'une paire à l'autre. On construit ensuite un intervalle de confiance avec 1,96 pour valeur critique : [3,59 ; 7,72].

L'IC s'appuyant sur la variance calculée à la Neyman est plus simple à calculer : [0,684 ; 10,629], mais légèrement plus large, comme nous l'avions évoqué plus haut. En prenant un seuil de 10 %, on obtient [1,471 ; 9,843].

3.4.3. ANOVA (Analyse de la Variance)

Avec la régression, c'est une méthode très utilisée dans le cadre d'expérimentations randomisées. On n'a pas besoin de se placer dans le cadre du MCR avec les $Y(0)$ et $Y(1)$, mais dans celui d'une régression. Dans l'exemple The Electric Company, sans variable explicative, et deux traitements, nous pouvons utiliser la commande **oneway**, pour dire que nous avons un ANOVA à un facteur (le traitement).

Figure 3.5 : ANOVA à un facteur (The Electric Company)

```
. oneway Y D

                    Analysis of variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1536.23755       1   1536.23755      4.97     0.0269
 Within groups      58699.7024     190   308.945802
------------------------------------------------------------------------
    Total           60235.9399     191   315.371413
```

Puis un IC avec **`regress`**, où l'on estime une constante, ce qui change le calcul des erreurs standards, mais est proche de Neyman (pas pairwise).

L'IC dans la régression est celui d'un **`ttest Y, by(D)`** avec variances égales (les ddl sont plus simples que sous l'hypothèse de variances inégales).

3.5. Exercices sur le chapitre 3

3.5.1) Reprenez le tableau de données de l'expérimentation « The Electric Company » de la section 3.4 et tester l'hypothèse d'un effet (causal) additif égal à -5. Pour cela, vous devrez modifier deux lignes du programme **`fisherexacttest_electriccompany.R`**. Rejetez-vous l'hypothèse?

3.5.2) Montrer que dans une EX avec un nombre d'individus noté N, où le nombre d'individus qui reçoivent le traitement actif est N_1 et le nombre de ceux qui reçoivent le traitement de contrôle $N_0 := N - N_1$, la probabilité qu'un individu i soit traité vaut N_1/N.

3.5.3) (Exercice élaboré à partir de « Entrée à l'université : la casse-tête de l'été », Le Monde, 18 juillet 2017, p. 8, et de la circulaire N° 2017-077 du 24-4-2017 du MENESR). 86969 étudiants n'avaient pas encore reçu de proposition d'admission au 14 juillet 2017. Est-ce étonnant ? Il y avait environ 50000 étudiants de plus en 2016 qu'en 2015 ! Le tirage au sort pratiqué à l'entrée de l'université pour départager des candidats peut être vécu comme une injustice par certains. D'après la loi, lorsque le total des candidats à une formation, Economie en L1 par exemple, ayant obtenu le BAC et résidant dans une académie excède les capacités d'accueil, la plateforme APB effectue un classement en tenant compte de l'académie du candidat, de l'ordre de ses vœux, de sa situation de famille et, en dernier recours, d'un tirage au sort. Appelons M0 le mécanisme du tirage au sort et M1 un autre mécanisme par lequel on pioche des paires successives d'étudiants dans la liste des *ex aequo* et inscrit l'étudiant qui a eu la meilleure note des deux au BAC. Vous devez réaliser une étude sur l'inscription des étudiants à l'Université (en fonction par ex. de la décision d'avoir suivi des options ou pas).

a) Quel mécanisme biaiserait le moins votre évaluation ?
b) Expliquez.

3.5.4) (Exercice élaboré à partir d'une expérimentation racontée dans Chevassus-au-Louis, N. (2016), Malsciences, Seuil) David Simmons-Duffin, de l'Institute for Advanced Studies de Princeton, a inventé un programme en ligne qui génère des titres d'articles de physique, http://snarxiv.org/vs-arxiv/. Le jeu consiste à choisir entre deux titres d'articles lequel est une authentique étude déposée sur ArXiv (un site de référence dans le domaine) et lequel est une création aléatoire de son programme. Sur 750000 internautes qui se sont essayés au jeu, le taux d'identification des articles authentiques n'est que de 59 %, soit à peine mieux que la performance que l'on attendrait de quelqu'un qui choisirait en jetant une pièce : 50 % en répondant au hasard. Quant aux différents spécialistes de physique théorique interrogés par l'auteur de l'ouvrage, ils obtiennent un score de 80 %.

a) En vous appuyant sur l'EX de la buveuse de thé, que pensez-vous de ce protocole ?
b) Quel est le seul de signification dans cette EX ?
c) Construire un IC pour le taux d'identification. Rejetteriez-vous l'hypothèse nulle que le taux vaut 57 % ?
d) Sans calculer, la p-valeur est-elle inférieure au seuil de la question b) ?

3.5.5) Il s'agit d'un exercice d'inférence autour de l'estimateur de Neyman de l'ECM (Imbens et Rubin, 2015, 105-112) dans le cas d'un MAT pleinement randomisé. On rappelle que l'ECM dans l'échantillon est égal à $\overline{Y(1) - Y(0)}$, que Imbens et Rubin (2015) notent τ_{fs}, 'fs' pour *finite sample*. On note l'estimateur de différences des moyennes, $\bar{Y}_1 - \bar{Y}_0 := \hat{\tau}_{\text{dif}}$. Montrer que $\hat{\tau}_{\text{dif}}$ est un estimateur sans biais de τ_{fs}.

3.5.6) Supposons que la population soit de très grande taille et le MAT ignorable, $E(Y(D)|D) = E(Y(D))$. Montrer que l'estimateur des différences des moyennes dans la population $E(Y|D = 1) - E(Y|D = 0)$ est sans biais.

3.5.7) (Exercice de révision du chapitre 2) Le collège Niki de Saint Phalle à Valbonne, souhaite évaluer une nouvelle méthode d'enseignement de l'anglais qui sera introduite avec la réforme des collèges en septembre 2016. Un an avant, à la rentrée 2015, un échantillon de deux classes de 6e fut tiré pour cette évaluation. Les nombres d'élèves de chaque classe sont, respectivement, 36 et 34. Dans chaque classe, des élèves ont été tirés au hasard pour suivre l'enseignement d'anglais en vigueur, tandis que les autres sont affectés à l'enseignement réformé ; en tout, 35 élèves ont été affectés à chaque type d'enseignement. À la fin de l'année, chacun passe un examen oral et obtient une note. Pour chaque groupe de chaque classe on compte le nombre d'élèves qui ont obtenu 7/10 ou plus à l'examen. On obtient le tableau suivant :

	Enseignement réformé	Enseignement classique
Classe 1	$\frac{8}{9}$	$\frac{23}{27}$
Classe 2	$\frac{19}{26}$	$\frac{5}{8}$
École	$\frac{27}{35}$	$\frac{28}{35}$

a) À la lecture des résultats au niveau de l'école, recommanderiez-vous la réforme de l'enseignement d'anglais ?
b) À la lecture des résultats au niveau de chaque classe, recommanderiez-vous cette fois-ci cette réforme ?
c) Comment appelle-t-on la contradiction dans vos réponses aux questions a) et b) ?
d) Après avoir fait une enquête socio-économique auprès des élèves et de leurs parents, vous observez que dans la classe 1, il y a un peu plus d'enfants de parents anglophones que dans la classe 2. Utilisez cette information supplémentaire pour expliquer votre réponse à la question c).

3.5.8) (Exercice difficile, à faire à la maison) Vérifier que dans le cas de trois traitements, le résultat observé dans l'équation de Rubin peut s'écrire

$$Y = Y(0)\tfrac{1}{2}(D-1)(D-2) - Y(1)D(D-2) + Y(2)\tfrac{1}{2}D(D-1),$$

où $Y(0)$, $Y(1)$ et $Y(2)$ sont les résultats potentiels associés aux traitements D égal à 0, à 1 ou à 2. Et, supposons un ECI constant du traitement 1 par rapport au traitement 0, et du traitement 2 par rapport au traitement 1, égal à ρ. Montrez que Y dans l'équation de Rubin peut s'écrire

$$Y = Y(0) + \rho D.$$

3.6. Annexe (estimateur de Neyman de la variance)

La variance de l'estimateur de Neyman est une approximation du 'vrai' estimateur sous l'hypothèse nulle d'un effet constant. En effet, l'estimateur complet de la variance est

$$\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0} - \frac{s_{10}^2}{N_1 + N_0}$$

Comme pour la variance d'une moyenne, ou de la variance d'une différence de moyennes, il suffit de calculer la variance de $\bar{Y}_1 - \bar{Y}_0 \overset{\text{def}}{=} \hat{\tau}_{\text{dif}}$. Le calcul donne comme résultat :

$$\mathrm{V}_D(\hat{\tau}_{\text{dif}}) = \frac{1}{N_1}\frac{1}{N-1}\sum_i \left(Y_i(1) - \overline{Y(1)}\right)^2 + \frac{1}{N_0}\frac{1}{N-1}\sum_i \left(Y_i(0) - \overline{Y(0)}\right)^2 + \frac{1}{N}\frac{1}{N-1}\sum_i (\mathrm{ECI}_i - \mathrm{ECM})^2.$$

Autrement dit, la variance de $\hat{\tau}_{\text{dif}}$ est égale à la somme des variances de $Y_i(1)$, de $Y_i(0)$ et de l'effet causal individuel. Si les ECI_i sont égaux à une valeur inconnue, τ par exemple, alors $\mathrm{ECM} = \mathrm{ECI}_i$. Il est alors naturel de considérer l'estimateur suivant : $s_1^2/N_1 + s_0^2/N_0$.

La démonstration est un peu longue (Imbens et Rubin, 2015, 105-107) : à finir

$$\begin{aligned}\mathrm{V}_D(\hat{\tau}_{\text{dif}}) &= \mathrm{V}_D(\bar{Y}_1 - \bar{Y}_0) \\ &= \mathrm{V}_D(N_1^{-1}\textstyle\sum_{i:D_i=1} Y_i - N_0^{-1}\sum_{i:D_i=0} Y_i) \\ &= \mathrm{V}_D\left(N_1^{-1}\textstyle\sum_{i:D_i=1} Y_i\right) + \mathrm{V}_D\left(N_0^{-1}\textstyle\sum_{i:D_i=0} Y_i\right) - \\ &2\mathrm{Cov}_D(N_1^{-1}\textstyle\sum_{i:D_i=1} Y_i\,, N_0^{-1}\sum_{i:D_i=0} Y_i)\end{aligned}$$

À partir d'ici, on utilise le fait que Y_i avec $i{:}\,D_i = 1$ vaut $Y_i(1)$ et $Y_i(0)$ pour $i{:}\,D_i = 0$, puis l'astuce qui consiste à remplacer $\sum_{i:D_i=1} Y_i(1)$ par $\sum_i D_i Y_i(1)$ et $\sum_{i:D_i=0} Y_i(0)$ par $\sum_i (1 - D_i) Y_i(0)$, où i va de 1 à N. Rappelons également que les $Y_i(d)$, $d \in \{0; 1\}$ sont des quantités fixes. Enfin, notons que $\mathrm{V}_D(D_i) = \mathrm{V}_D(1 - D_i) = N_1 N_0 N^{-2}$, et $\mathrm{Cov}_D\left(D_i, 1 - D_j\right) = -\mathrm{Cov}_D\left(D_i, D_j\right)$, dont la valeur dépend du fait que $i = j$ ou pas. Si $i = j$, $-\mathrm{Cov}_D\left(D_i, D_j\right) = -N_1 N_0 N^{-2}$. Mais si $i \neq j$, elle vaut $= N_1 N_0 N^{-2}(N-1)^{-1}$. En effet, $-\mathrm{Cov}_D\left(D_i, D_j\right)$ vaut

$$\begin{aligned}-\mathrm{E}_D\left(D_i D_j\right) + \mathrm{E}_D(D_i)\mathrm{E}_D\left(D_j\right) &= -\Pr(D_i = 1, D_j = 1) + N_1^2 N^{-2} \\ &= -\Pr\left(D_i = 1 \middle| D_j = 1\right)\Pr(D_j = 1) + N_1^2 N^{-2} \\ &= -\frac{N_1 - 1}{N-1}\frac{N_1}{N} + \frac{N_1^2}{N^2} \\ &= \frac{N_1 N_0}{N^2(N-1)}.\end{aligned}$$

Enfin, $\mathrm{Cov}_D\left(1 - D_i, 1 - D_j\right) = \mathrm{Cov}_D\left(D_i, D_j\right)$. Reprenons chacun des termes de $\mathrm{V}_D(\hat{\tau}_{\text{dif}})$.

$$\begin{aligned}\mathrm{V}_D\left(N_1^{-1}\textstyle\sum_{i:D_i=1} Y_i\right) &= N_1^{-2}\mathrm{V}_D(\textstyle\sum_i D_i\, Y_i(1)) \\ &= N_1^{-2}\textstyle\sum_i \mathrm{V}_D(D_i)\, Y_i^2(1) + 2\sum_i \sum_{j\neq i} Y_i(1) Y_j(1)\mathrm{Cov}_D(D_i, D_j) \\ &= N_1^{-1} N_0 N^{-2}\textstyle\sum_i Y_i^2(1) - 2N_1 N_0 N^{-2}(N-1)^{-1}\sum_i \sum_{j\neq i} Y_i(1) Y_j(1)\end{aligned}$$

$$\begin{aligned}\mathrm{V}_D\left(N_0^{-1}\textstyle\sum_{i:D_i=0} Y_i\right) &= N_0^{-2}\mathrm{V}_D(\textstyle\sum_i (1 - D_i)\, Y_i(0)) \\ &= N_0^{-2}\textstyle\sum_i \mathrm{V}_D(D_i)\, Y_i^2(0) + 2\sum_i \sum_{j\neq i} Y_i(0) Y_j(0)\mathrm{Cov}_D(D_i, D_j)) \\ &= N_0^{-1} N_1 N^{-2}\textstyle\sum_i Y_i^2(0) - 2N_1 N_0 N^{-2}(N-1)^{-1}\sum_i \sum_{j\neq i} Y_i(0) Y_j(0)\end{aligned}$$

$-2\mathrm{Cov}_D\left(N_1^{-1}\sum_{i:D_i=1} Y_i\,, N_0^{-1}\sum_{i:D_i=0} Y_i\right) = -2N_1^{-1}N_0^{-1}\mathrm{Cov}_D\left(\sum_i D_i\,Y_i(1), \sum_i(1-D_i)\,Y_i(0)\right)$,

où $\mathrm{Cov}_D\left(\sum_i D_i\,Y_i(1), \sum_i(1-D_i)\,Y_i(0)\right)$ est la somme de deux termes, $\sum_i Y_i(1)Y_i(0)\mathrm{Cov}_D(D_i, 1-D_i)$ et $\sum_i\sum_{j\neq i} Y_i(1)Y_j(0)\mathrm{Cov}_D\left(D_i, 1-D_j\right)$, qui valent $N_1N_0N^{-2}(N-1)^{-1}\sum_i Y_i(1)Y_i(0)$ et $N_1N_0N^{-2}(N-1)^{-1}\sum_i\sum_{j\neq i} Y_i(1)Y_j(0)$. Donc, $-2\mathrm{Cov}_D\left(N_1^{-1}\sum_{i:D_i=1} Y_i\,, N_0^{-1}\sum_{i:D_i=0} Y_i\right)$ est égal à $-2N_1^{-1}N_0^{-1}[N_1N_0N^{-2}(N-1)^{-1}\sum_i Y_i(1)Y_i(0) + N_1N_0N^{-2}(N-1)^{-1}\sum_i\sum_{j\neq i} Y_i(1)Y_j(0)]$ qui, après simplification, devient :

$$-2N^{-2}(N-1)^{-1}[\sum_i Y_i(1)Y_i(0) + \sum_i\sum_{j\neq i} Y_i(1)Y_j(0)].$$

Or, $-2N^{-2}(N-1)^{-1}[\sum_i Y_i(1)Y_i(0) + \sum_i\sum_{j\neq i} Y_i(1)Y_j(0)]$ est égal à :

$$-2N^{-1}(N-1)^{-2}[\sum_i(Y_i(1)-\overline{Y(1)})(Y_i(0)-\overline{Y(0)}) + \sum_i\sum_{j\neq i} Y_i(1)Y_j(0)].$$

Ainsi, si on ajoute tout, nous avons :

$N_1^{-1}N_0N^{-2}\sum_i Y_i^2(1) - 2N_1N_0N^{-2}(N-1)^{-1}\sum_i\sum_{j\neq i} Y_i(1)Y_j(1) +$

$N_0^{-1}N_1N^{-2}\sum_i Y_i^2(0) - 2N_1N_0N^{-2}(N-1)^{-1}\sum_i\sum_{j\neq i} Y_i(0)Y_j(0) +$

$-2N^{-1}(N-1)^{-2}[\sum_i(Y_i(1)-\overline{Y(1)})(Y_i(0)-\overline{Y(0)}) + \sum_i\sum_{j\neq i} Y_i(1)Y_j(0)].$

Corrections des exercices du chapitre 3

3.5.1) Les modifications à faire au programme sont les suivantes :

```
elec$y1 <- ifelse(elec$treatment==0, elec$y0-5, elec$y1)
elec$y0 <- ifelse(elec$treatment==1, elec$y1+5, elec$y0)
```

La p-valeur vaut 0,15.

3.5.2) La probabilité qu'un individu soit traité est la probabilité qu'il appartienne au groupe des N_1 individus parmi N qui reçoivent le traitement actif ; N_0 reçoivent le contrôle (évidemment, si $N_1 = N$, l'individu a 100% de chance d'être traité, mais il n'y aurait pas de groupe de contrôle). Au dénominateur nous avons donc le nombre de combinaisons avec N_1 individus traités parmi N : $\frac{N!}{N_1!N_0!}$.

Le calcul du numérateur suit la même logique et répond à la question suivante : dans un échantillon de N individus, où N_1 sont traités, combien il y a-t-il de combinaisons où l'individu i particulier reçoit toujours le traitement actif ? Puisque cet individu est traité, il fait partie des N_1. Je peux donc calculer le nombre de combinaisons sans lui. Pour cela, je le retire de N_1 et donc de N, de sorte qu'il me reste $N_1 - 1$ individus à affecter au traitement actif sur $N - 1$ et toujours N_0 qui ne le sont pas. Le numérateur vaut donc :

$$\frac{(N-1)!}{(N_1-1)!N_0!}.$$

Il suffit pour finir de diviser le numérateur par le dénominateur :

$$\frac{\frac{(N-1)!}{(N_1-1)!N_0!}}{\frac{N!}{N_1!N_0!}} = \frac{(N-1)!}{(N_1-1)!N_0!}\frac{N_1!N_0!}{N!} = \frac{(N-1)!}{(N_1-1)!}\frac{N_1!}{N!} = \frac{N_1}{N}.$$

3.5.3)
a) M0.
b) Car à aucun moment M0 fait dépendre la sélection du résultat potentiel. Ce n'est pas le cas de M1. En sélectionnant dans chaque paire l'étudiant/e qui a eu la meilleure note au BAC c'est bien parce que l'on pense qu'il va mieux réussir ses études sup. L'inscription en études supérieures des étudiants qui ont pris des options devrait alors être relativement supérieure que si c'est M0 qui est appliqué.

3.5.4)
a) Ce protocole est insuffisant, à moins de répondre au moins 4 fois, et en supposant que les articles sont complètement randomisés. Dans le cas d'une réponse, un internaute qui ni connaît rien, cliquerait au même endroit si les titres des deux articles avaient été permutés.
b) 50%. Ici, seuil de signification et p-valeur sont la même chose sous l'hypothèse nulle d'un internaute qui n'y connaît rien, car il s'agit du test exact de Fisher.
c)
$0{,}59 - 1{,}96 \times \sqrt{0{,}59(1-0{,}59)/750000} \leq p \leq 0{,}59 + 1{,}96 \times \sqrt{0{,}59(1-0{,}59)/750000}$.
$0{,}589 \leq p \leq 0{,}591$. Oui, on rejette.
d) La statistique de test vaut 158,47, obtenu en calculant $(0{,}59 - 0{,}50)/(0{,}59(1-0{,}59)/750000)^{1/2}$. Elle suit une loi normale centrée réduite. Or, elle est très grande. La p-valeur associée à cette statistique est donc très petite.

3.5.5) Il y a différentes manières de le démontrer. Une manière simple consiste à supposer des RP fixes et que les RO sont aléatoires à cause du MAT qui lui-même est aléatoire. Le RO Y est, pour chaque individu, relié aux RP par l'équation de Rubin $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$. Nous notons également, comme dans le cours, $\bar{Y}_1 \equiv {N_1}^{-1} \sum_{i:D_i=1} Y_i$ et $\bar{Y}_0 \equiv {N_0}^{-1} \sum_{i:D_i=0} Y_i$. Les espérances mathématiques sont prises par rapport aux variables aléatoires D_i. On rappelle que $E(D_i) = 0 \times \Pr(D_i = 0) + 1 \times \Pr(D_i = 1) = N_1 N^{-1}$, et $1 - E(D_i) = N_0 N^{-1}$ dans le cas d'une affectation aléatoire (voir exercice 3.5.2) pour le MAT pleinement randomisé. Calculons $E_D(\hat{\tau}_{\text{dif}})$:

L'opérateur d'espérance mathématique est linéaire, donc
$= \mathrm{E}(\bar{Y}_1) - E(\bar{Y}_0)$.
$= E({N_1}^{-1} \sum_{i:D_i=1} Y_i) - E({N_0}^{-1} \sum_{i:D_i=0} Y_i)$.

D'après l'équation de Rubin, nous pouvons écrire
$= E({N_1}^{-1} \sum_{i:D_i=1} Y_i(1)) - E({N_0}^{-1} \sum_{i:D_i=0} Y_i(0))$.

L'astuce consiste à faire intervenir les D_i en prenant les sommes de 1 à N :
$= E\left({N_1}^{-1} \sum_i D_i Y_i(1)\right) - E\left({N_0}^{-1} \sum_i (1 - D_i) Y_i(0)\right)$
$= \sum_i {N_1}^{-1} E(D_i) Y_i(1) - \sum_i {N_0}^{-1} \left(1 - E(D_i)\right) Y_i(0)$
$= \sum_i {N_1}^{-1} N_1 N^{-1} Y_i(1) - \sum_i {N_0}^{-1} N_0 N^{-1} Y_i(0)$
$= N^{-1} \sum_i Y_i(1) - N^{-1} \sum_i Y_i(0)$
$= N^{-1} \sum_i (Y_i(1) - Y_i(0))$
$= \tau$.

On va ajouter une implication importante, sachant que $\beta_2^{MC} = \bar{Y}_1 - \bar{Y}_0$, dans le modèle de régression linéaire, alors $E_D(\beta_2^{MC}) = \tau$.

3.5.6) Renvoyé au chapitre 4.

3.5.7) Pensez à écrire les fréquences relatives en pourcentage dans le tableau de l'énoncé : $8/9 = 88\%$, $19/26 = 73\%$, $23/27 = 85\%$, $5/8 = 62\%$.

a) Non, car les résultats au niveau de l'école (les résultats agrégés de la dernière ligne donc) montrent qu'avec l'enseignement classique de l'anglais, un élève de plus (28) a obtenu au moins 7/10.
b) Oui. Il suffit de comparer les résultats de l'enseignement réformé relativement à l'enseignement classique : $88\% > 85\%$ et $73\% > 62\%$.
c) Le paradoxe de Simpson.
d) Nous pouvons expliquer le paradoxe en combinant deux informations :
- être dans la classe 2 c'est avoir plus d'élèves traités avec l'examen réformé ;
- être dans la classe 2 c'est avoir plus d'enfants de parents non-anglophones.
Donc plus d'élèves de parents non anglophones ont passé l'examen réformé. La classe est un facteur commun, de confusion, qui, lorsqu'il change de valeur entraîne un changement important dans le nombre d'élèves qui ont la réforme et dans la langue maternelle des parents. Cette corrélation entre nombre d'élèves et langue se produit au niveau agrégé car j'utilise les deux lignes du tableau. Si je me place sur une ligne, j'empêche cette corrélation de se produire.

3.5.8) Correction à finir.

4. Les études observationnelles

Nous allons voir les avantages et les inconvénients des études observationnelles (EO) dans la section 4.1, après avoir situé ce type d'études relativement aux types d'EX vus dans le chapitre 2. Puisqu'il s'agit avant tout de régler les problèmes d'évaluation inhérents à ce type d'études, nous insisterons sur les inconvénients dans la section 4.2. Il est courant dans la littérature d'évoquer les biais de sélection qui se trament dans ce type d'études, qui sont généralement révélés par des situations de déséquilibre et non-recouvrement. La section 4.3 introduira une supposition, l'indépendance conditionnelle, que doivent satisfaire les EO avant d'estimer l'effet causal de la PP.

4.1. Inconvénients et avantages des études observationnelles

Dans le chapitre 2, **étude observationnelle** (EO) désignait trois types d'expérimentations : EXT, EXN et EXS. William Gemmell Cochran, qui fut le directeur de thèse de Rubin, définit *observational studies* à propos d'études portant sur la relation tabac-cancer (Rosenbaum, 2010, 72-73). Il définit l'EO à peu près ainsi (*Ibidem*, pp. vii, 1) :

Définition 4.1 : une EO est une étude ayant pour objet de répondre à une question causale, mais dans laquelle l'évaluateur contrôle imparfaitement, ou pas du tout, quel individu reçoit quel traitement.

En rapport avec la définition d'un MAT vue dans le chapitre 3, la forme fonctionnelle f est inconnue (Imbens et Rubin, 2015, 41). L'évaluateur sait néanmoins quelles variables rentrent dedans.

Par exemple, dans une EXT, la probabilité de recevoir tel ou tel traitement est mal contrôlée, parce que des individus tirés au sort ne veulent plus participer à l'EX, des individus veulent changer de groupe, ou bien, des individus sortent avant la fin de l'EX (**attrition**).

Dans une EXS, l'expérimentateur ne contrôle rien.

4.1.1. Les inconvénients des EO

On glisse ici de l'idée d'affectation (chapitre 3) à celle de sélection

Premier inconvénient : le **biais de sélection** (BS, *selection bias*). Les individus qui ont recours à un dispositif le font en partie parce qu'ils pensent en avoir plus besoin que ceux qui n'ont pas recours. On ne doit pas s'étonner de trouver un effet positif du dispositif. Ou bien, ce sont ceux qui comprennent le mieux son fonctionnement.

Le concept de BS remonte aux années 1950 dans les sciences médicales, et les années 1970 en économie. Heckman (1979) corrige le BS dans le cas particulier d'**échantillon tronqué**, une situation que nous étudierons dans le chapitre 9. Rosenbaum et Rubin (1983) corrigent également le BS. Tous ces auteurs parlent également de **sélection non-aléatoire** (non-contrôlée) des individus (Rosenbaum, 2010, 86 ; Imbens et Wooldridge, 2009, 6) au sens où ils s'**auto-sélectionnent** dans les groupes de traitement.

Pour comprendre la différence entre les approches de Heckman d'une part et de Rubin d'autre part, prenons le cas de la relation entre salaire (Y), formation (D), et participation au marché du travail (S). Des individus sans emploi ont suivi une formation professionnelle ($D = 1$) pendant que d'autres non ($D = 0$). Les deux approchent supposent que les personnes qui suivent la formation ont des caractéristiques (X) différentes de celles qui ne l'ont pas suivie, conduisant à une estimation biaisée de l'effet $D \rightarrow Y$. Heckman suppose en plus que des individus n'ont pas trouvé de travail ($S = 0$).

Avant d'introduire le deuxième inconvénient, soulignons que dans une EO, les deux groupes de traitement sont des échantillons aléatoires, chacun représentatif de sa population ; voir Dehejia et Wahba (2002, 152). Autrement dit, les facteurs de confusion dans les groupes peuvent être différents et ne pas être distribuées pareil. Or, nous n'avons pas d'autre choix que de contrôler ces facteurs, qui sont reflétés dans les caractéristiques X, avec des méthodes qui sont propres aux MSE. Il y a cependant un problème.

Deuxième inconvénient : les facteurs de confusion ne sont pas tous contrôlables. Soit parce qu'ils ne sont pas observables ou qu'ils ne sont pas observés.

En théorie, un MATAC atténue le BS à la Rubin (Angrist et Pischke, 2009, 15) ; on décide à la place des individus dans quel groupe de traitement ils vont se trouver. Quant aux caractéristiques individuelles (âge, genre, etc.), leurs distributions dans chaque groupe tendent vers celle dans la population unique d'où sont tirés ces individus.

Une phrase résume assez bien ces inconvénients :

> *« In an experiment, treatment effects are seen clearly because the environment is tightly controlled, whereas in [an observational study], control is, to a large extent, replaced by choice – The environment is carefully chosen. »* ; Rosenbaum (2010, 333).

4.1.2. Les avantages relativement aux études randomisées

Moins onéreuses (la plupart d'entre elles). C'est le cas des EXS et EXN. En effet, il n'y a pas de protocole d'affectation des traitements, d'information et d'accompagnement des individus. Dans une EXN par ex., c'est la nature qui randomise (Zubizarreta, Small et Rosenbaum, 2014) comme on l'a vu dans un encadré du chapitre 2. Dans une EXS, ce n'est personne. Le coût résiduel, dans les deux (EXN, EXS), est lié au temps de mise en forme des données observationnelles, d'estimation et test de l'effet causal.

Les EXT peuvent être coûteuses en effet. Par exemple, TICELEC l'EX introduite dans le tableau 1.1 avait un budget de 190600 € courants en 2010, salaires des évaluateurs et évaluatrices inclus. L'**efficience** du projet (résultat financier/ménage) :

- 190600 € / 75 ménages = environ 2540 € / ménage ;
- économie d'énergie *a priori* (estimation du coût avant traitement) en €/ménage équipé : avec une réduction de la facture électrique de l'ordre de 7 %, l'économie pourrait être d'env. 48 € / ménage·an (pour une facture électrique moyenne de 687 € par ménage·an, chauffage électrique compris).[30]
- l'efficience annuelle : 0,019 / an << 1 (48 € / ménage·an) / (2540 € / ménage).

Le projet n'est pas efficient, et il faudrait 52,9 ans pour l'amortir : (2540 € / ménage) / (48 € / ménage·an). La PP du compteur vert *LINKY* fut financée par une répartition du budget sur un grand nombre de ménages (5 milliards d'€ répartis sur 35 millions de ménages, ce qui fait 142 € par ménage, amortis en 142/48 ≅ trois ans).

Plus éthiques. L'EXT sur les discriminations à l'embauche comme celle de Bertrand et Mullainathan (2004) part du principe qu'il n'est pas éthique de randomiser les individus (assigner des travailleur·euse·s au hasard à des entretiens). En revanche, vous pouvez envoyer des CV manipulés. *Idem* pour l'exposition au CIR ; les entreprises n'y ont pas recours au hasard !

Les épidémiologistes sont un peu moins contraints lorsqu'elles étudient la relation tabac-cancer par ex., en forçant des souris tirées au hasard à « fumer » dans un aquarium. Autre ex. : l'évaluation de la relation consommation d'aliments ultra-transformés-cancer ; Rosenbaum (2010) liste des études majeures en dehors de l'économie.[31]

---

[30]. D'après l'Insee, en 2006 (date proche de l'année de l'étude), la dépense électrique des ménages valait 17,92 Mds € pour 26069046 ménages à la fin de la même année (ined) ; ça fait un coût de 687€ par ménage.

[31]. Touvier, M., Srour, B., Hercberg, S., 2018. « La consommation d'aliments ultra-transformés est-elle liée à un risque de cancer ? », The Conversation, 15/0/2018.

4.2. Illustration et détection du biais de sélection

Nous allons d'abord illustrer une situation de BS en reprenant un exemple d'Angrist et Pischke (2009). Dans cet exemple, on n'observe pas de facteurs de confusion permettant de détecter un BS. Il s'agit de données issues d'une enquête américaine, pour savoir si des personnes plutôt âgées et précaires s'étant rendues aux urgences hospitalières se sentent en meilleure santé que celles qui n'y sont pas allées.

Dans une deuxième illustration, en économie de l'éducation, nous verrons comment détecter le BS à partir des concepts de **déséquilibre** et de **non-recouvrement** des groupes de traitement (Krueger, 1999).

On parle de déséquilibre (*imbalance*) quand au moins une caractéristique individuelle qui rentre dans le MAT n'est pas également distribuée entre les groupes de traitement (par ex., il y a significativement plus de jeunes dans le groupe test). Le non-recouvrement (*lack of overlap*) correspond à la situation où une ou plus des modalités d'une caractéristique sont absentes d'un groupe (par ex., il n'y a pas d'hommes entre 20 et 22 ans dans le groupe témoin).

Le MATAC évite ces problèmes. Il **équilibre** les caractéristiques individuelles, et le déséquilibre d'une caractéristique entre groupes diminue d'autant plus que le nombre d'individus augmente. Dans les chapitres qui suivent, nous verrons des méthodes d'appariement (contrôle) permettant d'atténuer le problème de BS dans les EO.

4.2.1. Biais de sélection

L'évaluation d'Angrist et Pischke est dans le domaine de la santé publique. Nous allons voir qu'à cause d'un potentiel BS (ici, l'offre de services d'urgence hospitalière), l'estimation de l'effet causal peut être très en-dessous de ce à quoi on pourrait s'attendre. Nous avons repris l'illustration des auteurs dans l'encadré 4.1 à la page suivante.

[charger **`urgence_angristpischke.R`** ou **`urgence_angristpischke.do`**]

Encadré 4.1 : Santé publique (urgences à l'hôpital)

Cette illustration d'Angrist et Pischke (2009, 12-15) en économie de la santé pose la question causale de savoir si l'hôpital améliore la santé des gens (« *Do hospitals make people healthier ?* »). C'est un problème de moyens, mais pas que … À partir des critère de classification d'une évaluation que nous avons vus dans le tableau 1.1, on a :

Individu :	personnes âgées, précaires
Traitement :	a rendu visite ($D_i = 1$), ou pas ($D_i = 0$), sur les 12 dernier mois, au service d'urgence d'un hôpital pour des premiers soins
Résultat :	santé/bien-être ; la variable traitée expérimentale est un niveau de santé déclaré par l'individu dans le National Health Interview Survey de 2005, sur une échelle de 1 (« mauvaise santé ») à 5 (« excellente santé »)
Traité :	les données ne contiennent pas d'information sur les soins reçus par les personnes

Calculons l'EMT $\bar{y}_1 - \bar{y}_0 = 3{,}21 - 3{,}93 = -0{,}72$. Cette différence est-elle significative ? Le traitement n'est pas randomisé et les population sont indépendantes. Nous faisons un test de Student de l'hypothèse $H_0 : \overline{Y(1)} = \overline{Y(0)}$:

```
> t.test([redacted]$healthnew~[redacted]group, mu=0, paired=F, var.equal=T)

        Two Sample t-test

data:  data$healthnew by data$group
t = -58.912, df = 97821, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7387162 -0.6911453
sample estimates:
mean in group Treated mean in group Control
             3.213275              3.928206
```

La valeur de la statistique conduit à rejeter H_0, mais son signe est contre-intuitif !

À la main, avec $N_0 = 90049$, $N_1 = 7774$, et $97823 - 2 = 97821$ le nombre de degrés de liberté du test. L'écart-type du niveau de santé chez les traités vaut 1,2549 et 1,0044 dans l'autre groupe. Les erreurs standards sont suffisamment petites pour que nous rejetions la nullité des moyennes ($1{,}2549/7774^{1/2} \cong 0{,}0142$ et $1{,}0044/90049^{1/2} \cong 0{,}0033$). La statistique utilisée sous R pour le t de Student est celle pour un test d'égalité de moyennes de deux populations indépendantes avec variances égales ; on obtient $t = -58{,}9$ (en supposant des variances inégales on a $t = -48{,}9$).*

$$t = \frac{\bar{y}_1 - \bar{y}_0}{\left(\frac{1}{N_0}+\frac{1}{N_1}\right)^{1/2}\left(\frac{(N_0-1)S_0^2+(N_1-1)S_1^2}{N_0+N_1-2}\right)^{1/2}} = \frac{3{,}21-3{,}93}{\left(\frac{1}{90049}+\frac{1}{7774}\right)^{1/2}\left(\frac{(90049-1)1{,}0044^2+(7774-1)1{,}2549^2}{90049+7774-2}\right)^{1/2}} \cong -58{,}9.$$

Or, $|t| = 58{,}9 > T_{0{,}05}(97821) \sim T_{0{,}05}(\infty)$ qui tend en distribution vers la valeur critique d'une N(0,1) au seuil de 5%, c'est-à-dire, 1,96.

* On teste l'égalité des variances avant de tester celle des moyennes. Les variances $1{,}2549^2$ et $1{,}0044^2$ étant proches, nous avons calculé la statistique de Fisher pour l'hypothèse nulle $\sigma_0^2 \leq \sigma_1^2$ contre $\sigma_0^2 > \sigma_1^2$. La statistique vaut $\frac{1{,}0044^2 \times 90049/(90049-1)}{1{,}2549^2 \times 7774/(7774-1)} \cong 0{,}64$. Elle est inférieure à la valeur critique tabulée $F(90048, 7773) \sim F(\infty, \infty) = 1$. On ne rejette pas H_0. Si sous l'hypothèse nulle, $\sigma_0^2 = \sigma_1^2 := \sigma^2$ était connu, nous calculerions $t = (\bar{y}_1 - \bar{y}_0)/[\sigma(n_0^{-1} + n_1^{-1})^{1/2}]$; voir Newbold, Carlson et Betty (2007).

Ce résultat suggère que les personnes âgées et précaires qui sont allées aux urgences durant ces 12 derniers mois (le groupe test) se sentent significativement en moins bonne santé que celles qui n'y sont pas allées (le groupe témoin). Faut-il pour autant conclure qu'aller à l'hôpital rend plus malade ? Bien sûr que non !

N'ayant aucune information sur les problèmes de santé des individus, spéculons sur les causes plausibles de ce résultat. Nous allons nous limiter à des explications simples, sous l'hypothèse d'hétérogénéité des groupes de traitement (sous-entendu, l'hétérogénéité entre groupes est probablement plus grande que celle dans chacun).

1) Les urgences sont pleines de microbes.
2) Des individus du groupe test mettent du temps à récupérer après être sortis des urgences (ils prennent des médicaments qui ont des effets indésirables).
3) Le groupe test est « structurellement » en moins bonne santé.

1 : Ok mais nous n'avons pas d'information sur les conditions d'hygiène aux urgences.
2 : Ok mais ne peut concerner qu'une portion du groupe test car les effets indésirables concernent entre 1 personne sur 10000 (très rare) et 1/100 (peu fréquent).
3 : Pose la question de la santé des individus tests s'ils n'avaient pas été traités.

La quantité non-observable $\sum_{i:D_i=1} Y_i(0)/N_1$ est un élément du BS en rapport avec le point **3**. La question est de savoir si $\sum_{i:D_i=1} Y_i(0)/N_1 - \sum_{i:D_i=0} Y_i(0)/N_0 < 0$. Nous pourrions répondre un peu si nous connaissions la santé des personnes avant la période des 12 derniers mois. Cette différence mesure le terme de BS, comme le voit dans l'encadré 4.2.

Ce terme apparaît dans l'EMT $\hat{\tau}^{\text{diff}} = \tau_1 + BS$. Il peut être positif, ou négatif (comme le suggère l'exemple « Santé publique ». En effet, si $BS < 0$, $\hat{\tau}^{\text{diff}}$ sous-estime l'ECMT (τ_1), avec un chance qu'on ait aussi $\hat{\tau}^{\text{diff}} < 0$.

Le BS dans une EO vient du fait que des individus choisissent plus ou moins le traitement auquel ils seront exposés. À moins que les traitements soient randomisés, l'auto-sélection dans les groupes de traitement (Wooldridge, 606- de la version de 2003) se fait toujours sur la base de caractéristiques personnelles (des variables observables ou pas). L'auto-sélection est supposé être corrélé aux RP et à ces variables. Par exemple, il peut être lié à l'ECI, $Y_i(1) - Y_i(0)$, comme dans le tableau théorique de la section 3.1 où nous avions supposé $D_i = 1(Y_i(1) - Y_i(0) > 0)$. Le problème pour l'évaluateur est qu'il ne connaît pas la forme fonctionnelle de ce lien.

Nous allons donc devoir faire des hypothèses qui puissent capturer les idées suivantes :

- Jusqu'au chapitre 6, nous allons supposer que le biais de sélection est fonction de facteurs de confusion X observables. Ça veut dire que l'information contenue dans X est suffisante pour déterminer statistiquement quel individu reçoit quel traitement (D serait donc corrélé à X).
- L'hypothèse cruciale que nous allons poser est que D est aussi corrélé à $(Y(1), Y(0))$. Mais, conditionnellement à X, D sera indépendant de $(Y(1), Y(0))$.
- Si l'on souhaite estimer l'ECMT, les facteurs X qui rentrent BS sont relatifs aux caractéristiques des individus tests et témoins dans l'état où ils ne seraient pas traités (Angrist et Pischke, 2015, 11) ; il n'y a que du $Y(0)$ dans BS dans ce cas.
- Déséquilibre (sous-section 4.2.2) et non-recouvrement (sous-section 4.2.3), puisqu'ils portent sur X, permettront d'apprécier l'ampleur du BS.

Encadré 4.2 : Isoler le biais de sélection

1) Au niveau d'une population $i = 1, \dots, N$, avec $Y(1)$ et $Y(0)$ fixes, et D aléatoire. Rappelons les formules de l'ECMT et de l'estimateur des différences de moyennes :

$\frac{1}{N_1}\sum_{i:D_i=1}\big(Y_i(1) - Y_i(0)\big)$, qui n'est pas observable, et que l'on note aussi τ_1, et
$\frac{1}{N_1}\sum_{i:D_i=1} Y_i - \frac{1}{N_0}\sum_{i:D_i=0} Y_i$, que l'on note $\bar{Y}_1 - \bar{Y}_0$ ou aussi $\hat{\tau}^{\text{diff}}$.

Ecrivons l'équation de Rubin ainsi : $Y_i = Y_i(0) + (Y_i(1) - Y_i(0))D_i$. Alors

$$\begin{aligned}\hat{\tau}^{\text{diff}} &= \tfrac{1}{N_1}\textstyle\sum_{i:D_i=1}(Y_i(0) + (Y_i(1) - Y_i(0))D_i) - \tfrac{1}{N_0}\sum_{i:D_i=0}(Y_i(0) + (Y_i(1) - Y_i(0))D_i)\\ &= \tfrac{1}{N_1}\textstyle\sum_{i:D_i=1}(Y_i(1) - Y_i(0)) + \tfrac{1}{N_1}\sum_{i:D_i=1} Y_i(0) - \tfrac{1}{N_0}\sum_{i:D_i=0} Y_i(0).\end{aligned}$$

On constate que l'estimateur de la différence des moyennes est égal à τ_1 plus un terme qui mesure le **biais de sélection** :

$$\frac{1}{N_1}\sum_{i:D_i=1} Y_i(0) - \frac{1}{N_0}\sum_{i:D_i=0} Y_i(0) \equiv BS. \qquad (1)$$

Autrement dit,

$$\hat{\tau}^{\text{diff}} = \tau_1 + BS.$$

Remarques :
- Le terme BS compare $Y(0)$ entre les groupes test et témoin.
- Les déterminants qui pourraient faire que $BS \neq 0$ n'apparaissent pas dans la formule. Ces déterminants seront capturés par les caractéristiques des individus.
- Pour que BS soit claire, écrivons-le dans le cas $N = 2$ ($N_1 = N_0 = 1$). BS est soit $Y_1(0) - Y_2(0)$ si $D_1 = 1$ et $D_2 = 0$ (l'individu 1 est dans le groupe test, l'individu 2 dans le groupe témoin) ou $Y_2(0) - Y_1(0)$ si $D_1 = 0$ et $D_2 = 1$.

Avec un MATAC pleinement randomisé permet en théorie d'éliminer le BS. La démonstration repose sur une transformation importante :

$$BS = \frac{1}{N_1}\sum_i Y_i(0)D_i - \frac{1}{N_0}\sum_i Y_i(0)\,(1 - D_i).$$

Prenons l'espérance mathématique de BS sur le support de D. Avec un MATAC, $E_D(D_i) = N_1/N$. Alors, l'espérance du premier terme vaut $\frac{1}{N_1}\sum_i Y_i(0)E_D(D_i) = \frac{1}{N}\sum_i Y_i(0)$. L'espérance du second : $\frac{1}{N_0}\sum_i Y_i(0)\,E_D(1 - D_i) = \frac{1}{N}\sum_i Y_i(0)$. Par conséquent, $E_D(BS) = 0$.

■

2) Avec les RP aléatoires, on a une loi jointe $j(y(1), y(0), d)$. Alors $\hat{\tau}^{\text{diff}} = E(Y|D = 1) - E(Y|D = 0) = ECMT + BS$, avec

$$BS = E(Y(0)|D = 1) - E(Y(0)|D = 0). \qquad (2)$$

Le traitement étant randomisé, $Y(0) \perp D$, donc $\Rightarrow E(Y(0)|D = d) = E(Y(0))$, $\forall\, d \in \{0, 1\}$. Littéralement, la décision de participer est indépendant de ce que les individus gagneraient, en moyenne, en l'absence de programme. Par conséquent, $BS = E(Y(0)) - E(Y(0)) = 0$.

■

4.2.2. Déséquilibre

Dans une EO, l'existence de BS peut être décelée avant d'estimer un effet causal.

Définition 4.1 : les groupes de traitement sont **déséquilibrés** (*unbalanced*) quand la distribution d'au moins un facteur de confusion diffère entre ces groupes ; Imbens et Rubin (2015, 32).

Le plus souvent, on compare les moments d'ordre 1. On doit pouvoir justifier pourquoi cette différence est le reflet d'un BS. Par ex., si les entreprises qui ont recours à une aide sont plus grandes en moyenne que celles du groupe de contrôle, c'est peut-être parce que recourir à une aide coûte moins pour une grande entreprise (la taille est souvent une proxy d'autres facteurs de ressources économiques).

Notons que dans une EXL, le déséquilibre reflète moins un BS qu'une randomisation défaillante, quand l'échantillon est trop petit.

Le déséquilibre porte sur des facteurs de confusion observables avant sélection dans les groupes (ils ne sont pas affectés par le traitement). On les appelle **variables de prétraitement** (*covariates*). Rosenbaum (2010) parle de **biais de sélection manifeste** (*overt bias*, *selection on observables*) ; la source du biais n'est pas cachée.[32]

Nous illustrons la situation de déséquilibre avec l'EO menée par Alan Krueger, reproduite partiellement par Angrist et Pischke (2009, 17-22). L'encadré 4.3 « Economie de l'éducation » résume quelques points du protocole, qui est plus sophistiqué que la description que nous en faisons.

C'est une EO car les écoles et des caractéristiques telles que public/privé, taille, quartier, etc., ne sont pas randomisés (ce sont des facteurs de confusion). C'est une source probable de BS. Et pour être retenue dans l'EX, une école devait avoir au moins trois classes dans chaque niveau car il y a trois traitements.

En revanche, l'affectation des élèves et des professeur·e·s dans les classes de tailles différentes suit un MATAC. Krueger ne prend pas des classes déjà constituées car, souvent, les élèves sont regroupés selon leurs résultats passés (les élèves en difficulté dans des petites classes) et des parents font du *lobbying* pour que leur enfant soit dans une petite classe ou ne soit pas avec un·e prof qui ne leur convient pas (un prétexte ?).

Le tableaux I, p. 503 dans Krueger (1999) est fréquent dans les EO (Rosenbaum, 2010, 20). Les facteurs possiblement déséquilibrés sont la couleur de peau (*White-Asian vs Black-African*), *Age in 1985*, et une variable *Free lunch* (cantine gratuite ou à prix réduit) renseignant sur le revenu des parents (les ménages aisés n'y ont pas droit ; voir le fichier **`krueger1999_table1.png`**.

Une mesure synthétique de la 'qualité' d'apprentissage avant affectation (***baseline outcome***) aurait été intéressante, mais n'était pas disponible. La variable utilisée est le *Percentile score* … (SAT) pour l'année d'entrée dans l'expérimentation STAR. On constate une différence significative du SAT entre les groupes. Dans le cas des *kindergarten* (partie A du tableau I), les différences entre les groupes pour les facteurs *Free lunch*, *White/Asian* et *Age in 1985* sont petites. La plus petite p-valeur est pour *Free lunch* (0,09), puis *White/Asian* (p-valeur = 0,26) et *Age in 1985* (p-valeur = 0,32). Au seuil de 5 %, on rejette l'égalité des moyennes au seuil de 5 % pour aucun facteur.

[32]. Ce biais est moins problématique que le ***hidden bias*** résultant de facteurs non-observables (cf. supra).

Encadré 4.3 : Economie de l'éducation

Aux États-Unis, le Congrès fit passer en 2002 une loi obligeant les 50 États fédérés à effectuer des expérimentations (financées par l'État fédéral) avant de réformer leurs systèmes d'éducation. Le programme STAR est l'une de ces expérimentations. Il s'agit d'une EXT menée au Tennessee et évaluée par Krueger (1999), puis reprise par de nombreux auteurs dont Angrist et Pischke (2009, 16-22).

La question causale posée par STAR est de savoir quels sont les effets de classes allégées (sur la qualité d'apprentissage des élèves) en primaire ? (« What are the effects of smaller classes in primary school? »)

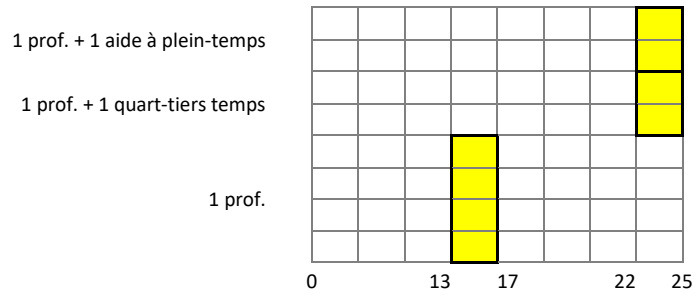
La théorie économique sur laquelle s'appuie l'étude considère l'éducation comme un investissement rentrant dans une fonction de salaire. L'auteur ne s'intéresse pas au nombre d'années dans le système scolaire, mais à l'apprentissage (*learning*). L'apprentissage est l'*output* d'une fonction de production transformant des intrants (*inputs*) au sein de l'école : effectif de la classe et autres caractéristiques du système éducatif : nombre d'enseignant·e·s par classe, leur expérience, les caractéristiques des élèves, etc.

L'allègement des classes accroit le nombre d'enseignant·e·s (à nb. d'élèves constant), et donc la part des dépenses d'éducation à la charge du citoyen américain. Est-ce une réforme efficace ? Efficiente ? Krueger (1999, 530-531) aborde l'efficience.

STAR est une EX à 12 M de $ réalisée dans les années 1980. Ella a duré quatre années, jusqu'à ce que les élèves de grande section (GS) de maternelle (*kindergarten*) atteignent le CE2 (*third grade*). Entre les deux, il y a le cours préparatoire (*first grade*) et le CE1 (*second grade*). Environ 11600 enfants participèrent sur les quatre ans.

Individus : élèves de grande section de maternelle en 1985-1986.
Traitement : effectif petit (13-17 élèves), grand avec aide partielle (22-25 élèves, un·e professeur·e aidant à quart-tiers-temps, du CP au CE2), grand avec aide à plein temps (22-25, professeur·e à plein temps). En 1985-86, l'effectif moyen est d'environ 22,3 élèves (plutôt proche de 22 que de 25).
Traité : apprentissage mais le résultat (la variable traitée expérimentale) est une mesure de la qualité d'apprentissage, une note d'examen.

La variable de traitement varie en fonction du nombre d'élèves et de professeur ETP de la classe.



Cependant, la randomisation n'a pas équilibré toutes les variables de confusion.

Penchons-nous un sur le taux d'**attrition** (*Attrition rate*) : la proportion d'élèves quittant l'EX au moins une fois avant même d'avoir complété le CE2 (c'est le pourcentage de *jours* avant la fin du *third grade* pour tous les élèves). Ce taux est significativement plus petit dans les petites classes (0,49), pour les élèves rentrés dans STAR en GS.

Nous remarquons que plus le niveau monte, moins la taille des classes a d'effet sur l'attrition. La p-valeur en *kindergarten*, 1st et 2nd grade est 0,02, 0,07 et 0,58. Une raison est que les élèves qui sont rentrés dans l'EX en *kindergarten* ont plus de chance de quitter l'échantillon puisque pour eux, l'EX s'étale sur une plus longue période.

L'effectif (*Class size in …*) reflète bien les différentes valeurs du traitement. Mais il y a eu du lobbying de parents pour que leurs enfants soient dans les petites classes. L'effectif moyen des classes *Small* en *kindergarten* est supérieur à l'effectif médian théorique ($15{,}1 > 15$), ce qui a eu pour effet de rééquilibrer un peu les tailles des grandes classes ($22{,}8 < 23{,}5$).

Dans l'article de Krueger (1999), la détection du déséquilibre repose sur des statistiques de test connues. Le calcul des p-valeurs de la dernière colonne (pour trois traitements), se fait avec un test de Fisher qui suit une $F_{2\alpha}(G-1, N-G)$ où G est le nombre de groupes de traitement ; donc $F_{2\alpha}(2, N-3)$.

La statistique de test se simplifie, $F_{2\alpha}(1, N-2) = T_{\alpha}^{2}(N-2)$; Cacoullos (1965). Voir l'exercice 4.4.3. C'est le T de Student pour variances égales.

[Faire le a) de l'exercice 4.4.2]

On recourt aussi à des « tests » moins sensibles aux effectifs que Fisher et Student. Voyons d'abord le cas à deux traitements :

La **différence normalisée** (Imbens et Wooldridge, 2009, p. 24) :

$$\frac{\bar{X}_1 - \bar{X}_0}{\sqrt{\hat{S}_{0,X}^2 + \hat{S}_{1,X}^2}} := \Delta'_X,$$

où $\hat{S}_{d,X}^2$ est un estimateur de $\sigma_{d,X}^2$, $d = 0{,}1$. Imbens et Wooldridge (2009) suggèrent la règle $|\Delta'_X| \leq 1/4$. Au-delà de $1/4$, l'évaluateur qui utilise des méthodes de contrôle des facteurs, court le risque d'un problème de spécification. Imbens et Rubin (2015) suggèrent une formule proportionnelle à la précédente, c'est celle que nous allons utiliser

$$\frac{\bar{X}_1 - \bar{X}_0}{\sqrt{(\hat{S}_{0,X}^2 + \hat{S}_{1,X}^2)/2}} := \Delta_X.$$

Δ_X se compare à la valeur 1 (le numérateur est forcément égal au dénominateur). Si $\Delta_X > 1$, le déséquilibre commence à être important. Les groupes ne sont pas très comparables.

Remarquons que Δ'_X est presque comme une statistique T de Student pour variances inconnues et inégales où les effectifs interviennent au dénominateur :

$$\frac{\bar{X}_1 - \bar{X}_0}{\sqrt{\hat{S}_{0,X}^2/N_0 + \hat{S}_{1,X}^2/N_1}}.$$[33]

[33] Le calcul du nombre de degrés de libertés est compliqué dans ce cas.

Mais, on préfère Δ_X à T, car T est trop sensible à la taille de l'échantillon. En effet, doublons N_0 et N_1 et appelons T' la nouvelle statistique. Nous avons :

$$T' = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{\hat{S}^2_{0,X}/2N_0 + \hat{S}^2_{1,X}/2N_1}} = \sqrt{2}T.$$

Doubler N_0 et N_1 conduit à une statistique 1,414 fois plus grande, augmentant ainsi la probabilité de rejet de l'équilibre. Le test gagne en **puissance** (rejet de l'équilibre à raison). Or, avec une taille d'échantillon au départ plus grande, il est plus facile d'équilibrer les variables par des méthodes d'appariement (cf. chapitre 6). En effet, plus N_0 est grand, plus je pourrais trouver des individus $i: D_i = 0$ comparables à $i: D_i = 1$.

Les formules que l'on vient de voir s'appliquent sur une variable à la fois. Dans le cas de deux groupes et plusieurs variables potentiellement déséquilibrées, nous verrons une mesure de déséquilibre, la **distance de Mahalanobis**, dans le chapitre 6.

4.2.3. Absence de recouvrement (lack of overlap)

Comme précédemment, on part du principe que lorsque l'on évalue une PP, l'évaluation est plus nette quand les groupes exposés aux différents traitements sont comparables. C'est la philosophie du MCR (on est censé avoir une population d'individus qui sont exposés ou pas au traitement actif). Dans STAR, il n'y a pas de problème de non-recouvrement pour l'origine ethnique, sinon *White/Asian* serait soit égal à 0 soit à 1 dans le tableau I, p. 503 de l'article. Illustrons le concept de **non-recouvrement** dans le cas fictif d'une variable à trois catégories codées 0, 1 et 2 et deux groupes de traitements.

Exemple 1

	X			
D_i	0	1	2	
0	25	25	50	100
1	50	25	25	100

Exemple 2

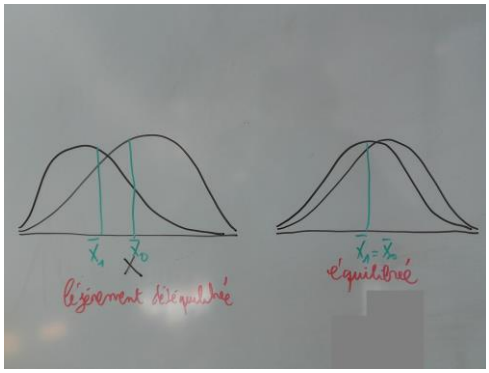
	X			
D_i	0	1	2	
0	50	0	50	100
1	0	100	0	100

Exemple 1 : il y a recouvrement. Quelle que soit la valeur de X, j'ai des individus dans les deux groupes. X est-elle déséquilibrée ? Calculons Δ_X à partir des données du tableau ($\bar{x}_1 = 0{,}75$, $\bar{x}_0 = 1{,}25$, $s_1^2 = s_0^2 = 0{,}6875$, écart-type $\sqrt{0{,}6875} \cong 0{,}83$) ; $\Delta_X \cong -0{,}60 < 1$ en valeur absolue (déséquilibre pas important).

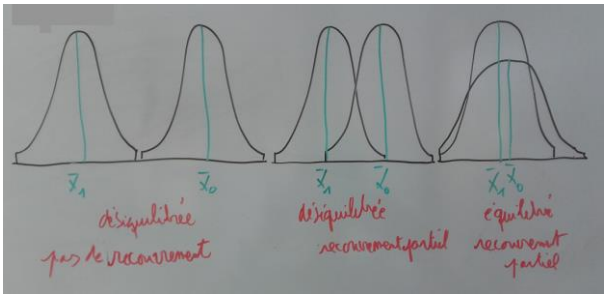
Exemple 2 : il n'y a pas recouvrement, X est équilibré. Pour chaque valeur de X, je n'ai que des individus tests (pour $X = 1$) ou témoins ($X \in \{0; 2\}$).

Quand X est continue, une méthode de détection visuelle est de superposer les histogrammes des groupes de traitement, ou faire d'autres calculs (cf. chapitre 6).

Recouvrement total



Recouvrement nul ou partiel



4.3. Supposition d'indépendance conditionnelle et recouvrement

Rappelons quelques éléments que nous venons de voir, de manière plus formelle.

1) Le MATAC atténue la valeur du terme de BS, en moyenne. Il suffit que l'affectation (D) soit indépendante des RP contrefactuels en l'absence de traitement ; $Y(0) \perp D$.

2) Nous avons privilégié les facteurs de confusion X observables (*selection on observable*) et non-déterminés par D_i.

3) Le MATAC rend les X indépendants des traitements : $\forall\, d \in \{0\,;1\},\ j(x,d) = m(x)m(d) \Rightarrow E(X|D=1) = E(X|D=0)$, *balanced covariates*. Si le MAT n'est pas AC, et $E(X|D=1) \neq E(X|D=0)$, alors on a détecté un BS.

4) Donc, si entre groupes de traitement, X prend des valeurs moyennes différentes, nous supposerons que $D \leftarrow X \rightarrow Y$.

5) Comme nous l'avions écrit dans la sous-section 2.1.3, il faut contrôler les facteurs de confusion X. Dans « Santé publique », X est l'état de santé avant les 12 derniers mois.

(i) les individus âgés et précaires, dont la santé depuis la naissance, ou à cause d'une maladie chronique, est fragile, vont plus souvent aux urgences ($X \rightarrow D \rightarrow Y$).

(ii) les individus âgés et précaires vont aux urgences dans l'espoir d'avoir une meilleure santé : $D|X = 1(Y(1) - Y(0) - X > 0)$. Plus l'état de santé de départ est faible, plus il y a de chance d'aller aux urgences (BS négatif).

La solution dans les chapitres 5 à 8 consistera à supposer que conditionnellement aux facteurs de confusion, eh bien l'ECM, l'ECMT ou toute autre fonction des RP, sont indépendants de la décision de participer, ce que l'on écrit ainsi :

$$(Y_i(1), Y_i(0)) \perp D_i|X_i \text{ (\textbf{Indépendance conditionnelle en distribution})}$$

Cette supposition d'indépendance conditionnelle (*conditional independence assumption*, CIA), ou non-confusion (***unconfoundedness***) ou sélection sur observables, ou *ignorability of treatment*, est plus forte que la supposition de MAT ignorable dont nous avons parlé dans la sous-section 3.3.3 (X n'intervenait pas).

CIA concerne les distributions de probabilités. Une supposition un peu moins forte est l'indépendance conditionnelle en moyenne (Wooldridge, 2010, 909) :

$$E(Y_i(d)|D_i, X_i) = E(Y_i(d)|X_i),\ \forall\, d \in \{0;1\} \text{ (\textbf{Indépendance conditionnelle en moyenne})}$$

Une supposition supplémentaire, de recouvrement (***overlap***, *common support*) est nécessaire pour identifier l'ECM :

$$\Pr(D_i = 1|X_i) < 1, \forall\, X_i \text{ sur son support}$$

4.4. Exercices de TP (4.4.1-4.4.4), et à l'oral (4.4.5)

4.4.1) **Biais de sélection dans le cas sans PFIC**. Exemple fictif du chapitre 3.

Tableau 3.1 et 3.2 combinés : RP

i	$Y(0)$	$Y(1)$	ECI_i	D_i	Y_i
1	**14**	13	−1	0	14
2	0	**6**	6	1	6
3	1	**4**	3	1	4
4	2	**5**	3	1	5
5	3	**6**	3	1	6
6	1	**6**	5	1	6
7	**10**	8	−2	0	10
8	**9**	8	−1	0	9
Total	40	56	16		−5,6
Moyenne	5	7	2 (ECM)		

- ECM $= 2$
- $\bar{Y}_1 - \bar{Y}_0 = -5{,}6$

Rappelons que nous avions supposé dans cet exemple que l'affectation de la subvention (en plus du CIR) aux huit entreprises était fonction de l'ECI (le traitement dépendait de la différence $Y_i(1) - Y_i(0)$). Le groupe test est constitué des i : $\text{ECI}_i > 0$.

a) On note $\sum_{i:D_i=1}(Y_i(1) - Y_i(0))/N_1$ par τ_1, l'effet causal moyen sur les traités (ECMT). Combien vaut τ_1 ?

$20/5 = 4$.

b) On note l'ECM par τ et $\bar{Y}_1 - \bar{Y}_0$ par $\hat{\tau}^{\text{diff}}$. Retrouver la valeur du terme de biais de sélection, noté BS, à partir de l'identité qui relie $\hat{\tau}^{\text{diff}}$ à τ_1 et BS

On peut utiliser l'identité $BS = \hat{\tau}^{\text{diff}} - \tau_1 = -5{,}6 - 4 = -9{,}6$, ou calculer directement :

$$\frac{1}{N_1}\sum_{i:D_i=1} Y_i(0) - \frac{1}{N_0}\sum_{i:D_i=0} Y_i(0) = \frac{7}{5} - \frac{33}{3} = 1{,}4 - 11 = -9{,}6.$$

4.4.2) a) *Déséquilibre*. Retrouver les p-valeurs de l'article de Krueger (1999) pour les variables *Free lunch* et *White/Asian*, des élèves rentrés dans STAR en *kindergarten*, à partir des du fichier **`krueger1999_balanceoverlap.do`** appliqué aux données **`webstar.dta`**.

Les variables à utiliser sont les suivantes :

- **`cltypek`** : le traitement renseignant les différentes tailles de classe en *kindergarten*, codés de 1 à 3 ;
- **`sesk`** : la variable nommée *Free lunch*, une proxy des conditions économiques de l'élève, code 1 ou 2 ;
- **`srace`** : la variable White/Asian, l'affiliation ethnique, codées de 1 à 6.

b) *Recouvrement*. Montrer avec un tableau la distribution de **`sesk`**, et avec un graphique celle de **`srace`** dans les trois groupes de traitement (sur le même graphique).

4.4.3) Supposons que nous ayons $G \geq 2$ groupes de traitement pour lesquels on veuille tester l'équilibre d'un facteur X. Si $G = 2$, on a le cas avec deux traitements.
Plutôt qu'une analyse de la variance (ANOVA), on construit un modèle de régression multiple ayant pour variables explicatives les variables dichotomiques des différents groupes :

$$X_i = \gamma_0 D_{i0} + \gamma_1 D_{i1} + \cdots + \gamma_{G-1} D_{iG-1} + V_i, \qquad (4.4.1)$$

où $D_{ig} = 1$ si i appartient au groupe g, et 0 sinon, V_i les termes d'erreur sont supposés i.i.d., et X_i est une variable aléatoire d'espérance inconditionnelle μ. On peut montrer que tester l'absence d'effet groupe (l'équilibre) pour le facteur X, formellement $G-1$ hypothèses (contraintes) d'égalité des moyennes conditionnelles $\mu_1 = \mu_0$, ..., $\mu_{G-1} = \mu_0$ revient à tester $\gamma_1 = \gamma_0$, ..., $\gamma_{G-1} = \gamma_0$.

Il est habituel de se ramener à des tests d'égalités à 0. Remarquez que dans le cas où nous n'avons que deux groupes, on retrouve les notations $D_{i1} = 1$ si i appartient au groupe du traitement actif, et 0 sinon. Or, $D_{i0} + D_{i1}$ est égal à 1 quel que soit i (i appartient nécessairement à l'un des deux groupes), ce qui justifie la transformation suivante : on remplace D_{i0} par $1 - \sum_{g\geq 1} D_{ig}$, d'où

$$\begin{aligned} X_i &= \gamma_0(1 - \textstyle\sum_{g\geq 1} D_{ig}) + \gamma_1 D_{i1} + \cdots + \gamma_{G-1} D_{iG-1} + V_i \\ &= \gamma_0 1 + (\gamma_1 - \gamma_0) D_{i1} + \cdots + (\gamma_{G-1} - \gamma_0) D_{iG-1} + V_i \end{aligned}$$

On peut définir γ' par $\gamma - \gamma_0$, de sorte que :

$$X_i = \gamma'_0 + \gamma'_1 D_{i1} + \cdots + \gamma'_{G-1} D_{iG-1} + V_i. \qquad (4.4.2)$$

Les $G-1$ hypothèse sont $\gamma'_1 = 0$, ..., $\gamma'_{G-1} = 0$.

a) Montrer que $\hat{\gamma}_g^{MCO} = \bar{X}_g$, $g = 0, \ldots, G-1$. En déduire que $\hat{\gamma}_g'^{MCO} = \bar{X}_g - \bar{X}_0$, $g = 1, \ldots, G-1$.

Soit :

$$\boldsymbol{D} := \begin{matrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & & \vdots \\ 0 & 1 & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & 1 & & \vdots \\ \vdots & 0 & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 0 \\ 0 & 0 & & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ N_0 \\ \\ \\ N_1 \\ \\ \\ \\ \\ N_{G-1} \end{matrix}$$

L'estimateur des MCO, noté $\hat{\boldsymbol{\gamma}}^{MCO}$, est $\hat{\boldsymbol{\gamma}}^{MCO} = (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{X}$. On a :

$$\boldsymbol{D}'\boldsymbol{D} = \begin{matrix} N_0 & 0 & \cdots & 0 \\ 0 & N_1 & & \\ \vdots & & \ddots & \\ 0 & & & N_{G-1} \end{matrix},$$

donc :

$$(\boldsymbol{D}'\boldsymbol{D})^{-1} = \begin{matrix} 1/N_0 & 0 & \cdots & 0 \\ 0 & 1/N_1 & & \\ \vdots & & \ddots & \\ 0 & & & 1/N_{G-1} \end{matrix}.$$

Notons maintenant que le produit $X_i D_{ig}$ vaut X_i si i appartient au groupe g. On a $\sum_{i=1}^{N} X_i D_{ig}$ qui est égal à $\sum_{i:D_{ig}=1} X_i$. Ainsi, le produit $\boldsymbol{D}'\boldsymbol{X}$, qui vaut :

$$\boldsymbol{D}'\boldsymbol{X} = \begin{bmatrix} \sum_{i:D_{i0}=1} X_i \\ \sum_{i:D_{i1}=1} X_i \\ \cdots \\ \sum_{i:D_{iG-1}=1} X_i \end{bmatrix}.$$

Par conséquent, en notant la moyenne du facteur X dans le groupe g, $N_g^{-1} \sum_{i:D_{ig}=1} X_i$, par $\bar{X}_g$, l'estimateur des MCO vaut simplement :

$$\widehat{\boldsymbol{\gamma}}^{MCO} = (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{X} = \begin{bmatrix} N_0^{-1} \sum_{i:D_{i0}=1} X_i \\ N_1^{-1} \sum_{i:D_{i1}=1} X_i \\ \cdots \\ N_{G-1}^{-1} \sum_{i:D_{iG-1}=1} X_i \end{bmatrix} = \begin{bmatrix} \bar{X}_0 \\ \bar{X}_1 \\ \cdots \\ \bar{X}_{G-1} \end{bmatrix}.$$

■

Concernant la deuxième partie de la question, il suffit de partir de la prédiction dans le modèle sans constante (<u>on omet 'MCO' dans la notation des estimateurs</u>) :

$$\hat{X}_i = \hat{\gamma}_0 D_{i0} + \hat{\gamma}_1 D_{i1} + \cdots + \hat{\gamma}_{G-1} D_{iG-1}. \qquad (4.4.3)$$

Si nous estimions au contraire le modèle avec une constante, la prédiction serait :

$$\hat{X}_i = \hat{\gamma}'_0 + \hat{\gamma}'_1 D_{i1} + \cdots + \hat{\gamma}_{G-1} D_{iG-1}. \qquad (4.4.4)$$

Or, puisque les D_{ig} somment à 1, on peut transformer 4.4.3 comme suit :

$$\hat{X}_i = \hat{\gamma}_0 + (\hat{\gamma}_1 - \hat{\gamma}_0) D_{i1} + \cdots + (\hat{\gamma}_{G-1} - \hat{\gamma}_0) D_{iG-1}.$$

Par conséquent, par identification, $\hat{\gamma}_0 = \hat{\gamma}'_0$, et $\hat{\gamma}'_g = \hat{\gamma}_g - \hat{\gamma}_0$ quel que soit $g = 1, \ldots, G-1$. Or, $\hat{\gamma}_g^{MCO} = \bar{X}_g$ d'après la réponse à la première partie de la question. Donc, $\hat{\gamma}_g'^{MCO} = \bar{X}_g - \bar{X}_0$.

On peut fait un test de Wald (statistique de Fisher) pour le modèle (4.4.2)

b) Le test d'équilibre du facteur X dans les différents groupes de traitement correspond au test standard des $G-1$ hypothèses $\gamma'_1 = 0$, ..., $\gamma'_{G-1} = 0$. La statistique du test de Fisher peut s'écrire en fonction des coefficients de corrélation multiple (ou coefficients

de détermination) dans le modèle non-contraint 4.4.2, R^2_{nc}, et dans le modèle contraint (X_i est égal à une constante + un terme d'erreur), R^2_c (qui vaut zéro par définition) :

$$f = \frac{R^2_{nc}}{1-R^2_{nc}} \frac{N-G}{G-1} \sim F(G-1; N-G).$$

c) Montrer que dans le cas $G = 2$, la statistique f devient

$$f = \frac{\bar{X}_1 - \bar{X}_0}{\left(\frac{1}{N_1}+\frac{1}{N_0}\right)\left(\frac{(N_0-1)\,\hat{S}_0^2 + (N_1-1)\,\hat{S}_1^2}{N_0+N_1-2}\right)},$$

et dire quelle loi suit cette statistique.

C'est la statistique du test de Student d'égalité des moyennes, avec variances supposées égales mais inconnues, vue dans la **sous-section 4.2.1**. Elle suit un $T^2(N-2)$.

4.4.4) Supposons le modèle linéaire vrai $E(Y|D,X) = \beta_1 + \beta_2 D + \beta_3 X$. Les paramètres $\beta_1, \beta_2, \beta_3$ prennent des valeurs quelconques. Vous estimez le modèle $E(Y|D) = \beta_1' + \beta_2' D$. Le coefficient qui nous intéresse est β_2, mais il va falloir se contenter de β_2'. Bien que la variable de confusion X soit observable, nous l'avons omise par ignorance, faute de connaissance du modèle vrai (ou, on connaît ce modèle, mais X n'est pas observable).

a) Montrer que le calcul de β_2' par les moindres carrés est $E(Y|1) - E(Y|0)$.
Déjà vu dans les exercices du chapitre 1.

b) Montrer à l'aide de la loi des espérances itérées que si X est équilibrée, alors le calcul par les moindres carrés de β_2' est β_2.

$$\begin{aligned}
\beta_2^{MC\prime} &= E(Y|1) - E(Y|0) \\
&= E_{X|D}(E(Y|1,X)|1) - E_{X|D}(E(Y|0,X)|0) \\
&= E_{X|D}(\beta_1 + \beta_2 + \beta_3 X|1) - E_{X|D}(\beta_1 + \beta_3 X|0) \\
&= \beta_1 + \beta_2 + \beta_3 E_{X|D}(X|1) - [\beta_1 + \beta_3 E_{X|D}(X|0)] \\
&= \beta_2 + \beta_3 [E_{X|D}(X|1) - E_{X|D}(X|0)].
\end{aligned}$$

Or d'après l'énoncé, X est équilibré, c'est-à-dire $E_{X|D}(X|1) = E_{X|D}(X|0)$. Par conséquent, $\beta_2^{MC\prime} = \beta_2$.

c) Montrer que l'estimateur $E_X\big(E(Y|1,X) - E(Y|0,X)\big) = \beta_2$.

$$\begin{aligned}
&E_X[E(Y|1,X) - E(Y|0,X)] \\
&= E_X(\beta_1 + \beta_2 + \beta_3 X) - E_X(\beta_1 + \beta_3 X) \\
&= \beta_1 + \beta_2 + \beta_3 E_X(X) - [\beta_1 + \beta_3 E_X(X)] \\
&= \beta_2.
\end{aligned}$$

4.4.5) **À la maison**. Biais de sélection et paradoxe de Simpson (inspiré de Lee, 2016, p. 25). Montrer que l'estimateur du c) à la question précédente règle le paradoxe de Simpson de l'exercice du chapitre 2 sur l'insertion des économistes. Le traitement D_i vaut 1 pour une étudiante en économie, et 0 pour une étudiante en droit. $X \in$ {"Paris", "Banlieue"} et $Y = 1$ si l'étudiante trouve un travail, $Y = 0$ sinon.

Hint. On a la situation $E(Y|D=1) - E(Y|D=0) < 0$, *mais* $E(Y|D=1,X) - E(Y|D=0,X) > 0$, $\forall\, X \in$ {"*Paris*","*Banlieue*"}. *Par la suite, vous pouvez noter* $X = 1$ si l'étudiante vient de Paris et $X = 0$ si elle vient de banlieue. On rappelle aussi que Y étant Bernoulli, $E(Y) = \Pr(Y = 1)$.

5. Stratification exacte

On peut utiliser cette méthode d'évaluation, que les données proviennent d'un protocole randomisé stratifié (*stratified randomized experiment*) ou pas (les études observationnelles). Dans la **section 5.1** nous introduisons la stratification dans le cas de l'évaluation randomisée, The Electric Company. Au sein de chaque niveau (grade), il y a eu affectation aléatoire des deux classes retenues dans le traitement, mais il pourrait y avoir un « effet niveau ». Nous nous appuierons sur un supposition d'indépendance conditionnelle permettant de considérer le niveau comme strate de confusion. La **section 5.2** présente les estimateurs de l'ECM et de l'effet causal moyen sur les traités (ECMT) dans le cas de deux strates. Nous introduirons aussi une nouvelle quantité à estimer, l'effet causal moyen sur les individus témoins/non-traités (ECMnT). Le passage à plus de deux strates sera alors évident, ce que nous verrons dans la **section 5.3**, où nous reviendrons sur l'expérimentation STAR.

5.1. Introduction

La stratification est une méthode de contrôle d'un ou plusieurs facteurs, au même titre que d'autres méthodes telles que l'appariement, la régression en discontinuité ou la méthode de différence de différences que nous verrons ultérieurement.

C'est une bonne méthode d'atténuation du BS lorsque les facteurs responsables de ce biais sont évidents. Elle fait partie de ces méthodes utilisées après un MATAC, mais qui peuvent aussi être utilisées dans des EO (Rosenbaum, 2010, 70).

[**Figure 5.1** sur les résultats de The Electric Company]

Dans The Electric Company, expérimentation randomisée au niveau des classes de chaque grade, une estimation de l'effet causal moyen dans le Grade 4 est $114 - 110 = 4$ sur le graphique (3,71 en arrondissant pas).

Cette estimation est assez éloignée de la différence de moyenne entre les deux groupes sans contrôler les Grades : 6 sur le graphique (5,65 au centième près).

La statistique de Student dans le Grade 4 pour l'hypothèse nulle (H_0) d'absence d'effet causal moyen (pour un test bilatéral), selon l'approche de Neyman, vaut 2,02 environ.

[Tester l'effet causal au seuil de 5 % pour les Grade 4 et 2 dans R]

La p-valeur est légèrement supérieure au seuil de 5 %, de sorte que nous ne rejetons pas H_0 dans le Grade 4 (dans le Grade 3 non plus). En revanche, dans le Grade 2, la p-valeur est inférieure à ce seuil (p-valeur $\cong 0{,}002$). Nous rejetons l'hypothèse.

Figure 5.1 : résultats de l'évaluation The Electric Company par niveau

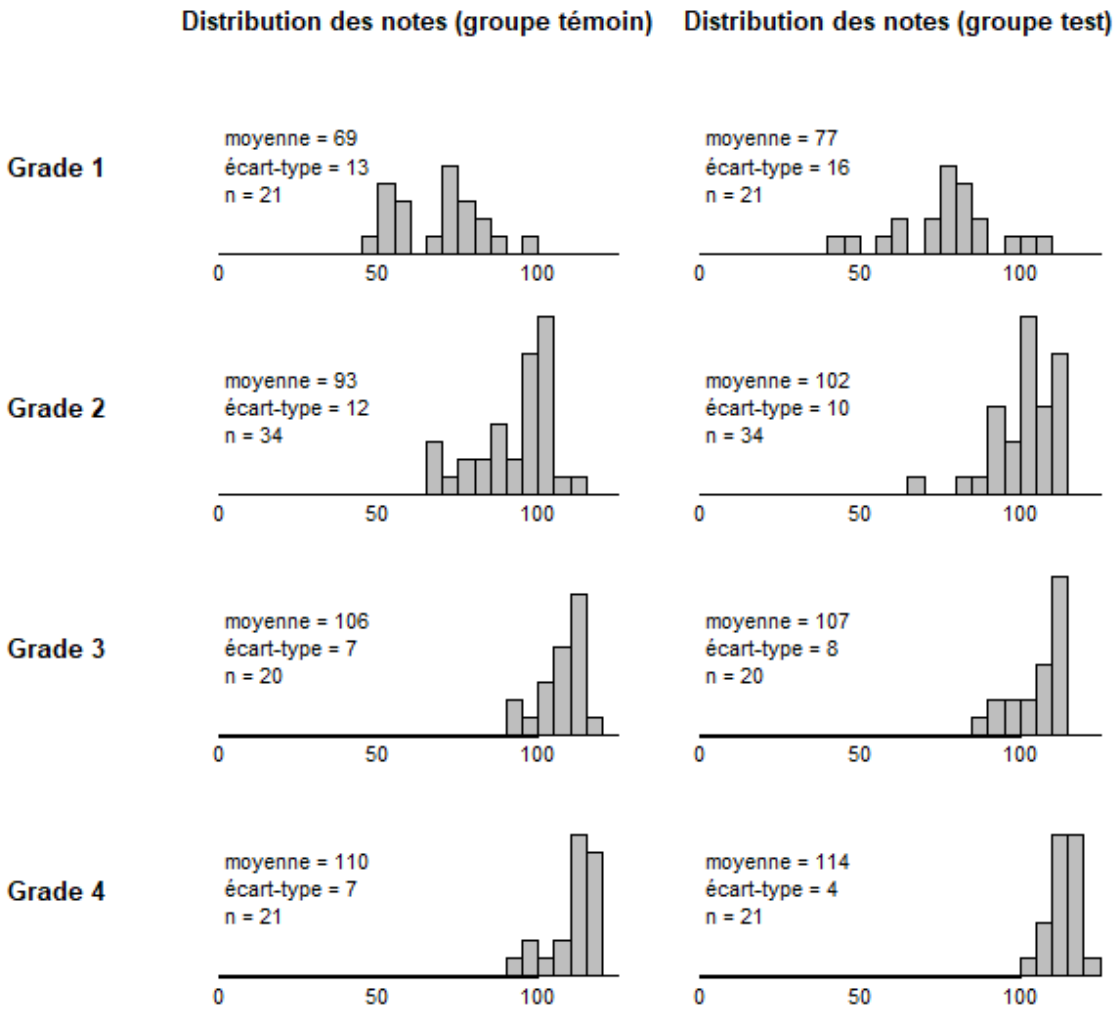


Tableau 5.1 : Résultats de l'intervention par Grade

	Individus tests ($\bar{Y}_{j1}$)	Individus témoins ($\bar{Y}_{j0}$)	N_1	N_0	
Grade 1	77,09	68,79	21	21	42
Grade 2	101,57	93,21	34	34	68
Grade 3	106,51	106,17	20	20	40
Grade 4	114,06	110,35	21	21	42
			96	96	192

Les effets par Grade ($77{,}09 - 68{,}79 = 8{,}3 > 0$, …, $114{,}06 - 110{,}35 = 3{,}71 > 0$) sont du même signe que l'effet agrégé (pas de paradoxe de Simpson),

$$(77{,}09 - 68{,}79)\frac{42}{192} + \cdots + (114{,}06 - 110{,}35)\frac{42}{192} \cong 5{,}65 > 0.$$

Pour revenir au concept de biais manifeste, les différents Grade sont connus avant l'affectation des élèves aux traitements. La stratification va consister à comparer le groupe test au groupe témoin, pour les mêmes valeurs des facteurs observés, c'est-à-dire des individus dans la même strate (on dit aussi appariés). Les individus sont stratifiés (*blocked*) pour différentes valeurs d'une variable X, ici une variable catégorielle, le Grade, dont on indice les valeurs par $j = 1, \ldots, 4$.

5.2. L'estimateur de l'ECM sur données stratifiées (approche à la Neyman)

Nous introduisons la stratification dans le cas de deux strates (**sous-section 5.2.1**). Cette section est aussi l'occasion d'introduire l'ECMT et celui sur les non-traités, noté ECMnT (**sous-section 5.2.2**). Les différences de calculs de l'ECM, l'ECMT et de L'EMCnT viennent simplement des pondération des différences des moyennes dans chaque strate.

5.2.1. Effet causal moyen

On introduit la méthode dans le cas de deux strates. La variable X_i définit la valeur de la strate pour l'individu i. Donc, $X_i \in \{x_1; x_2\}$. Par exemple, x_1 est la strate des demandeurs d'emploi qui habitent la banlieue, et x_2 celle des demandeurs d'emploi qui habitent Paris. Ou bien, x_1 est la strate des petites entreprises, et x_2 celle des grandes.

X_i est une variable de confusion, mais supposée ne pas être affectée par le traitement. On fait une supposition d'indépendance conditionnelle en moyenne :

$$E(Y_i(d)|X_i, D_i) = E(Y_i(d)|X_i), \forall\, X_i \in \{x_1; x_2\}, \forall\, D_i \in \{0; 1\}$$

Autrement dit, conditionnellement à X_i, les résultats potentiels ne dépendent pas du traitement D_i reçu, comme si ce dernier avait été randomisé.

Pour chaque strate nous pouvons définir un ECM :

$$\tau_{x_1} \equiv \frac{1}{N_{x_1}} \sum_{i:X_i=x_1} (Y_i(1) - Y_i(0)), \text{ et } \tau_{x_2} \equiv \frac{1}{N_{x_2}} \sum_{i:X_i=x_2} (Y_i(1) - Y_i(0)).$$

Ce sont des quantités que nous allons devoir estimer. Mais étant donné que l'on souhaite estimer l'ECM, il nous faut plutôt trouver une quantité qui agrège les strates (Imbens et Rubin, 2015, 201), une combinaison convexe de τ_{x_1} et τ_{x_2} par exemple :

$$\alpha\tau_{x_1} + (1 - \alpha)\tau_{x_2} := \tau^{\text{strat}}$$

α est une nombre qui doit être strictement compris entre 0 et 1. Naturellement, la littérature propose $\alpha \equiv \frac{N_{x_1}}{N}$, la proportion d'individus de l'échantillon dans la première strate. Par déduction $1 - \alpha = \frac{N_{x_2}}{N_{x_1} + N_{x_2}}$ est la proportion d'individus dans l'autre strate. Bien sûr, $N_{x_1} + N_{x_2} = N$. L'ECM pour données stratifiée est donc un ECM pondéré :

$$\boxed{\tau^{\text{strat}} = \frac{N_{x_1}}{N_{x_1} + N_{x_2}} \tau_{x_1} + \frac{N_{x_2}}{N_{x_1} + N_{x_2}} \tau_{x_2}}.$$

Il suffit de développer cette somme pour voir que c'est l'ECM que nous avons vu dans les **chapitres précédents**.

[Exercice 5.4.1]

Dans la strate x_1, un nombre d'individus N_{1,x_1} reçoit le traitement actif, et les autres, $N_{0,x_1} := N_{x_1} - N_{1,x_1}$ le traitement de contrôle. Idem pour la strate 2. Il y a N_{1,x_2} traités et N_{0,x_2} exposés au traitement de contrôle.

Les estimateurs des ECM de chaque strate sont du même type que celui lorsqu'on ne contrôle pas pour le facteur stratifié. Il s'agit de différences de moyennes :

$$\hat{\tau}^{\text{diff}}_{x_1} = \bar{Y}_{1,x_1} - \bar{Y}_{0,x_1} = \frac{1}{N_{1,x_1}}\sum_{i:X_i=x_1} Y_i D_i - \frac{1}{N_{0,x_1}}\sum_{i:X_i=x_1} Y_i(1-D_i),$$
$$\hat{\tau}^{\text{diff}}_{x_2} = \bar{Y}_{1,x_2} - \bar{Y}_{0,x_2} = \frac{1}{N_{1,x_2}}\sum_{i:X_i=x_2} Y_i D_i - \frac{1}{N_{0,x_2}}\sum_{i:X_i=x_2} Y_i(1-D_i).$$

Notons que au lieu de $\frac{1}{N_{1,x_1}}\sum_{i:X_i=x_1} Y_i D_i$, nous aurions pu écrire $\frac{1}{N_{1,x_1}}\sum_{i:X_i=x_1,D_i=1} Y_i$, qui est la moyenne des Y pour les individus qui sont dans la strate x_1 et qui reçoivent le traitement actif. Et, au lieu de $\frac{1}{N_{0,x_1}}\sum_{i:X_i=x_1} Y_i(1-D_i)$, nous aurions pu écrire $\frac{1}{N_{0,x_1}}\sum_{i:X_i=x_1,D_i=0} Y_i$; idem pour ces quantités dans l'autre strate.

> Le première façon a l'avantage de faire ressortir la variable aléatoire D. Et donc, de pouvoir conditionner les estimateur par rapport à cette variable dans les démonstrations.

Etant donné la forme de τ^{strat}, un estimateur « pondéré » de τ^{strat} est naturellement

$$\boxed{\frac{N_{x_1}}{N_{x_1}+N_{x_2}}\hat{\tau}^{\text{diff}}_{x_1} + \frac{N_{x_2}}{N_{x_1}+N_{x_2}}\hat{\tau}^{\text{diff}}_{x_2} := \hat{\tau}^{\text{diff}}_{\text{strat}}}.$$

Passons directement à l'estimateur de la variance de $\hat{\tau}^{\text{diff}}_{\text{strat}}$ dont nous aurons besoin si nous voulons tester des hypothèses sur l'effet causal du traitement. La variance a la même allure que celle que nous avons vu dans la **sous-section 3.3.2** où nous ne contrôlions pour aucun facteur. Ici, c'est une variance pondérée (la pondération joue le rôle de contrôle) :

$$\left(\frac{N_{x_1}}{N}\right)^2\left(\frac{s^2_{1,x_1}}{N_{1,x_1}} + \frac{s^2_{0,x_1}}{N_{0,x_1}}\right) + \left(\frac{N_{x_2}}{N}\right)^2\left(\frac{s^2_{1,x_2}}{N_{1,x_2}} + \frac{s^2_{0,x_2}}{N_{0,x_2}}\right).$$

Rappelons que les variances $s^2_{\cdot,\cdot}$ sont les estimateurs qui corrigent la taille de l'échantillon ; par exemple, $s^2_{1,x_1} = \frac{1}{N_{1,x_1}-1}\sum_{i:X_i=x_1,D_i=1}(Y_i - \bar{Y}_1)^2$. La formule ci-dessus est un estimateur sans biais de la variance de $\hat{\tau}^{\text{diff}}_{\text{strat}}$.

5.2.2. Effet causal moyen sur les traités et les non-traités

Ces effets causaux et leurs estimateurs suivent la même logique d'écriture que pour l'ECM. Nous avons toujours nos deux strates. La variable X_i a la même définition. Il n'y a que les pondérations qui changent :

$$\alpha_1\tau_{x_1} + (1-\alpha_1)\tau_{x_2} := \tau^{\text{strat}}_1,$$

les traités

où $\alpha_1 = \frac{N_{1,x_1}}{N_{1,x_1}+N_{1,x_2}}$, la proportion d'individus traités dans la première strate relativement au nombre total d'individus traités. L'ECMT pour données stratifiée est donc :

$$\tau^{\text{strat}}_1 = \frac{N_{1,x_1}}{N_{1,x_1}+N_{1,x_2}}\tau_{x_1} + \frac{N_{1,x_2}}{N_{1,x_1}+N_{1,x_2}}\tau_{x_2}.$$

Un estimateur naturel de l'ECMT est :

$$\frac{N_{1,x_1}}{N_{1,x_1}+N_{1,x_2}}\hat{\tau}^{\text{diff}}_{x_1} + \frac{N_{1,x_2}}{N_{1,x_1}+N_{1,x_2}}\hat{\tau}^{\text{diff}}_{x_2} := \hat{\tau}^{\text{diff}}_{1,\text{strat}}.$$

Pour la variance, nous avons :

$$\left(\frac{N_{1,x_1}}{N_{1,x_1}+N_{1,x_2}}\right)^2\left(\frac{s^2_{1,x_1}}{N_{1,x_1}}+\frac{s^2_{0,x_1}}{N_{0,x_1}}\right)+\left(\frac{N_{1,x_2}}{N_{1,x_1}+N_{1,x_2}}\right)^2\left(\frac{s^2_{1,x_2}}{N_{1,x_2}}+\frac{s^2_{0,x_2}}{N_{0,x_2}}\right).$$

Remarques :

- évidemment, nous pouvons aussi nous intéresser à l'effet causal moyen sur les non-traités (ECMnT), τ_0^{strat}, qui répond à la question de savoir quel serait l'effet moyen du traitement pour les individus témoins, s'ils avaient été traités. Il suffit de remplacer N_{1,x_1} par N_{0,x_1} et N_{1,x_2} par N_{0,x_2} partout dans les formules de cette sous-section ;

- l'utilisation des variances corrigées pour le calcul des ECM et ECMT, voire de l'ECMnT, implique nécessairement que le nombre d'individus traités dans chaque strate soit au moins égal à 2 ($N_{1,x_1} \geq 2, N_{1,x_2} \geq 2$). Un raisonnement similaire pour les non-traités doit être fait. Par conséquent, le nombre d'individus dans l'échantillon est au moins être égal à quatre fois le nombre de strates.

5.2.3. Biais de l'estimateur de l'ECMT stratifié

Examinons le biais dans le cas de l'estimation de l'ECMT, qui est le paramètre généralement retenu dans les travaux d'évaluation. La supposition d'indépendance conditionnelle va jouer un rôle crucial. Il s'agit donc de montrer que $E\left(\hat{\tau}^{\text{diff}}_{1,\text{strat}}\right) = \tau_1$.

Nous suivons l'approche du modèle d'échantillonnage, plus épurée que celle en termes de population de taille finie, où il faut en permanence manipuler le signe de sommation. On suppose donc que les quadruplets $(Y_i(0), Y_i(1), X_i, D_i) \equiv A_i$ indépendants et identiquement distribués, tirés dans une super-population infinie.

Notez que c'est l'approche adoptée pour l'écriture de CIA.

Réécrivons l'estimateur $\hat{\tau}^{\text{diff}}_{1,\text{strat}}$. Dans chaque strate, nous avons l'estimateur $E(Y_i|D_i = 1, X_i) - E(Y_i|D_i = 0, X_i)$. Cette différence doit être prise sur l'ensemble des strates (ici deux, $X_i \in \{x_1; x_2\}$), mais conditionnellement au fait d'être traité ($D_i = 1$). Nous avons donc :

$$\begin{aligned}
\hat{\tau}^{\text{diff}}_{1,\text{strat}} &= E_{X|D}(E(Y_i|D_i = 1, X_i) - E(Y_i|D_i = 0, X_i)|D_i = 1) \\
&= E_{X|D}(E(Y_i(1)|D_i = 1, X_i) - E(Y_i(0)|D_i = 0, X_i)|D_i = 1) \\
&= E_{X|D}(E(Y_i(1)|D_i = 1, X_i) - E(Y_i(0)|D_i = 1, X_i)|D_i = 1) && \Leftarrow \text{CIA} \\
&= E_{X|D}(E(Y_i(1) - Y_i(0)|D_i = 1, X_i)|D_i = 1). \\
&= E(Y_i(1) - Y_i(0)|D_i = 1). \\
&= \tau_1
\end{aligned}$$

■

Dans le cadre d'une population de taille finie, avec les RP fixes, où seuls les D_i et X_i sont aléatoires, on peut commencer par montrer que $E\left(\bar{Y}_{d,x}\right) = \overline{Y(d)}$, pour $d = 0$ et $d = 1$. De sorte que

$$E\left(\hat{\tau}^{\text{diff}}_{1,\text{strat}}\right) = \sum_x \frac{N_{1,x}}{N_1}\left(E\left(\bar{Y}_{1,x}\right) - E\left(\bar{Y}_{0,x}\right)\right) = \sum_x \frac{N_{1,x}}{N_1}\left(\overline{Y(1)} - \overline{Y(0)}\right) = \overline{Y(1)} - \overline{Y(0)}$$

5.3. Application au projet STAR

Nous allons voir l'intérêt de contrôler le facteur « école » par la méthode de stratification que nous venons de voir. C'est l'occasion de discuter un peu la supposition SUTVA introduite dans le chapitre 2.

Théoriquement, un élève d'une classe « Small » devrait être en tout point comparable à un élève d'une classe « Regular » et « Regular/Aide ». Rappel de la double randomisation :

> Les élèves qui furent affectés aléatoirement dans des classes de tailles différentes.
>
> Cela après que les professeurs ont été affectés aléatoirement dans ces classes.
>
> Cette double randomisation garantie que, s'il y a un effet *Small class* positif (resp. négatif), ce n'est pas parce que ces classes ont de meilleurs (moins bons) élèves et professeurs.

Mais, seules les écoles ayant au moins trois classes dans chaque niveau étaient éligibles (3 classes au moins en GS et/ou 3 classes au moins en CP, …).

> L'affectation aléatoire des élèves dans les classes au sein d'une école, ne garantit pas l'équilibre des caractéristiques des écoles (Angrist et Pischke, 2009, 23).
>
> Cela soulève des questions. Par exemple, si les élèves habitant à la campagne ont en général plus de chances d'être dans de petites classes que les élèves habitant en ville, les zones rurales étant moins denses en général.
>
> > L'estimation de l'effet des petites classes, relativement aux deux autres types de classes, serait biaisée par un facteur non-pris en compte, la localisation géographique de l'école.

Krueger (1999) a réglé ce problème en contrôlant pour le facteur « école » dans un modèle de régression avec un effet fixe par école. Son évaluation révèle un effet *Small class* supérieur avec contrôle que sans (5,37 pdp au lieu de 4,82 pdp).

> Ce résultat, reporté par Angrist et Pischke (2009, 19-20), s'interprète comme suit. La performance d'un élève est mesurée par le Stanford Achievement Test (SAT) en pourcentage. Si un élève d'une école a un SAT de 50, cela veut dire que 50 % des élèves ont plus que lui. Si cet élève était placé dans une petite classe, ce ne serait plus 50 % des élèves de l'école qui auraient plus, mais $50\ \% - 5{,}37\ \text{pdp} = 44{,}63\ \%$ des élèves. À la page 20 de l'ouvrage, on voit que le contrôle de l'effet « école » ne change pas beaucoup le résultat, qui passe de 45,18 % à 44,63 %, mais l'estimation plus précise (l'erreur-type passe de 2,19 à 1,26).

Remarque sur le protocole de STAR et l'hypothèse de non-interférence dans SUTVA, à peine abordée dans le **chapitre 2** :

> On aurait pu prendre les classes pour « individus » plutôt que les élèves. Dans ce cas, l'école serait la strate. Ce protocole résoudrait un problème important de STAR, celui de l'interaction entre les élèves. En effet, le SAT d'un élève n'est pas indépendant des résultats de ses camarades (l'interaction est bénéfique en général, chacun pouvant profiter des réponses apportées par le professeur à ses camarades). Dans le cas où l'individu est la classe, la supposition d'interférence est moins fragile, mais alors le problème d'interférence est impossible à contrôler.
>
> > Du point de vue du professeur : l'apprentissage est-il meilleur dans une petite classe (moins d'interactions, mais moins de dispersion) ? Est-ce mieux pour le professeur ou les élèves ? Les deux ! Est-ce que 15 élèves-un professeur c'est pareil que 30 élèves-deux professeurs ?

5.3.1. Version de STAR d'Imbens et Rubin (2015)

Imbens et Rubin (2015) font une analyse des données de l'expérimentation STAR où l'individu est la classe.

L'intérêt, nous le rappelons, est que l'on peut supposer qu'il y a moins d'interactions entre les classes d'une école qu'entre les élèves d'une classe !

La variable de traitement est la taille qui peut prendre deux valeurs : petite, ou grande (avec une aide à mi-temps). Ils suppriment donc les grandes classes où un professeur aide à plein temps. Ils ne gardent que les niveaux *kindergarten* et les écoles qui ont au moins deux classes de chaque type (il y a donc au moins quatre classes par école) ; ça donne 16 écoles (strates) et 68 classes avec l'affectation suivante, pour reprendre nos écritures, $N_0 = 32$ et $N_1 = 36$.

Remarque sur la notation W_i en colonnes, qui en fait le D_i dans ce cours.

[Tableaux 9.1 et 9.2 de Imbens et Rubin (2015, 190,204)]

13/16 écoles ont deux classes de chaque type.

Après avoir centré et réduit les notes des élèves du sous-échantillon, la variable traitée est la note moyenne de la classe en mathématique. Cela donne une étendue de 7,07 pour les notes des classes (entre $-4{,}13$ et 2,94). La note moyenne dans le groupe de contrôle est de $-0{,}13$ (écart-type 0,56) et de 0,09 dans le groupe des petites classes (écart-type 0,61).

A partir de la **section 5.2**, l'approche pour estimer l'ECM, en contrôlant pour le facteur école, consiste à calculer les différences de moyennes $\hat{\tau}_x^{\text{diff}}$ pour chacune des strates $x = x_1, \dots, x_{16}$, puis, de prendre une moyenne pondérée de ces différences de moyennes. Nous pouvons voir dans le tableau 9.2 que la valeur de cette moyenne pondérée,

$$\sum_{j=1,\dots,16} \frac{N_{x_j}}{N} \hat{\tau}_x^{\text{diff}} = 0{,}241.$$

Sous l'hypothèse que τ_x^{diff} est constant d'une école à l'autre, alors l'estimateur sans biais de l'écart-type de cette statistique est une extension de celui vu à la sous-section précédente à 16 strates :

$$\left(\sum_{j=1,\dots,16} \left(\frac{N_{x_j}}{N} \right)^2 \left(\frac{s_{1,x_j}^2}{N_{1,x_j}} + \frac{s_{0,x_j}^2}{N_{0,x_j}} \right) \right)^{1/2} = 0{,}092.$$

On en déduit l'intervalle de confiance suivant au seuil de significativité de 5 % :

$$(0{,}061;\ 0{,}421).$$

Sans stratification, l'intervalle est $(-0{,}053;\ 0{,}500)$ qui comprend la valeur 0 ! L'intervalle est plus large (moins précis).

5.3.2. Réplication sous Stata

Répliquer l'évaluation de l'expérimentation STAR sous Stata est compliqué dans la mesure où les identifiants des classes ont été retirés de la version publique des données.

Angrist et Pischke (2009) proposent une réplication, avec un bout de code permettant de reconstruire des identifiants de classes.

La résolution ressemble à la reconstruction d'une clé qu'on aurait perdu.

Nous nous sommes appuyés en partie sur leur code, afin d'essayer de nous approcher de l'exercice menée par Imbens et Rubin (2015).

[Présenter le code « **krueger1999.do** »]

5.4. Exercices

5.4.1) En passant par les résultats potentiels (RP), montrer que $\tau^{strat} = \tau$, $\tau_1^{strat} = \tau$. En déduire que $\tau_0^{strat} = \tau$, sans utiliser les RP cette fois-ci.

5.4.2 (Données tirées de Gelman et Hill (2006)). On veut étudier l'effet d'un programme médico-social sur le développement d'enfants prématurés (qui ont un faible poids) nés dans les années 1980. Ce programme, appelé Infant *Health and Development Program (HDP)*, consiste en une assistance médicale et sociale renforcée (suivi pédiatrique, visites au domicile, suivi des apprentissages, etc.). La variable traitée est une note à un test de QI à l'âge de 3 ans. Dans le tableau ci-dessous figure l'effet moyen du programme HDP pour chaque niveau d'éducation de la mère du foyer de l'enfant.

			Taille de l'échantillon		
Niveau d'éducation de la mère	Effet moyen du traitement	Erreur type	Test	Témoin	Total
Avant baccalauréat	9,3	1,3	126	1358	1484
Baccalauréat	4,0	1,8	82	1820	1902
Licence 1 à 3	7,9	2,3	48	837	885
Master	4,6	2,1	34	366	400
			290	4381	4671

a) De quels individus est constitué le groupe témoin ?

b) Il y a-t-il recouvrement pour la variable « Niveau d'éducation de la mère » ?

c) Estimer l'ECM pour les données stratifiées.

d) Comparer ce résultat à l'estimation de l'ECMT.

e) Construire un intervalle de confiance au seuil de 5 % pour l'ECMT.

Corrections des exercices du chapitre 5

5.4.1

$$\begin{aligned}\tau^{\text{strat}} &= \frac{N_{x_1}}{N_{x_1}+N_{x_2}}\tau_{x_1} + \frac{N_{x_2}}{N_{x_1}+N_{x_2}}\tau_{x_2} \\ &= \frac{N_{x_1}}{N_{x_1}+N_{x_2}}\frac{1}{N_{x_1}}\sum_{i:X_i=x_1}(Y_i(1)-Y_i(0)) + \frac{N_{x_2}}{N_{x_1}+N_{x_2}}\frac{1}{N_{x_2}}\sum_{i:X_i=x_2}(Y_i(1)-Y_i(0)) \\ &= \frac{1}{N}\sum_{i:X_i=x_1}(Y_i(1)-Y_i(0)) + \frac{1}{N}\sum_{i:X_i=x_2}(Y_i(1)-Y_i(0)) \\ &= \frac{1}{N}\sum_i(Y_i(1)-Y_i(0)) \\ &= \tau\end{aligned}$$

■

$$\begin{aligned}\tau_1^{\text{strat}} &= \frac{N_{1,x_1}}{N_{1,x_1}+N_{1,x_2}}\tau_{x_1} + \frac{N_{1,x_2}}{N_{1,x_1}+N_{1,x_2}}\tau_{x_2} \\ &= \frac{N_{1,x_1}}{N_{1,x_1}+N_{1,x_2}}\frac{1}{N_{x_1}}\sum_{i:X_i=x_1}(Y_i(1)-Y_i(0)) + \frac{N_{1,x_2}}{N_{1,x_1}+N_{1,x_2}}\frac{1}{N_{x_2}}\sum_{i:X_i=x_2}(Y_i(1)-Y_i(0)) \\ &= \frac{1}{N_{1,x_1}+N_{1,x_2}}0{,}5\sum_{i:X_i=x_1}(Y_i(1)-Y_i(0)) + \frac{1}{N_{1,x_1}+N_{1,x_2}}0{,}5\sum_{i:X_i=x_2}(Y_i(1)-Y_i(0)) \\ &= \frac{1}{0{,}5N}0{,}5\sum_{i:X_i=x_1}(Y_i(1)-Y_i(0)) + \frac{1}{0{,}5N}0{,}5\sum_{i:X_i=x_2}(Y_i(1)-Y_i(0)) \\ &= \frac{1}{N}\sum_i(Y_i(1)-Y_i(0)) \\ &= \tau\end{aligned}$$

■

Concernant $\tau_0^{\text{strat}} = \frac{N_{0,x_1}}{N_{0,x_1}+N_{0,x_2}}\tau_{x_1} + \frac{N_{0,x_2}}{N_{0,x_1}+N_{0,x_2}}\tau_{x_2}$, la déduction est facile. En effet,

$$N_0\tau_0^{strat} + N_1\tau_1^{strat} = N_{x_1}\tau_{x_1} + N_{x_2}\tau_{x_2}, \qquad (1)$$

qui vaut $N\tau$. Et nous savons que $\tau_1^{strat} = \tau$. Donc, (1) devient $N_0\tau_0^{strat} + N_1\tau = N\tau$. Par conséquent, $\tau_0^{strat} = \tau$.

■

5.4.2

a) D'enfants qui n'ont pas participé à ce programme

b) Oui.

c) L'estimateur de l'ECM vaut :

$$9{,}3 \times \frac{1484}{4671} + \cdots 4{,}6 \times \frac{400}{4671} = 6{,}47.$$

d) 7,02.

d) Construire un intervalle de confiance au seuil de 5% pour l'ECMT.

Il nous faut d'abord une estimation de l'écart-type de l'estimateur trouvé à la question « c) » :

$$\left(\left(\frac{126}{290}\right)^2 \times 1{,}3^2 + \cdots + \left(\frac{34}{290}\right)^2 \times 2{,}1^2\right)^{1/2} = 0{,}885$$

L'intervalle de confiance est $(7{,}02 - 1{,}96 \times 0{,}885; 7{,}02 + 1{,}96 \times 0{,}885) = (5{,}28; 8{,}75)$.

6. Appariement

Comme la stratification vu précédemment, l'appariement est une méthode d'atténuation du BS quand les facteurs de confusion sont observables. Nous allons d'abord motiver cette méthode avec une population abstraite (**section 6.1**). Nous verrons deux types d'appariement. Le premier, appelé estimateur d'appariement, permet de sélectionner et apparier de manière **exacte** et **inexacte** les individus des groupes de traitements (**section 6.2**), puis les comparer avec un test de moyennes. Nous l'appliquerons aux données de l'article de Card et Krueger (1994).

Bien que simple, cette méthode sacrifie beaucoup d'observations, donc de degrés de libertés. Nous pouvons utiliser un maillage des variables (*coarsening*), permettant de résoudre ce problème, mais pour un nombre limité de variables de confusion. Le second type d'appariement est plus flexible, mais aussi plus compliqué. C'est l'appariement par le score de propension (**section 6.3**). L'estimation de l'effet du traitement est ensuite libre (estimateur d'appariement, test de moyenne, stratification, régression, etc.). La **section 6.4** porte sur le score de propension généralisé pour le cas d'un traitement continu. Nous verrons quelques applications, et boucleront le chapitre avec une utilisation à la Horvitz-Thompson du score (**section 6.5** d'exercices).

6.1. Motivations théoriques

Supposons la population suivante, avec X déséquilibré, qui n'est pas indépendant de la sélection dans les groupes de traitement. Par exemple, X mesure le coût d'opportunité du recours à une PP pour l'individu (montage d'un dossier CIR, frais de transport pour aller à une formation, coût psychologique, etc.). Les individus (pas le statisticien) connaissent les RP. Le MAT est tel qu'un individu a recours à la PP si $Y(1)$, net du coût de recours, $Y(1) - X$, est plus grand que $Y(0)$:

$$D(X) = \mathbb{I}\big(Y(1) - X \geq Y(0)\big).$$

Pour une modélisation simple d'un problème de décision dans le MCR, avec X qui rentre dans le coût de recours, on peut voir Florens *et alii* (2008) et Abadie *et alii* (2004).

X prend trois valeurs, $x > x' > x''$, avec x qui n'est pas choisi au hasard : $x = 3x'' - 2x'$.[34] L'effet de l'intervention est $\delta > x$. Ces valeurs sont distribuées comme suit.

i	X	$Y(0)$	$Y(1)$	D	Y	$j(x, y(0), y(1), d)$
1	x	$2a$	$2a + \delta$	1	$2a + \delta$	1/5
2	x	0	δ	1	δ	1/5
3	x'	b	$b + \delta$	1	$b + \delta$	1/5
4	x''	c	$c + \delta'$	0	c	1/5
5	x	a	$a + \delta'$	0	a	1/5

Note : $j(x, y(0), y(1), d)$ est la probabilité de chaque quadruplet dans la super-population.

Les individus $i \in \{1; 2; 3\}$ ont recours à la PP, mais $i \in \{4; 5\}$ non. Par ex., pour $i = 2$, on a $Y_2(1) - Y_2(0) - X_2 = \delta - 0 - x = \delta - x > 0$ par définition, donc $D_2 = 1$. Pour $i = 4$, $Y_4(1) - Y_4(0) - X_4 = (c + \delta') - c - x'' = \delta' - x'' < 0$ par définition, donc $D_4 = 0$.

[34] On a calculé $E(X|D = 1) = (2x + x')/3$ et $E(X|D = 0) = (x + x'')/2$, qui sont égaux pour l'équilibrage si $x(x', x'') = 3x'' - 2x'$. Par ailleurs, $x(x', x'') > x' \Leftrightarrow x'' > x'$, et $x(x', x'') > x'' \Leftrightarrow x'' > x'$.

On va montrer que si on évalue dans la sous-population pour laquelle $X = x$, on mesure δ sans biais. Notons que par définition, δ c'est l'ECMT, mais aussi l'ECMnT et l'ECM. Vérifions le premier :

$$E(Y(1) - Y(0)|D = 1) = \frac{2a + \delta - 2a}{3} + \frac{\delta - 0}{3} + \frac{b + \delta - b}{3} = \delta.$$

Sans randomisation, la différence de moyennes $E(Y|D = 1) - E(Y|D = 0)$ ne permet pas d'obtenir l'effet causal. C'est l'ECMT + BS :

$$\frac{2a+\delta+\delta+b+\delta}{3} - \frac{c+a}{2} = \delta + \frac{a+2b-3c}{6}.$$

L'appariement consiste à remplacer les RP non-observables (RP contrefactuels) des unités telles que $D = 1$ (les valeurs en gris dans la colonne $Y(0)$) par le RP observable (RO) du groupe témoin ($D = 0$) au « point » $X = x$ seulement, $Y(0) = a$. Le seul point pour lequel il y a équilibre et recouvrement.

Etudions la sous-population des $X = x$.

Notons auparavant que $\Pr(D = 1, X = x) = \Pr(D = 1|X = x)\Pr(X = x) = (2/3)(3/5) = 2/5$. Et que $\Pr(D = 0, X = x) = (1 - 2/3)(3/5) = 1/5$.

CIA est vérifiée en $X = x$, la seule valeur pour laquelle il y a recouvrement et équilibre.

- Recouvrement : $\boldsymbol{x}$ est la seule valeur de $\boldsymbol{X}$ présente dans les deux groupes.
- Equilibre :

$$E(X|D = 1) = \frac{2}{3}x + \frac{x'}{3} = \frac{2}{3}(3x'' - 2x') + \frac{x'}{3} = 2x'' - x'.$$
$$E(X|D = 0) = \frac{1}{2}x'' + \frac{1}{2}x = \frac{1}{2}x'' + \frac{1}{2}(3x'' - 2x') = 2x'' - x'.$$

■

- CIA :

$$E(Y(1)|D = 1, X = x) = (2a + \delta)\frac{1}{5}/\frac{2}{5} + \delta\frac{1}{5}/\frac{2}{5} = a + \delta, E(Y(1)|D = 0, X = x) = (a + \delta)\frac{1}{5}/\frac{1}{5} = a + \delta$$
$$E(Y(0)|D = 1, X = x) = 2a\frac{1}{5}/\frac{2}{5} + 0\frac{1}{5}/\frac{2}{5} = a, E(Y(0)|D = 0, X = x) = a\frac{1}{5}/\frac{1}{5} = a.$$

■

Le BS ne peut être calculé que pour $X = x$ (on contrôle), et dans toute la population (on ne contrôle pas).

- Pour $\boldsymbol{X = x}$, BS est nul ($\boldsymbol{a - a}$) ; voir **ci-dessus**.
- Dans toute la population :

$$E(Y(0)|D = 1) - E(Y(0)|D = 0) = E(E(Y(0)|D = 1, X)|1) - E(E(Y(0)|D = 0, X)|0).$$
$$= a\frac{2}{3} + b\frac{1}{3} - \frac{1}{2}a - \frac{1}{2}c = \frac{a+2b-3c}{6}.$$

Calculons maintenant les différences de moyennes en $X = x$.

$$E(Y|D = 1, X = x) - E(Y|D = 0, X = x) = E(Y(1)|D = 1, X = x) - E(Y(0)|D = 0, X = x)$$

par définition. C'est donc égal à $a + \delta - a = \delta$.

La méthode d'appariement est quasi-expérimentale (au sens de non-randomisée). Elle peut être employée de deux manières (Imbens et Rubin, 2015, 275-276)

Estimateur d'appariement, **comme dans l'illustration précédente**
Estimer l'effet causal pour chaque individu pioché :

1) Pour chaque individu, sélectionner des individus comparables
2) Estimer un effet causal

La sélection peut être avec ou sans remise

Appariement via le score de propension
Équilibrer les groupes de traitement dans une première étape, estimer l'effet causal dans une deuxième étape, à partir d'une méthode standard (MC, stratification, estimateur d'appariement, etc.) :

1) Retirer des individus pas comparables à différente strates du score
2) Utiliser une méthode d'estimation au choix de l'effet causal

Dans tous les cas, l'appariement vise à atténuer des différences entre les individus, avant de comparer les résultats observés afin que $\mathrm{E}(X|D=1) \cong \mathrm{E}(X|D=0)$. Et plus fortement, $\Pr(X=x|D=1) \sim \Pr(X=x|D=0)$. **Exercice 6.5.6** : $e(X) \perp X$ est suffisant.

Définition : des individus appariés (*matched*) possèdent des caractéristiques individuelles observables proches voire identiques

En pratique, les unités appariées idéales sont « jumelles », c'est-à-dire, pour chaque valeur du vecteur de caractéristiques d'une unité i traitée $\boldsymbol{X}_i = \boldsymbol{x}$, il y a en face un individu non-traité j ayant des caractéristiques identiques ($\boldsymbol{X}_j = \boldsymbol{x}$). L'unité j est donc « jumelle » de i, modulo la valeur du traitement qui n'est pas la même.

Lorsque l'on apparie i et j, sur la base de $\boldsymbol{X}_i$ et $\boldsymbol{X}_j$ on le fait conditionnellement à $\boldsymbol{X}_i$ et $\boldsymbol{X}_j$.

Qui doit ressembler à qui (*who matches with whom*) dépend de l'effet causal que l'on souhaite estimer (ECMT, ECMnT, ECM). Par exemple, pour estimer l'ECMT, c'est une unité traitée que l'on tire en premier, et pour qui nous allons piocher une jumelle dans le groupe témoin.

Attention aux termes **contrôler**, **conditionner**, **ajuster**.

L'estimateur naturel de l'ECMT au point $X = x$ est la différence des moyennes conditionnelles en ce point :

$$E(Y|D = 1, X = x) - E(Y|D = 0, X = x)$$
$$= E(Y(1) - Y(0)|D = 1, X = x) + \underbrace{E(Y(0)|D = 1, X = x) - E(Y(0)|D = 0, X = x)}_{0}, \forall x$$

C'est l'ECMT plus le terme de BS dans la sous-population des $X = x$. Or, sous CIA, $E(Y(0)|D = 1, X = x) = E(Y(0)|D = 0, X = x) = 0$! Wooldridge (2003, 620).

CIA sur $Y(0)$ suffit pour estimer l'ECMT

Si on intègre par rapport à la loi conditionnelle $\Pr(X = x|D = 1)$; Lee (2016, 16, 30) :

$$\int [E(Y|D = 1, X = x) - E(Y|D = 0, X = x)]C(x|1)dx.$$

C'est comme pour la stratification, démontré dans la **sous-section 5.2.3**. À la différence que dans l'estimateur stratifié, X était catégoriel.

$$\sum_x [E(Y|D = 1, X = x) - E(Y|D = 0, X = x)]\, N_{1,x}/N_1.$$

À contrario, si c'est l'ECMnT qui nous intéresse, il faudra piocher dans le groupe test des individus semblables à ceux du groupe témoin (on conditionne sur les caractéristiques de ces derniers). À ce moment-là, CIA pour Y(1) suffit (**exercice 6.4.4**), et l'estimateur est :

$$\int [E(Y|D = 1, X = x) - E(Y|D = 0, X = x)]C(x|0)dx.$$

On comprend que l'appariement puisse être utilisé en combinaison à différentes méthodes de contrôle et d'estimation. Ces méthodes diffèrent essentiellement au niveau du calcul de $E(Y|D = 1, X = x) - E(Y|D = 0, X = x)$, qui peut être obtenu par régression par exemple (dans **`matching.do`** ; voir plus loin). Dans la sous-section « Regression meets matching », Angrist et Pischke (2009, 69-77) écrivent que la régression est un estimateur d'appariement pondéré. Mais différents estimateurs de $E(Y|D = d, X = x)$ donneront différents résultats ! Il n'y a que $E(Y|D = 1)$ et $E(Y|D = 0)$ qui seront pareil.

Comme toute méthode, les méthodes d'appariement exact ont leur <u>limites</u>.

<u>L'appariement est toujours inexact, pour deux raisons :</u>

Contrairement à ce que suggèrent les équations ci-dessus, l'appariement exact est impossible quand X est continue ; c'est trop fin. Il y a peu de chance de trouver deux individus avec le même x. On va travailler sur des intervalles, des strates.

Quand le nombre de caractéristiques dans $\boldsymbol{X}$ est important, il y a encore moins de chance de trouver deux individus pareils. On introduit une mesure de « distance entre les individus ».

Quel critère de ressemblance (proximité) utiliser en cas d'appariement inexact ? Nous allons avoir besoin d'une mesure précise, que l'on peut calculer assez rapidement, et qui ne soit pas trop compliquée (**distance euclidienne**, **distance de Mahalanobis**, par exemple).

Faut-il remettre un jumeau dans son groupe une fois pioché ? Si oui (TAR), l'ordre dans lequel on pioche les individus ne compte pas. Mais sinon (TSR), il compte.

> Supposons que deux individus traités $i = 1{,}2$ ont des caractéristiques telles qu'ils ont le même jumeau dans le groupe de contrôle. Dans un tirage sans remise, une fois ce jumeau pioché pour $i = 1$, je ne peux plus l'utiliser pour $i = 2$ ($i = 2$ ne peut pas être apparié).

Appariement sur la base de caractéristiques observables : nous allons comparer des individus sur la base des caractéristiques observées. Dans une EO, il y a probablement des caractéristiques non-observées pour lesquelles les individus ne sont plus appariés. Tant qu'il ne s'agit pas de facteurs de confusion, ce n'est pas problématique. Holland (1986) les appelle *attributes*.

> Contrôler autant de caractéristiques – de confusion – que possible.
>
> > Cette stratégie est rapide si l'appariement est via le score de propension, mais lente dans le cas de l'estimateur d'appariement.
>
> Des caractéristiques pas déterminées par le traitement (Wooldridge, 2005).

L'appariment revient à évaluer « localement » : contrairement à la régression, on extrapole peu (validé interne élevée), mais puisqu'il faut retenir des individus « proches », l'échantillon **représente** une population plus étroite (validité externe faible). Ce n'est pas juste un problème de biais de taille d'échantillon.

Recouvrement : le cas de recouvrement partiel est très courant (x' n'est pas dans le groupe $D = 0$, et x'' n'est pas dans le groupe $D = 1$; **l'exemple d'introduction**). On peut même envisager un recouvrement nul pour certaines variables au sens où quelle que soit la valeur de la variable, les individus ont une probabilité nulle d'être dans l'un des deux groupes. Dans ce cas, il n'y a rien à faire (Rubin, 1977). Accroître la taille de l'échantillon n'améliore pas le recouvrement.

[Donner l'exemple de l'EX **TICELEC**, « **`ticelec_veryshort.do`** »]

Remarque sur cet exemple : il y a un missing pour METER et DIST. Donc $60 - 1 = 59$ (N). On a $N_1 = 31$ et $N_0 = 28$.

$\Pr(D = 1|d) \neq 1 \Leftrightarrow \Pr(D = 0|d) \neq 0$, mais $\Pr(D = 1|w) = 0 \Leftrightarrow \Pr(D = 0|w) = 1$.

$31/41$	$10/41$ Mais en fait, il y en a que 3 qui auraient pu être traités, car 3 ont DIST < 20, et 7 ont DIST > 20.	$0/10$	$18/18$

6.2. Estimateur d'appariement de l'ECMT

L'estimateur d'appariement pour l'ECMT fait partie des *matching estimators* dans la littérature académique.

Nous allons voir l'appariement par paires.

Il s'agit d'un appariement au plus proche voisin (*nearest-neighbor matching*). Dès lors qu'il y a plusieurs facteurs de confusion, on n'a pas le choix, le concept mathématique d'égalité cède la place à celui de « voisinage », proximité. En effet, dans un ensemble à une dimension, $\mathbb{R}$ par exemple, comparer une caractéristique pour deux individus (genre, âge, nombre d'années d'études, redoublement, etc.), est très facile, mais comment comparer le vecteur (garçon, 23 ans, 5 années d'études supérieures) avec (fille, 22 ans, 5 années d'études) ?

Géométriquement, avec deux caractéristiques catégorielles, ça ressemble au dessin ci-dessous.

[Graphique « **matchingregions.png** »]

6.2.1. Appariement exact et inexact

On suppose que :

- Une fois un « jumeau » trouvé dans l'autre groupe, on ne le remet pas dans ce groupe (appariement par paire sans remise).
- Lorsque plusieurs individus sont *a priori* des jumeaux potentiels, on utilisera une métrique (la distance euclidienne) pour les départager.
  Il existe d'autres métriques : Mahalanobis, Minkowski, etc.

L'estimation d'appariement par paire de l'ECMT va consister :

1) pour chaque individu test, à piocher un « jumeau » dans le groupe témoin
2) calculer la différence de leur RO
3) calculer la moyenne de ces différences

L'estimation d'appariement par paire de l'ECMnT va consister :

1) pour chaque individu témoin, à piocher un « jumeau » dans le groupe test
2) calculer la différence de leur RO
3) calculer la moyenne de ces différences

L'estimation de l'ECM inclut les deux étapes (Imbens et Rubin, 2015).

Nous allons considérer une évaluation très fameuse en économie du travail. Elle porte sur la hausse du SMIC en 1992 dans l'État américain du New Jersey.

Après le New Jersey en 1992, c'est la Californie qui s'engagea à relever le salaire horaire de 50 % entre 2016 et 2018. Dans un article du Parisien du 29/03/2016, on peut lire « Le salaire horaire minimum va augmenter de 50 % … en Californie ».

L'article dit que les législateurs californiens ont conclu un accord pour augmenter le salaire horaire minimum. Dans les entreprises de plus de 26 employé/e/s, le SMIC devrait passer de 10 $/h en 2016 à 10,5 $/h en 2017, puis à 15 $ d'ici à 2022 (il est à 13 $/h actuellement en 2020). Les entreprises de moins de 26 employés devront mettre en place la mesure en 2023, i.e. avec un décalage d'une année. Pour le gouverneur démocrate de l'époque, Jerry Brown, cette hausse doit être assez 'flexible' pour pouvoir la suspendre si la conjoncture économique l'imposait.

[**Economie du travail : impact d'une hausse du SCMIC**]

Economie du travail : impact d'une hausse du SMIC
Card et Krueger (1994) ; Imbens et Rubin (2015, 404-)

La hausse du SMIC est une question récurrente dans les débats sur le chômage.

En Allemagne, le salaire minimum devrait augmenter d'un peu plus de 22 % d'ici 2022.

En France, des économistes comme Pierre Cahuc sont plutôt contre (ils lui préfèrent une combinaison prime d'activité + RSA), tandis que les économistes atterrés sont plutôt pour.

Dans un marché du travail où il existe un salaire minimum, **est-ce qu'une hausse de ce salaire, décrétée par l'État, accroît le chômage ?** L'article est par là, si ça vous intéresse.

[Article « **1994_CardKrueger.pdf** »]

La Commission a produit une illustration du travail de David Card et Alan Krueger : The Effect of Increasing the Minimum Wage.

L'article montre qu'augmenter le salaire minimum n'accroît pas le chômage.

Les auteurs ont répondu à cette question pour les États-Unis dans un marché particulier : celui de la restauration rapide dans l'État du New Jersey où le premier avril 1992, le salaire minimum passa de 4,25$ de l'heure à 5,05$.

Unités : restaurants de type fast-food ; il y a quatre chaînes de restaurants (Burger King, KFC, Roys, Wendys)

Traitement : hausse du salaire minimum. La variable expérimentale est $D_i = 1$ si le restaurant est dans l'Etat du New Jersey après le premier avril 1992 où le salaire minimum augmente, 0 si c'est un restaurant en Pennsylvanie durant la même période.

Traité : chômage. La variable du résultat expérimental est l'emploi post-intervention.

Après un peu de nettoyage de la base, il y a 6 variables sans missing et 347 unités d'échantillonnage :

279 dans le groupe test
68 dans le groupe témoin

Remarque : il y a plus de variables (46) que celles que l'on a retenues, notamment la part des salariés dont le salaire horaire est supérieur au salaire minimum.

6.2.2. L'évaluation par Card et Krueger (1994) de la hausse du SMIC

La variable emploi final (observée pour novembre 1992) est le résultat Y. C'est le nombre d'employés à temps-plein plus ½ du nombre d'employés à mi-temps. L'emploi initial, qui est observé pour février 1992, avant la hausse du SMIC, est construit de la même façon.

[visualiser les données de départ, « **`cardkrueger1994.R`** »]

Dans cette introduction à la méthode nous allons travailler sur le sous-ensemble des données de Card et Krueger (1994) sélectionné par Imbens et Rubin (2015). L'intérêt est double :

- Afin de bien cerner les calculs, pas besoin d'un grand échantillon.
- Les données originales de Card et Krueger (1994) ont l'inconvénient d'inclure relativement peu d'individus témoins (voir l'encadré). Or, dans le cas d'un appariement par paire sans remise, il vaut mieux que le réservoir d'individus témoins (on veut estimer l'ECMT) soit suffisamment grand pour que nous soyons sûrs de trouver un jumeau pour chaque individu test.

[Insérer feuille Excel des 20 restaurants « **2021_CardKrueger.xlsx** »]

Notons que dans les données originales de Card et Krueger (1994), on n'observe pas de déséquilibre sur la variable d'emploi initial, mais peut-être un léger sur la variable catégorielle de chaîne de restaurant

- Parmi les 279 restaurants traités, il y en a certains dont la valeur de la variable d'emploi initial est éloignée de la moyenne de l'échantillon (des individus traités ou pas), de sorte que l'on sait déjà qu'il va être difficile de les apparier.
- C'est corroboré par une mesure du recouvrement aux extrêmes. On remarque que le plus petit des restaurants du NJ a 1,5 salarié de moins (avant intervention) que le plus petit de PA. En revanche, le plus grand a 12,5 salariés de plus.
  Donc, on sait déjà qu'il existe des valeurs de l'emploi initial telles que je n'ai que des individus tests (au moins un) et que des individus témoins (*idem*)

Dans le sous-échantillon d'Imbens et Rubin (2015) il y a :
- 20 restaurants : 5 du New Jersey et 15 de Pennsylvanie
- deux chaînes de restaurant : Burger King (BK) et KFC
- deux facteurs de confusion (la chaîne de restaurant et l'emploi initial)

Nous allons numéroter toutes les étapes de l'estimation,
et supposer momentanément un tirage avec remise ⟵

1) On prend le premier restaurant, un BK du NJ, où l'emploi initial est 22,5
2) Quel est son jumeau ?
   Remarquons qu'il y a 11 BK et 4 KFC (resp. 3 et 2) dans le groupe de contrôle (test). Il y a recouvrement pour la variable de chaîne de restaurant.
3) Il y a deux candidats : 25,5 employés chez l'individu 9 et 20 employés chez l'individu 11 ; $\mathbf{20 \neq 22,5 \neq 25}$ (appariement inexact)
4) En différence absolue ou en logarithme (afin de corriger les problèmes d'échelle), c'est l'individu 11 qui est plus proche (appariement au plus proche voisin) :

$$|25{,}5 - 22{,}5| = 3 \; > \; |20 - 22{,}5| = 2{,}5$$

$$|3{,}238 - 3{,}113| = 0{,}125 \; > \; |2{,}995 - 3{,}113| = 0{,}117$$

5) Calcul de l'effet causal pour $\boldsymbol{i = 1}$: $\boldsymbol{y_1 - y_{11} = 40{,}0 - 19{,}5 = 20{,}5}$ ($\hat{\tau}_{1,1}^{\mathbf{appa}}$).
6) On reproduit les étapes 1 à 5 pour les individus 2 à 4 …

… le dernier individu est l'individu 5, un KFC, où l'emploi initial est 8,0

7) Aucun KFC témoin n'a l'emploi initial à 8,0. Il y a 8,5 (individu 8). En revanche, il y a un BK à 8,0 (individu 20). Dilemme !
8) Prenons la distance euclidienne au carré comme mesure de proximité (Imbens et Rubin, 2015, 413). On code la chaîne de restaurant $\mathbf{KFC} = \mathbf{1}, \mathbf{BK} = \mathbf{0}$. Focalisons-nous dans le petit tableau suivant sur les deux individus 8 et 20, qui sont nos candidats pour un appariement à l'individu 5.

Tableau 6.2 : codage numérique des variables

$\boldsymbol{i}$	État ($\boldsymbol{D}$)	Emploi ini. ($\boldsymbol{X_1}$)	Chaîne ($\boldsymbol{X_2}$)	(Chaîne $\boldsymbol{X_2}$ codé)
5	NJ	**8,0**	KFC	**1**
8	PA	**8,5**	KFC	**1**
20	PA	**8,0**	BK	**0**

On note $\boldsymbol{x_{i,k}}$ la valeur de la variable $\boldsymbol{X_k}$ pour l'individu $\boldsymbol{i}$, avec $\boldsymbol{k = 1}$ (l'emploi initial) et $\boldsymbol{k = 2}$ (la chaîne). Alors,

$$x_5 := (x_{5,1}; x_{5,2}) = (8{,}0; 1), \text{ et } x_8 := (x_{8,1}; x_{8,2}) = (8{,}5; 1),\ x_{20} := (x_{20,1}; x_{20,2}) = (8{,}0; 0).$$

Donc, $\boldsymbol{x_5 - x_8} = (\mathbf{-0,5; 0})$ et $\boldsymbol{x_5 - x_{20}} = (\mathbf{0; 1})$.
Par conséquent, $\|\boldsymbol{x_5 - x_8}\|^2 = (\mathbf{-0,5})^2 + \mathbf{0^2} = \mathbf{0,25} < \|\boldsymbol{x_5 - x_{20}}\|^2 = \mathbf{0^2} + \mathbf{1^2} = \mathbf{1}$. L'individu 8 est le plus proche.

Si nous avions codé KFC $= 0$ et BK $= 1$, nous aurions la même sélection.

9) L'effet causal pour l'individu traité 5 vaut $\hat{\boldsymbol{\tau}}_{\mathbf{1,5}}^{\mathbf{appa}} = \boldsymbol{y_5 - y_8} = \mathbf{5,5 - 10,5 = -5}$.
10) **Arrêtons-nous ici**.

L'estimation d'appariement de l'ECMT est $5^{-1}\left(\sum_{i=1}^{5} \hat{\tau}_{1\,i}^{\text{appa}}\right) \equiv \hat{\tau}_1^{\text{appa}}$, qui vaut 0,3. L'**exercice 6.4.2** reprend cette évaluation avec un appariment sans remise (TSR) ou *greedy matching*. Et si on ajoute la contrainte que les distances ne doivent pas dépasser un certain seuil, on parle de *caliper matching* ; voir Moczall (2014, 51-52).

Remarques :

1) Le critère de distance devrait normaliser les observations afin qu'il n'y ait ni problème de translation, ni d'échelle (Imbens et Rubin, 2015, 410). Par exemple, supposons que la métrique soit toujours euclidienne mais avec un changement d'unité de mesure (on passe de l'effectif salarié à un nombre d'heures travaillées, chaque salarié travaillant en moyenne 10h par jour. La mesure de proximité entre $\boldsymbol{i}$ et $\boldsymbol{j}$ devient $\|(\mathbf{10}\boldsymbol{X_{i,1}}; \boldsymbol{X_{i,2}}) - (\mathbf{10}\boldsymbol{X_{j,1}}; \boldsymbol{X_{j,2}})\|^2$. Alors,

$$\|\boldsymbol{x_5 - x_8}\|^2 = \|(\mathbf{-5; 0})\|^2 = \mathbf{25} > \|\boldsymbol{x_5 - x_{20}}\|^2 = \|(\mathbf{0; -1})\|^2 = \mathbf{1}.$$

   L'estimation de l'ECMT dans ce cas est $\mathbf{-0,4}$. Il est donc judicieux de centrer et **de réduire** les $\boldsymbol{X}$ avant appariement (c'est ce que font les commandes Stata, comme `nnmatch`, par exemple ; voir plus loin).
2) L'ordre suivant lequel on sélectionne les individus traités à apparier est crucial dans un TSR. L'estimation de l'ECMT avec un ordre différent (1, 2, 3, 5, 4) est 0,8. Ces différents résultats pour l'estimation de l'ECMT montrent que l'appariement séquentiel n'est pas optimal. Il existe une approche d'appariment optimal.

[Page 414 d'Imbens et Rubin (2015)]

6.2.3. Implémentation de l'estimateur dans Stata

Il existe différents programmes Stata et R implémentant des estimateurs d'appariement. Je retiens ceux pour Stata.

La première implémentation de ce type d'estimateurs pour Stata est le programme **pscore.ado** de Becker et Ichino (2002), **psmatch2.ado** de Leuven et Sianesi (2003) et **nnmatch.ado** d'Abadie, Drukker, Herr et Imbens (2004). Nichols (2007, 2008) montre comment utiliser les deux derniers programmes dans un travail économétrique. Mis à part Leuven et Sianesi (2003), ces programmes sont accompagnés d'un article dans *Stata Journal*.

On va appliquer **nnmatch.ado** à (i) la version simplifié de l'évaluation de Card et Krueger (1994), afin de vérifier que nous trouvons bien 0,3 pour $\hat{\tau}_1^{\text{appa}}$, l'estimation de l'ECMT. La commande propose aussi un test de nullité de l'ECMT, ce qui suppose le calcul de l'erreur-type de $\hat{\tau}_1^{\text{appa}}$. Puis, (ii) nous verrons l'exemple construit par Abadie, Drukker, Herr et Imbens (2004) pour comprendre l'estimateur d'appariement sur lesquels s'appuie **nnmatch**.

Nous nous servirons de **nnmatch** pour illustrer l'estimation des trois effets causaux, l'ECM, l'ECMT et l'ECMnT. Il vous sera facile d'appliquer la commande à vos données. On détaille les estimateur des effets causaux, pas les variances. Les enjeux sont assez importants, je vous invite à voir la section 4 de Abadie, Drukker, Herr et Imbens (2004). Pour plus de théorie, vous pouvez consulter Abadie et Imbens (2011). L'estimateur utilise une correction qu'on ne détaillera pas dans le cours. Dans les exemples (i) et (ii), nous effectuons un appariement par paires, au plus proche voisin ; un seul voisin !

[**matching.do**]

(i) *Card et Krueger*

On utilise la commande :

```
nnmatch JOBFIN D CHAIN JOBINI, ///
        tc(att)    m(1)    exact(CHAIN)    keep(matchingsave,
replace)
```

(ii) *Exemple artificiel d'Abadie et alii (2004)*

Cet exemple ne comporte qu'une seule variable X supposée confondante.

Tableau 6.3 : estimateur d'appariement avec $N = 7$ et un jumeau ($M = 1$)

i	D_i	X_i	Y_i	$J_M(i)$	$\frac{1}{\lvert J_M(i)\rvert}$	$\hat{Y}_i(0)$	$\hat{Y}_i(1)$	$K_M(i)$
1	0	2	7	{5}	1	7	8	3
2	0	4	8	{4 ; 6}	1/2	8	7,5	1
3	0	5	6	{4 ; 6}	1/2	6	7,5	0
4	1	3	9	{1 ; 2}	1/2	7,5	9	1
5	1	2	8	{1}	1	7	8	1
6	1	3	6	{1 ; 2}	1/2	7,5	6	1
7	1	1	5	{1}	1	7	5	0

8/1
(9+6)/2
(9+6)/2
(7+8)/2
7/1
(7+8)/2

Pour chaque individu, on peut avoir plus d'une jumelle (colonne 5), ce qu'on appelle un *tie* en anglais (deux *nearest neighbors*). S'il n'y a pas de tie, $J_M(i) = J_1(i)$ n'a qu'un élément.
Par exemple, l'individu test $i = 6$ ($X = 3$) a deux témoins : $j = 1$ ($X = 2$) et $j' = 2$ ($X = 4$) à égales distances (la val. abs. de la différence des X) de i. Alors, $|4 - 3| = |2 - 3| = 1$. Pour éviter un *tie*, on peut comme avant prendre ln(X).

> Qu'est-ce que ça donnerait pour $J_1(6)$? $|ln(4) - ln(3)| = 0{,}28 < |ln(2) - ln(3)| = 0{,}40$. On rejette le témoin $j = 1$. On aurait $J_1(6) = \{2\}$, $1/|J_1(6)| = 1$.
> Le cas $J_M(2)$ est différent. Il y a deux témoins/voisins ($i = 4{,}6$) sans que l'on ne puisse trancher entre eux car $X = 3$ dans les deux cas.

La théorie est contenue dans Abadie et Imbens (2011), Abadie, Drukker, Herr et Imbens (2004). On a le choix d'une formalisation de l'estimateur d'appariement. Cette formalisation a évolué avec Imbens et Rubin (2015) où les auteurs semblent proposer un cadre unifié des différents cas (TAR, TSR, exact, inexact, correction du biais). Abadie et Imbens (2011) est plus proche de l'ouvrage que Abadie, Drukker, Herr et Imbens (2004).

On note $I \equiv \{1, \dots, N\}$, $I_0 \equiv \{i : D_i = 0\}$ et $I_1 \equiv \{i : D_i = 1\}$, et

$d_M(i) : \sum_{j:D_j=1-D_i} \mathbb{1}_{[0,d_M(i)]}\left(\|X_i - X_j\|_V\right) < M$ et $\sum_{j:D_j=1-D_i} \mathbb{1}_{[0,d_M(i)]}\left(\|X_i - X_j\|_V\right) \geq M$.

Au moins M individus du groupe opposé à celui de i sont appariés à i. Par exemple, pour l'individu test $i = 6$ discuté ci-dessus, la distance aux deux témoins $j = 1$ et 2 vaut $d_M(6) = 1$ avec la norme valeur absolue, et $M = 1$. Le premier terme inclut aucun témoin : M valant 1, une somme de témoins plus petite est forcément nulle, et la distance doit être plus petit que 1.

$J_M(i) = \{j \in I | D_j = 1 - D_i, \|X_i - X_j\|_V \leq d_M(i)\}$. C'est l'ensemble des individus appariés à i.

$K_1(i)$ est le nombre total de fois que i est utilisé comme 'contrôle'. On applique une pondération (la colonne $1/|J_1(i)|$) qui tient compte du fait que i n'est pas l'unique témoin d'un individu test. Par exemple, $K_1(1) = 3$ car $i = 1$ est utilisé comme témoin pour $j \in \{4,5,6,7\}$. Mais chaque fois qu'il y a eu un *tie*, i n'a compté qu'à moitié. On a :

$$K_1(1) = \frac{1}{|J_1(4)|} + \frac{1}{|J_1(5)|} + \frac{1}{|J_1(6)|} + \frac{1}{|J_1(7)|} = \frac{1}{2} + 1 + \frac{1}{2} + 1.$$

Autre cas : $K_1(5) = 1$. En effet, $i = 5$, qui est dans le groupe test, n'a été apparié à un témoin qu'une fois, à $j = 1$, donc $K_1(5) = \frac{1}{|J_1(1)|} = \frac{1}{1} = 1$.

On a donc :

$$K_1(i) = \sum_j \frac{\mathbb{1}_{J_1(j)}(i)}{|J_1(j)|}$$

peut remarquer que $\sum_{i:D_i=1} K_1(i) = N_0$. Notons $\mathbb{1}(i,j)$ l'indicatrice valant 1 si i est utilisé comme individu d'appariement de j, et 0 sinon.

$\sum_{i:D_i=1} K_1(i) = \sum_{i:D_i=1} \sum_{j:D_j=0} \mathbb{1}(i,j)$.
$\sum_{i:D_i=1} K_1(i) = \sum_{i:D_i=0} |J_1(i)|/|J_1(i)|$.

En effet, le nombre de fois que chaque individu test est utilisé, c'est pour être apparié aux individus témoins. On peut donc plutôt parcourir chaque individu témoin (il y en a N_0), et pour chacun prendre le cardinal de $J_1(i)$, qui est le nombre d'individus tests

appariés à i, $|J_1(i)|$, et compte tenu du fractionnement, diviser par $|J_1(i)|$. Symétriquement, on a $\sum_{i:D_i=0} K_1(i) = N_1$.

[Compléter les colonnes $\hat{Y}_i(0)$ et $\hat{Y}_i(1)$]

A priori, l'EMT $\bar{Y}_1 - \bar{Y}_0$, est un estimateur biaisé de l'ECM. Il vaut :

$$\frac{1}{4}(9+8+6+5) - \frac{1}{3}(7+8+6) = 7 - 7 = 0.$$

Or, on peut voir que $\bar{X}_1 - \bar{X}_0 = 2{,}25 - 3{,}67 = -1{,}41$, et la différence normalisée, 1,27, dépasse 1 ! Il y a déséquilibre et recouvrement partiel ; normal, c'est un mini-échantillon !

Les estimateurs alternatifs, qui tiennent compte du facteur confondant, X, sont :

Celui de l'ECMT, qui repose sur l'estimation du résultat contrefactuel pour les individus tests s'ils n'avaient pas été traités,

$$\begin{aligned}\hat{\tau}_1^{appa} &= \bar{Y}_1 - \frac{1}{N_1}\sum_{i:D_i=1}\hat{Y}_i(0),\\ &= \bar{Y}_1 - \frac{1}{N_1}\sum_i (1-D_i)K_M(i)Y_i.\end{aligned}$$

L'estimateur de l'ECMnT est plus rarement recherché, mais n'est pas très pertinent dès lors que des individus témoins n'ont aucune chance d'être traités (Wooldridge, 2003, 604) :

$$\begin{aligned}\hat{\tau}_0^{appa} &= \frac{1}{N_0}\sum_{i:D_i=0}\hat{Y}_i(1) - \bar{Y}_0,\\ &= \frac{1}{N_0}\sum_i [D_i K_M(i) - (1-D_i)]Y_i - \bar{Y}_0.\end{aligned}$$

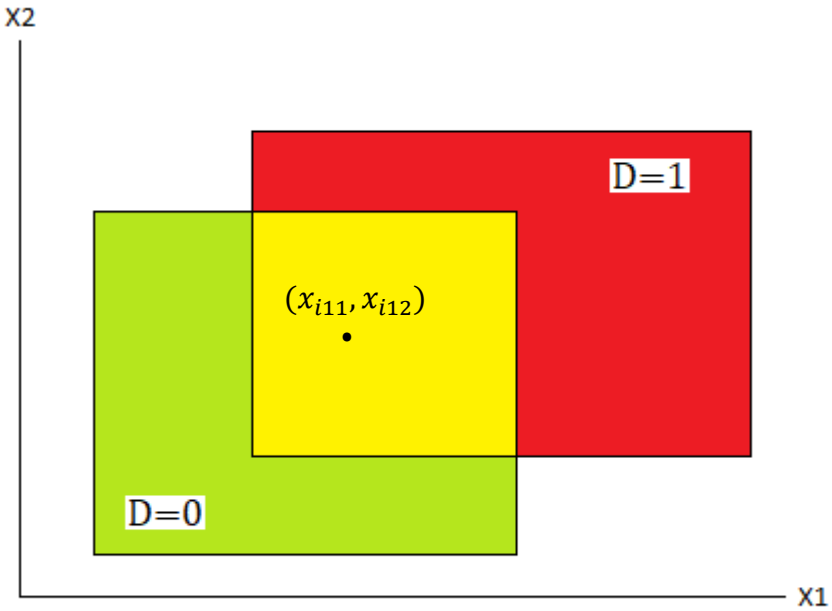
Et enfin, l'estimateur de l'ECM :

$$\begin{aligned}\hat{\tau}^{appa} &= \frac{1}{N}\sum_i(\hat{Y}_i(1) - \hat{Y}_i(0)),\\ &= \frac{1}{N}\sum_i(2D_i - 1)(1 + K_M(i))Y_i.\end{aligned}$$

6.2.4. Grossissement du maillage des X

Que peut-on faire lorsque l'on a une ou plusieurs variables d'appariement continues ? Va-t-on faire une comparaison pour chaque point du recouvrement ? Non.

Figure 6.1 : appariement exact : diminution de la région de recouvrement



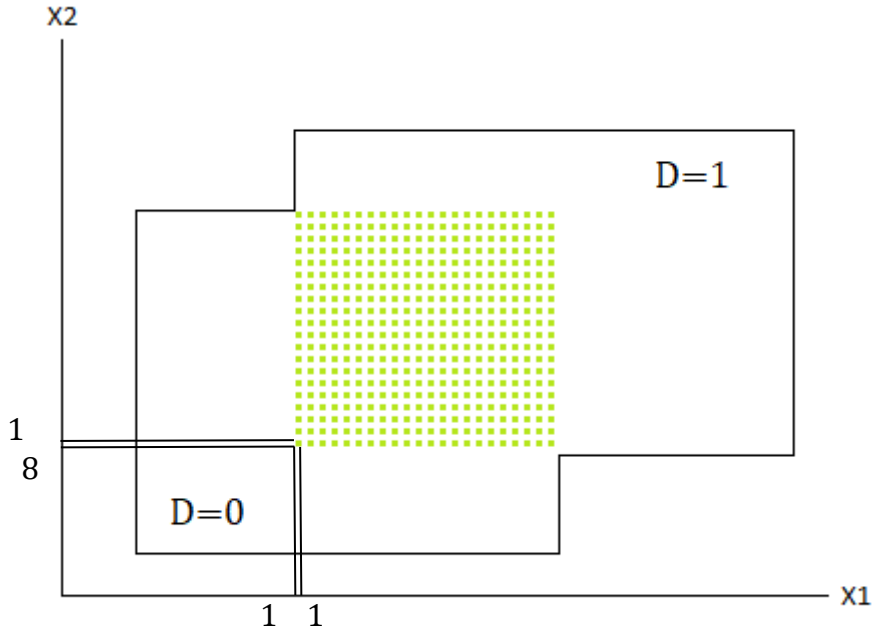
Note : x_{idk} est la valeur de la variable k pour l'unité i du groupe de traitement $D_i = d$

Dans la Figure, les valeurs pour le couple de variable (X_1, X_2) chez les traités $(D = 1)$ correspond au rectangle rouge. Et, les valeurs pour couple de variable (X_1, X_2) chez les non-traités $(D = 0)$ correspond au rectangle vert.

Même partiel, le recouvrement est tel qu'il y a trop de points qui devraient rentrer dans l'appariement.

Une solution consiste à mailler la région ; on parle de *coarsening* : rendre la région plus grossière, moins fine.

Figure 6.2 : *coarsening* de la région de recouvrement.



L'appariement sera possible en chacun des « points » verts. Dans cette figure, il y en $22 \times 22 = \mathbf{484}$.

> Nous pouvons réaliser ce type maillage avec la commande **autocode** de Stata.
>
> Il y a 22 valeurs pour X_1, par exemple $(10,12], (12,14]$, … Et pour X_2, $(8,10], (10,12]$, …
>
> Tous les individus $i: D_i = 1$ et qui sont tels que $X_1 \in (10,12]$ et $X_2 \in (8,10]$ seront appariés avec les individus $j: D_j = 0$ et qui sont tels que $X_1 \in (10,12]$ et $X_2 \in (8,10]$. Cela fait quand même beaucoup d'appariements.

Mais, cette approche a ses limites quand le nombre de variables d'appariement est important. À cause essentiellement du problème de non-recouvrement. En effet, plus il y a de variable, plus cette région rétrécit, comme on peut le voir dans la figure précédente. Forcément, le nombre d'individus considérés pour l'évaluation aussi va aussi se réduire.

D'où une préférence pour un autre approche dans ce cas, l'appariement via la score de propension. Le score de propension va servir à : *propensity score matching*

> Vérifier l'équilibrage et le recouvrement dans l'intervalle des valeurs du score retenues pour comparer les résultats des groupes de traitement
>
> Estimer des effets causaux à l'intérieur de chaque segment/intervalle (il s'agira en fait de strates, comme pour la stratification)
>
> Mais aussi pour faire de l'élagage : triming

6.3. Estimateur de l'ECMT après équilibrage par le score de propension (SP)

On se replace à l'étape du protocole (*design*). La création d'un sous-échantillon plus robuste pour l'estimation de l'ECMT par appariement via le SP, fait également partie du protocole, et peut être menée juste après la constitution de l'échantillon représentatif de la population.

Même dans des protocoles randomisés.

Le score de propension (**SP**) (*propensity score*) est une probabilité conditionnelle. Dans les applications, c'est sa valeur prédite qui nous intéresse. Dans le cas de deux groupes de traitements, on considère, pour chaque individu, l'évènement discret : être dans le groupe test ou pas. On va s'intéresser à la distribution de probabilité de cet événement conditionnellement aux variables de confusion X, les mêmes que celles que nous utiliserions pour construire l'estimateur d'appariement :

$$\Pr(D = 1|X) \coloneqq e(X).$$

La méthode remonte à Rosenbaum et Rubin (1983), qui ont proposé le théorème du score de propension. Dehejia et Wahba (2002) ont un algorithme d'estimation du score. Imbens (2004) revisite le théorème précédent, ainsi que Angrist et Pischke (2009). On peut aussi consulter Becker et Ichino (2002).

Ce n'est pas un théorème facile car le conditionnement est par rapport à $e(X)$. On conditionne par rapport à une probabilité, aléatoire du coup ...

Cette probabilité conditionnelle sert à apparier les individus, l'idée étant que les individus dont la probabilité – conditionnelle – d'être traités est proche de 1 et ceux dont la probabilité – conditionnelle – de ne pas être traités est proche de 1 (d'être traités proche de 0) ont des caractéristiques trop éloignées pour être comparés. En revanche, les individus dont la probabilité – conditionnelle – d'être traité ou pas est dans un intervalle $[\underline{e}\,;\,\overline{e}]$, ont forcément des caractéristiques proches.

L'hypothèse de recouvrement (*overlap*) dont nous avions déjà parlé prend tout son sens.

Au lieu de vérifier cette hypothèse, variable de confusion par variable de confusion, on la vérifie que pour le score !

$$0 < \Pr(D = 1|\boldsymbol{X}) < 1. \qquad (6.1)$$

Imbens et Rubin (2015) appellent ce type de MAT, un MAT probabiliste.

Quelle que soit la valeur de $\boldsymbol{X}$, « il y a au moins un » individu test. On peut vérifier qu'elle implique la même chose pour les individus témoins : quelle que soit la valeur de $\boldsymbol{X}$, « il y a au moins un » individu témoin (et tous les individus ne sont pas témoins). En effet, multiplions par -1 et ajoutons 1, **(6.1)** est équivalent à $0 < \Pr(D = 0|\boldsymbol{X}) < 1$. Sinon, la vraisemblance pour un **MAT non-confondu** serait nulle !

$$\Pr\left(D_1 = d_1, \dots, D_N = d_N | x, \cancel{y(0)}, \cancel{y(1)}\right) \propto \Pi_{i=1}^{N} e(x_i)^{d_i}\left(1 - e(x_i)\right)^{1-d_i}.$$

6.3.1. Modèles pour le score de propension : logit, probit, ...

Le modèle statistique le plus utilisé pour estimer $\Pr(D = 1|\boldsymbol{X})$ est le modèle logit.

Nous pourrions considérer un probit (et un multinomial si le nombre de groupes de traitements dépassait 2). Nous allons nous focaliser sur le logit.

L'événement que nous étudions est l'appartenance à un groupe de traitement. Selon le problème économique que nous évaluons, l'évènement est une *qualité*. Par exemple,

- « suivre la formation », « ne pas suivre la formation »
- « être localisé dans le New Jersey », « être localisé en Pennsylvanie »
- « percevoir un crédit d'impôt », « ne pas en percevoir »

Ou, une *quantité* :

- « avoir trois années d'études post-baccalauréat (Licence) », « avoir cinq années d'études post-baccalauréat (Master 2) ».

Dans ces exemples, on a un choix binomial. Le cas multinomial (le nombre de groupes de traitement peut être supérieur à deux, que la variable aléatoire soit associée à un événement qualitatif ou quantitatif) :

- « CIR », « subvention », « CIR et subvention », « rien »
- « avoir un enfant », « avoir deux enfants », « avoir un enfant puis des jumeaux »

Dans le premier exemple, les choix ne sont pas ordonnés (Greene, 2008, 811-), contrairement au second exemple. Dans le dernier exemple, le passage de la modalité « avoir un enfant » à « avoir deux enfants » ou « avoir un enfant puis des jumeaux » relève d'un choix de l'individu et de la nature. Cf. **le chapitre** où j'ai évoqué cette Expérimentation Naturelle.

La modélisation de ces événements utilise une variable (dépendante) discrète qui prend les valeurs 0, 1 quand nous avons deux groupes. Ces modèles, dont la variable dépendante est discrète, appartiennent à la famille des modèles à variable dépendante limitée (*limited dependent variable models*), au sens où entre les valeurs de D, ici 0 ou 1, il n'y a rien. Entre 0 et 1 il n'y a pas d'évènement, ni à gauche de 0, ni à droite de 1.

Avec trois groupes de traitement ou plus, nous considérons alors les valeurs 0, 1, 2, 3, … Il n'y a pas de valeur entre 1 et 2, 2 et 3 et à droite de 3. Je vous renvoie à votre **cours d'économétrie des variables qualitatives**. Une idée importante est que :

L'estimation de première étape ne vise pas à modéliser un comportement (contrairement aux études de choix en économétrie), mais à équilibrer les groupes de traitement (Imbens et Rubin, 2015, 284) ; Dehejia et Wahba (2002, 161).

Les méthodes d'appariment via le score de propension s'appuient sur l'estimation de la probabilité que $D_i = 1 \vert \boldsymbol{X}_i$ pour tous les i.

6.3.2. Théorème du score de propension : conditionner sur le SP atténue le BS

L'intérêt de la méthode (comparé aux estimateurs d'appariement de la **section 6.2**) est qu'en présence de beaucoup de variables de confusion, on peut atténuer le BS en ne conditionnant que sur une variable, $e(\boldsymbol{X})$. Pour être plus précis, quand on écrit la CIA, $(Y(1), Y(0)) \perp D \vert \boldsymbol{X}$, on est obligé de considérer tous les $\boldsymbol{X}$, un à un (c'est ce qu'on a vu avec l'appariement au plus proche voisin). En s'appuyant sur la CIA, le théorème du score de propension de Rosenbaum et Rubin (1983, 43) nous dit que

$$(Y(1), Y(0)) \perp D \vert e(\boldsymbol{X}).$$

et $e(\boldsymbol{X})$ est un score d'équilibrage.

Evidemment, $\boldsymbol{X}$ est aussi un score d'équilibrage. C'est même le plus fin ! $e(\boldsymbol{X})$ est le plus grossier (*the coarsest*) ; Rosenbaum et Rubin (1983, 43). La démonstration du théorème dans Angrist et Pischke (2009, 81) est celle de $\Pr(D = 1 \vert Y(d), e(\boldsymbol{X})) = e(\boldsymbol{X}), \forall\, d \in \{0; 1\}$.

6.3.3. Applications

Sur le sous-échantillon d'Imbens et Rubin (2015) des données de Card et Krueger (1994)

Appariement d'équilibrage (recouvrement) : *triming*

Vérification graphique (on plot les distributions) de l'équilibrage sur l'emploi initial et la chaîne

- Concernant l'emploi initial, le triming via le score permet de resserrer un peu les distributions en virant des observations dans les queues (trop à gauche et trop à droite), mais il y a trop peu d'observations dans le New Jersey (de 5 à est passé à 3) pour que les distributions s'épousent.
- Concernant les chaînes de magasins, le résultat après appariement sur le score permet d'obtenir la même proportion de BG dans chaque État. Parfait !

Sur les données originales de Card et Krueger (1994)

Appariement d'équilibrage (recouvrement) : *triming*

Vérification de l'équilibrage directement au niveau du logit (« score linéarisé »)

Sur un sous-échantillon (Dehejia et Wahba, 2002) des données de Lalonde (1986) portant sur l'évaluation d'un programme de formation. Dehejia et Wahba (2002) suivent la même stratégie de test que celle de Lalonde (1986) : utiliser le groupe de contrôle randomisé du programme afin d'obtenir une évaluation *benchmark* de l'effet du programme. Puis, remplacer ce groupe par un groupe issu d'une enquête. On reproduit des résultats figurant dans Imbens et Rubin (2015, p. 144-145). L'échantillon est le même que celui de Dehejia et Wahba (2002), mais avec variables de salaire nul quand re74=0 ou re75=0, et l'unité de mesure du salaires est le millier. Plusieurs étapes d'évaluation :

- Investigation du déséquilibre des groupes
- Comparaison des variables une à une avec **`estpost tabstat`** … et **`esttab`**
- Distance de Mahalanobis, notée M :

$$\frac{1}{2}\left(\frac{\sum_{i:D_i=0}(\boldsymbol{X}_i-\overline{\boldsymbol{X}_0})\,(\boldsymbol{X}_i-\overline{\boldsymbol{X}_0})'}{N_0-1}+\frac{\sum_{i:D_i=1}(\boldsymbol{X}_i-\overline{\boldsymbol{X}_1})\,(\boldsymbol{X}_i-\overline{\boldsymbol{X}_1})'}{N_1-1}\right)\equiv\widehat{\boldsymbol{\Sigma}}$$

$$M\equiv(\overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_0)'\widehat{\boldsymbol{\Sigma}}^{-1}(\overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_0).$$

Dans le cas d'une variable, M est le carré de la différence normalisée !

- Histogramme et différence normalisée pour le score linéarisé dans le cas avec groupe de contrôle randomisé et non-randomisé
- Test de l'hypothèse d'indépendance conditionnelle
- Estimation de l'effet causal avec les deux types de groupes de contrôle

L'ECMT, « … *the primary treatment effect of interest in nonexperimental settings …* » (Dehejia et Wahba, 2002, 152). **En appariant (*via* le score de propension) les individus tests et témoins sur la base des caractéristiques des premiers, Dehejia et Wahba parviennent à s'approcher de l'effet causal**.

C'est une confirmation empirique du théorème du score de propension évoqué dans la **sous-section 6.3.2**.

[**Encadré** de l'évaluation de Lalonde (1986)]

The National Supported Work* (NSW) *Program

Sous-échantillon dans Dehejia et Wahba (2002) de Lalonde (19
Aussi utilisé dans le ch. 8.2 d'Imbens et Rubin (2015) et Imbens (2015) ; voir également Angrist et Pischke (2009).*

Le NSW est un programme temporaire de formation professionnelle de personnes au chômage et exposées à des difficultés supplémentai (contraintes familiales, problèmes de santé, casiers judiciaires, etc.). La formation dure de neuf à 12 mois. Le NSW s'appuie sur un protocole randomisé (le groupe témoin ne bénéficie d'aucune formation, ni d'assistance). Lalonde (1986, 605) note que c'est la première évaluation d'un programme randomisé de ce type.

L'objet de l'étude de Robert Lalonde est double : évaluer le programme (estimer son effet) et confronter deux approches de l'évaluation :

(i) L'approche économétrique standard à l'époque, sans groupe de contrôle randomisé, mais tiré du *Current Population Survey* (CPS).** Une **spécification économétrique transpose une théorie économique**. Il y a des **paramètres structurels** et **réduits**.

(ii) L'approche avec groupe de contrôle randomisé, plus coûteuse (autour de 30 000 € de 2019), mais s'appuyant sur une spécification et des estimateurs plus simples (simple différence, différence de différences ; cf. **chapitre 8**).

Lalonde part du postulat que (ii) fournit des évaluations crédibles car la randomisation assure un appariement des facteurs (programmes auxquels les individus ont pu participer avant, et autres facteurs non-observables de confusion). La question est de savoir si (i) est capable de fournir une évaluation aussi crédible du programme en question (Imbens et Rubin, 2015, 328). L'évaluation dans (i) est sensible au choix d'un groupe de contrôle et à la spécification du modèle, essentiellement le choix des variables de contrôle (Lalonde, 1986, 614). Pour l'auteur, le meilleur test de spécification consiste à comparer les estimations avec les différents groupes de contrôle et les différentes spécifications dans (i) non pas entre elles, mais à celles obtenues avec (ii). Heckman et *alii* (1987) pointent des erreurs de Lalonde. La controverse randomisé/structurel continue (Heckman, 2020).

Unités d'observation : On distingue quatre catégories de personnes accompagnées

Femmes bénéficiant de l'aide pour les familles ayant un enfant à charge, des hommes et femmes ayant été soignées pour des addictions, jugées pour des agressions, …, enfants déscolarisés, etc. Dans la version simplifiée, on n'a que des hommes.

Il y eu des critères d'éligibilité à l'EX supplémentaires.

Une fois dans l'ensemble d'éligibilité, les individus sont affectés aléatoirement dans les groupes test et témoin.

Le traitement : formation ($D = 1$) ou pas ($D = 0$)
La variable de résultat : le salaire de 1978 ($)
Les variables de pré-traitement : âge (entier), #années d'études (entier), déjà marié (1/0), pas de diplôme de niveau lycée + (1/0), afro-descendant (1/0), hispanique (1/0), salaire de 1975 ($), 1974 ($), indicatrice de salaire nul en 1975 (1/0), et en 1974 (1/0)

Relevées au moment de la randomisation (*at the baseline*), puis tous les neuf mois

Dans le sous-échantillon de Dehejia et Wahba (2002), les effectifs des hommes répartis dans les groupes de traitement sont : $N_1 = 185$, $N_{0,NSW} = 260$. Et nous avons $N_{0,CPS} = 15992$.

Les salaires sont bas, de l'ordre de 1700 $ avant traitement, et autour de 5000 $ après (et plutôt 8400 $ quand on se place dans la sous-population des salaires non nuls des traités).

Dans le cas où les individus témoins sont ceux du CPS, l'ECM estimé est un effet « différentiel » du NSW, l'écart à l'effet d'au moins un autre programme caché (*hidden*).

* URL de la photo : https://newsroom.iza.org/en/archive/news/iza-fellow-robert-lalonde-1958-2018/

** Ainsi que du *Panel Study of Income Dynamics* (PSID), auquel on ne s'intéressera pas dans ce cours.

On peut voir que les groupes de comparaison expérimentaux sont assez équilibrés (on ne fait pas apparaître les écart-types corrigés, mais les erreurs types, comme dans l'article de Dehejia et Wahba (2002), ce qui ne permet pas de calculer les différences normalisées, mais de visualiser rapidement des intervalles de confiance asymptotiques).

```
estpost tabstat      age […] re75, by(treat) statistics(mean
     semean) columns(statistics) listwise nototal
esttab          ., main(mean) aux(semean) unstack
```

	0	1
age	25.05	25.82
	(0.438)	(0.526)
education	10.09	10.35
	(0.100)	(0.148)
black	0.827	0.843
	(0.0235)	(0.0268)
hispanic	0.108	0.0595
	(0.0193)	(0.0174)
married	0.154	0.189
	(0.0224)	(0.0289)
nodegree	0.835	0.708
	(0.0231)	(0.0335)
re74	2107.0	2095.6
	(352.7)	(359.3)
re75	1266.9	1532.1
	(192.4)	(236.7)
N	445	

mean coefficients; semean in parentheses

Dehejia et Wahba (2002, 161) proposent l'algorithme suivant d'estimation du SP :

1) Estimer un `logit` parcimonieux … pourquoi pas un probit ? Test de Vuong !
2) Trier le score par valeurs croissantes et le stratifier de manière à avoir $\boldsymbol{k}$ intervalles (strates) de scores, à l'intérieur desquels les valeurs sont proches *a priori*
3) Vérifier l'équilibre à l'intérieur de chaque strate
   a. Si équilibre dans les strates (critère à définir : différence normalisée, …), stop
   b. Si déséquilibre dans un strate, y réduire le grossissement (on coupe l'intervalle en deux) et recommencer à a
   c. Si au moins un $\boldsymbol{X_k}$ n'est pas équilibré dans au moins une strate, reprendre à 1 avec une spécification du logit plus flexible, incluant des termes d'ordre 2 pour $\boldsymbol{X_k}$, i.e. $\boldsymbol{X_k^2}$, des interactions $\boldsymbol{X_k} \times \boldsymbol{X_l}$, etc. En pratique, on évite l'ordre supérieur, car cela n'améliore pas l'ajustement (Imbens et Rubin, 2015, 283). La sélection des variables finalement insérées dans le logit repose sur un algorithme *stepwise* standard, *lasso*, l'algorithme d'Imbens et Rubin (2015, 285-288), etc.

      À côté des variables du tableau, on ajoute :
      - les indicatrice de salaire nul en 1975 (**`RE750`**), et en 1974 (**`RE740`**)
      - l'interaction **`age*age`**, **`RE750*RE740`** et autres interactions …

On utilise **pscore.ado** de Becker et Ichino (2002), très utilisée avec **psmatch2.ado** de Leuven et Sianesi (2003) pour faire de l'appariement via le score de propension. Ces commandes ont eu des mises à jours depuis. Il y a aussi **psreg.ado**, **pstrata.ado**, etc.

6.4. Score de propension généralisé

Plutôt que deux traitements qualitatifs, on pourrait en envisager trois ou plus, afin de pouvoir évaluer des non pas l'effet d'une intervention, mais cet effet selon l'intensité de l'intervention.

Pour résumer, il y a trois familles de traitements dans les études, selon le nombres de doses :

Binaire : $D_i \in \{0; 1\}$, avec pour exemple Card et Krueger (1994) dans lequel un Etat américain hausse le salaire minimum dans un secteur ;

Catégoriel : $D_i \in \{0; 1; 2; \dots\}$, Krueger (1999) sur les classes à effectif réduit ou pas, avec une aide à plein temps ou pas, et l'étude de Marino, Lhuillery, Parrotta et Sala (2016) sur les entreprises qui font de la recherche et ont des subventions avec CIR ou pas, ou rien du tout ;

Continu : $D_i \in [D_0; D_1] \subset \mathbb{R}$, Bia et Mattei (2012) qui reprennent un exemple que nous n'avons pas encore étudié, sur des aides aux entreprises, les aides étant une variable continue ! Les différentes valeurs du traitement sont appelées des doses.

Cette année, nous voyons le dernier cas, parce qu'il fait partie des méthodes encore peu connues.

Le cas catégoriel peut toujours être (mal) approché, mais c'est mieux que rien, par des comparaisons deux-à-deux. Par exemple, comme si on avait trois valeurs du traitement, on peut faire $(3 \times 2)/2$ comparaisons, c'est-à-dire C_3^2, le nombre de combinaisons de deux éléments (l'ordre ne compte pas). C'est aussi ½ fois le nombre d'arrangements de deux éléments. Donc, plus généralement, si on a P valeurs : $P(P-1)/2$ évaluations !

Marino, Lhuillery, Parrotta et Sala (2016) découpent la population des entreprises recevant des aides directes en trois groupes, en fonction du centile d'aide directe (trois terciles) : « small », « medium », « large ». Il y a un quatrième groupe, qui est celui des entreprises non aidées. Il y a donc $4 \times 3/2 = 6$ effets calculés.

L'estimation de l'effet du traitement, dans le cas continu (comme le cas binaire), passe par un **protocole d'appariement sur dose**, qui traduit approximativement « dose-response matching », expression que l'on trouve dans Marino, Lhuillery, Parrotta et Sala (2016). Appliqué à l'évaluation des aides aux entreprises, le cas continu permet d'étudier ce que ces auteurs appellent « the proper modulation » du financement public de la R&D privée.

La dose est par exemple un supplément d'aide de 1000 € pour des entreprises qui sont aidées à hauteur de 100000 €. Le contrefactuel est les entreprises qui sont aidées à hauteur de 100000 € et qui ont des caractéristiques proches des premières. On compare aussi les entreprises qui reçoivent 102 000 € à celles qui reçoivent 101000 €, etc.

Plutôt que de mesurer l'effet de doses de 1000 €, quel que soit le niveau d'aide contrefactuel, on peut préférer comparer des doses en pourcentage, 1 % par

exemple, d’où les comparaisons 101000 € vs 100000, 102010 € vs 101000 €, etc.

6.4.1. Le modèle statistique

Les éléments théoriques qui suivent sont tirés Bia et Mattei (2008), Imbens (2000) et Hirano et Imbens (2004).

On définit :

Individus traités $\boldsymbol{i} \in \{\mathbf{1}, \dots, \boldsymbol{N_1}\}$, avec $\boldsymbol{N} \equiv \boldsymbol{N_1}$ (il n'y a pas de $\boldsymbol{N_0}$).
Variables de confusion : $\boldsymbol{X_i}$.
Le traitement reçu : $\boldsymbol{T_i}$ (c'est le $\boldsymbol{D_i}$ du cas binaire)
RO : $\boldsymbol{Y_i}$.
RP : $\boldsymbol{Y_i(t)}$, avec $\boldsymbol{t} \in [\boldsymbol{t_0}; \boldsymbol{t_1}] \equiv \boldsymbol{\mathcal{T}} \subset \mathbb{R}$. On a donc l'équivalent de l'équation de Rubin :

$$\boldsymbol{Y_i} = \boldsymbol{Y(t)} \text{ si } \boldsymbol{T_i} = \boldsymbol{t}, \text{ que l'on pourrait écrire } \boldsymbol{Y_i} = \boldsymbol{Y(t)}\mathbf{1}(\boldsymbol{T_i} = \boldsymbol{t}).$$

Bia et Mattei (2008) appellent $\boldsymbol{t}$ un traitement potentiel.

L'ensemble des RP de $\boldsymbol{i}$, $\{\boldsymbol{Y_i(t)}\}_{\boldsymbol{t} \in [\boldsymbol{t_0}; \boldsymbol{t_1}]}$ est la fonction unitaire de réponse à une dose (*unit-level dose-response function*). Je ne pense pas que les accolades soient justes On s'intéresse à l'espérance $\boldsymbol{E}\big(\boldsymbol{Y_i(t)}\big) \equiv \boldsymbol{\mu(t)}$ prise sur le support du RP, pas du traitement potentiel.

Soit la probabilité conditionnelle, $c_{T|X}(t|x) \equiv r(t,x)$. Le score de propension généralisé (« généralisé » au sens où le nombre de traitements potentiels est infini), noté R chez Bia et Mattei (2008), mais $r(t,x)$ chez Imbens (2000), est

$$R \equiv r(T,X).$$

Notons que $r(t,x) = E(1(T_i = t)|X = x)$.

Hypothèse d'équilibrage. R est supposé avoir la propriété d'équilibrage :

$$X_i \perp 1(T_i = t)|r(t,X_i).$$

On a une variable indicatrice $1(T_i = t)$, comme dans le cas d'un traitement binaire.

On conditionne sur des strates du score, contrairement à ce que l'hypothèse suggère, car il s'agit d'une variable continue.

Hypothèse d'indépendance conditionnelle faible. T est non-confondu sachant X.

$$Y_i(t) \perp T_i|X_i, \text{ pour tous les } t \in \mathcal{T}.$$

Théorème d'indépendance conditionnelle faible (Hirano et Imbens, 2004, 75) :

$$c_{T|R,Y}\big(t\big|r(t,X), Y(t)\big) = c_{T|R}\big(t\big|r(t,X)\big).$$

Dans le cas binaire, le théorème équivalent est celui du score de propension. On pourrait donc l'appeler théorème du score de propension généralisé.

Théorème d'élimination du biais de sélection (Hirano et Imbens, 2004, 76). Définissons $\beta(t,r) \equiv E(Y|T = t, R = r)$, et $\mu(t) = E_Y(Y(t))$. Alors on a :

(i) $\boldsymbol{E(Y(t)|r(t,X)=r)=\beta(t,r)}$.
(ii) $\boldsymbol{E_R(\beta(t,r))=\mu(t)}$.

J'obtiens également une variante de la deuxième partie du théorème, qui est $E_{R|T}(\beta(t,r))=E(Y(t)|T=t)$.

Estimation. Elle se déroule en plusieurs étapes, comme pour le cas binaire.

1) On estime $\boldsymbol{r(t,x)}$ directement pour $\boldsymbol{T}$ ou une transformation $\boldsymbol{g(T)}$, afin de garantir (empiriquement) que l'on ait $\boldsymbol{g(T)\sim N(h(\gamma,X_i),\sigma^2)}$, où $\boldsymbol{h}$ doit satisfaire la propriété d'équilibrage. C'est un modèle semi-paramétrique avec des termes d'ordre 2 ou plus en $\boldsymbol{X}$. Les estimations de $\boldsymbol{\gamma}$ et $\boldsymbol{\sigma^2}$ sont obtenues par MV.

2) L'estimation du GPS est $\boldsymbol{\widehat{R}_i=\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}e^{-\frac{1}{2\hat{\sigma}^2}\left(g(T_i)-h(\hat{\gamma},X_i)\right)^2}}$.

6.5. Exercices

6.5.1) Dans l'évaluation de Card et Krueger (1994), quels facteurs de biais l'appariement sur les variables « chaîne de magasin » et « emploi initial », permet-il de contrôler?

6.5.2) (Exercice élaboré à partir du tableau 18.4 de Imbens et Rubin (2015, 414)) Calculer l'estimateur d'appariement sans remise de l'ECMT pour les cinq individus tests dans l'ordre suivant lequel ils apparaissent dans la feuille Excel.

Traité (i)	Contrôle (j)	Y_i	Y_j	$\hat{\tau}_i^{\text{appa}}$	
1					
2					
3					
4					
5					
				$\Sigma =$	$\hat{\tau}^{\text{appa}} =$

6.5.3) **Logit.** Soit (D, X) un vecteur de variables aléatoires, avec D une variable de Bernoulli et X dans $\mathbb{R}^K$. Par définition, $\Pr(D = 1|X = x) = E(D|X = x)$. Notons $\epsilon \equiv D - \mathrm{E}(D|X = x)$. L'identité $D = \Pr(D = 1|X = x) + D - \Pr(D = 1|X = x)$ entraine $D = \mathrm{E}(D|X = x) + \epsilon$. Montrer que :

(i) $\mathrm{E}(\epsilon|X = x) = 0$.

(ii) Notons $\Pr(D = 1|X = x) \equiv F(x'\beta)$. alors $V(\epsilon|X = x) = F(x'\beta)(1 - F(x'\beta))$

(iii) Montrez enfin que $\partial \Pr(D = 1|X = x)/\partial x_k = F(1 - F)\beta_k$.
Remarque : on a simplifié les notations en écrivant F pour $F(x'\beta)$.

6.5.4) Montrer que

$$\int [E(Y|D = 1, X = x) - E(Y|D = 0, X = x)]C(x|0)dx$$

est un estimateur sans biais de l'ECMnT, $E(Y(1) - Y(0)|D = 0)$.
Astuce : partez de l'ECMnT. Vous n'aurez besoin de la supposition d'indépendance conditionnelle pour $Y(1)$ seulement.

6.5.5) **Estimation par pondération à la Horvitz-Thompson.** Nous laissons cette méthode en exercice. Il s'agit d'estimer l'ECM en divisant les résultats Y_i des individus tests par $\hat{e}(x_i)$, et les y_j des individus témoins par $1 - \hat{e}(x_j)$. L'idée est de pondérer par l'inverse de la probabilité de sélection (Imbens et Rubin, 2015, 273-274). Cette pondération rend l'estimateur plus sensible que les autres à la spécification du score de propension. Il y a un risque que le rapport explose pour les unités telles que $1 - e$ est très proche de zéro (la supposition de recouvrement n'est pas valide).

(i) Montrer que $E_A\left(\frac{YD}{e(X)}\right) = E(Y(1))$ et $E_A\left(\frac{Y(1-D)}{1-e(X)}\right) = E(Y(0))$.
Remarque : A est le support d'intégration des variables $(Y(1), Y(0), D, X)$, comme dans Imbens et Rubin (2015, 73) où l'espérance est prise sur la distribution de randomisation de D et la distribution d'échantillonnage de $(Y(1), Y(0), D, X)$ dans la super-population infinie.

(ii) Estimer l'ECM sur le sous-échantillon de Dehejia et Wahba (2002) des données de Lalonde (1986) à l'aide de la commande

```
teffects ipw      (re78) ///
                  (treat age education …, logit),
                                        ///
                  atet ///
                  pstolerance(1e-5) ///
                  osample(OSAMPLE) ///
```

6.5.6) **Théorème du score de propension.** Il s'agit en fait d'un ensemble de théorèmes, montrant que le score de propension est un score d'équilibrage, avant de montrer que l'hypothèse d'indépendance conditionnelle est valide avec le score.
[…]
La dernière étape consiste à montrer, comme le font Angrist et Pischke (2009, 81), l'indépendance entre le RP et le traitement, conditionnellement au score, c'est-à-dire $C\big(1\big|Y(1), C(1|X)\big) = C(1|X)$. La démonstration avec $Y(0)$ au lieu de $Y(1)$ est similaire.

(i) Montrer $\boldsymbol{C\big(D\big|Y(1), C(1|X)\big) = C(D|X)}$ pour $\boldsymbol{D \in \{0; 1\}}$, et sous CIA.

6.4.7) **Distance de Mahalanobis**.

6.5.1)
(à l'oral)

6.5.2)

Traité (i)	Contrôle (j)	Y_i	Y_j	$\hat{\tau}_i^{\text{appa}}$	
1	11	40	19,5	20,5	
2	7	12,5	17	$-4,5$	
3	15	20	22,5	$-2,5$	
4	8	3,5	10,5	-7	
5	20	5,5	8	$-2,5$	
				4	$\hat{\tau}^{\text{appa}} = 0,8$

6.5.3)
(i) $E(\epsilon|X=x) = (1-F)F + (-F)(1-F) = 0.$
(ii) $V(\epsilon|X=x) = (1-F-0)^2F + (-F-0)^2(1-F).$
$= (1-F)[(1-F)F + (-F)^2]$ ⇐ Factorisation de $1-F$.
$= (1-F)F[1-F+F]$ ⇐ Factorisation de F.
$= (1-F)F.$
(iii) $\frac{\partial \Pr(D=1|X=x)}{\partial x_k} = \frac{\partial F(x'\beta)}{\partial x_k} = \frac{\partial F(x'\beta)}{\partial (x'\beta)} \times \frac{\partial (x'\beta)}{\partial x_k} = F'\beta_k$. Or, $F' = \left(\frac{e^{x'\beta}}{1+e^{x'\beta}}\right)'$, où l'argument par rapport auquel on dérive est $x'\beta$. On obtient :

$$\left(\frac{e^{x'\beta}}{1+e^{x'\beta}}\right)' = \frac{e^{x'\beta}\left(1+e^{x'\beta}\right)-e^{x'\beta}e^{x'\beta}}{\left(1+e^{x'\beta}\right)^2} = \frac{e^{x'\beta}}{\left(1+e^{x'\beta}\right)^2} = \frac{e^{x'\beta}}{1+e^{x'\beta}}\frac{1}{1+e^{x'\beta}} = F(1-F),$$

d'où $\frac{\partial \Pr(D=1|X=x)}{\partial x_k} = F(1-F)\beta_k$. Pour obtenir l'élasticité, il suffit de multiplier par x_k/F. Cette dernière vaut donc $(1-F)x_k\beta_k$.

6.4.4)
$E(Y(1)-Y(0)|D=0)$
$= E(Y(1)|D=0) - E(Y(0)|D=0)$
$= E(E(Y(1)|0,X)|0) - E(E(Y(0)|0,X)|0)$
$= E(E(Y(1)|1,X)|0) - E(E(Y(0)|0,X)|0)$
$= E(E(Y|1,X)|0) - E(E(Y|0,X)|0)$
$= E[E(Y|1,X) - E(Y|0,X)|0].$

■

6.4.5) (i) La démonstration reprend Angrist et Pischke (2009, 82), Imbens et Rubin (2015, 273-274). Avant de démontrer (i), notons que D joue le rôle de variable de sélection : en multipliant Y par D, seuls les Y pour $D=1$ vont rester, donc $Y(1)$. En effet, en postmultipliant l'équation de Rubin par D, on obtient $YD = Y(1)D^2 + Y(0)(1-D)D$, qui vaut $Y(1)$ si $D=1$, et 0 sinon ; Lee (2016, 62) propose des extensions de ce point.

$$E_A\left(\frac{YD}{e(X)}\right) = \int\int\int\int ywj(y(0),y(1),w,x)e^{-1}(x)dy(0)dy(1)dwdx$$
$$= \int\int\int\int ywc(y(0),y(1)|w,x)e(x)m(x)e^{-1}(x)dy(0)dy(1)dwdx \quad \text{LEI}$$
$$= \int[\int\int\int ywc(y(0),y(1)|w,x)dy(0)dy(1)dw]m(x)dx$$
$$= \int[\int\int y(1)c(y(0),y(1)|1,x)dy(0)dy(1)]m(x)dx \quad yw = y(1)$$
$$= \int[\int\int y(1)c(y(0),y(1)|x)dy(0)dy(1)]m(x)dx \quad \text{CIA}$$
$$= \int[\int y(1)c(y(1)|x)dy(1)]m(x)dx$$

$= \int y(1) m\big(y(1)\big) dy(1)$ ■

(ii)

6.4.6)

(i) Notons $e(x) \equiv C(D = 1|X = x)$. Il faut montrer que $C\big(1\big|Y(1), e(X)\big) = e(X)$. Je galère encore avec le choix X ou x dans la condition

$C(1|Y(1), e(X))$
$= E(D|Y(1), e(X))$
$= E(E(D|Y(1), e(X), X)|Y(1), e(X))$
$= E(E(D|Y(1), X)|Y(1), e(X))$ ⇐ $e(X)$ n'est pas plus fin que X, il saute
$= E(E(D|X)|Y(1), e(X))$ ⇐ CIA
$= E(e(X)|Y(1), e(X))$
$= e(X)$. ■

7. Ajustement par régression

En économétrie standard, l'ajustement par régression consiste à insérer $\boldsymbol{X}$ directement dans une fonction de régression. Cette approche est utilisée pour à la fois contrôler des variables $(1, D, \dots, X_{K-1})$ et mesurer l'effet de certaines d'entre-elles (D et X_3 par exemple). Donc $\boldsymbol{X} \coloneqq (1, D, \dots, X_{K-1})$ inclus ces deux types de variables. La forme générique d'un modèle de régression linéaire multiple, par exemple, est :

$$E(Y|\boldsymbol{X}) = \boldsymbol{X\beta} = \beta_2 D + \sum_{k\neq 2} \beta_k X_k.$$

Nous allons voir le cas bivarié sur une page (**section 7.1**). Ce n'est pas la peine ici de rappeler le cours d'économétrie que vous avez-vu en Licence. Le problème de cette approche, est qu'elle ne permet d'estimer qu'un effet « moyen » de D, une combinaison de l'ECMT et de l'ECMnT. On ne sait pas bien ce qu'on estime.

Dans les MSE, la régression peut être utilisée de trois autres manières :

1) Calculer spécifiquement trois effets au choix : les deux ci-dessus, ou l'ECM. Or, pour calculer ces effets, nous devons être capable de prédire un ou deux résultats potentiels contrefactuels : $\boldsymbol{E(Y(0)|D=1)}$, $\boldsymbol{E(Y(1)|D=0)}$, ou les deux. On prédira ces quantités grâce à la méthode des moindres carrés par exemple, sous l'hypothèse CIA. Ces prédictions font partie du protocole. En notant $\boldsymbol{X_0}$ et $\boldsymbol{X_1}$ les vecteurs des variables des groupes de traitement, et $\boldsymbol{D}$ le traitement, on se place dans une sous-population (c'est donc un *matching*), celle $\boldsymbol{X_0}$, ou celle de $\boldsymbol{X_1}$, ou les deux, et effectuons une ou plusieurs régressions intermédiaires. Il suffit ensuite de calculer les paramètres du modèle linéaire simple suivant

   $$E(Y|D) = \beta_1 + D\beta_2,$$

   sans n'avoir plus besoin d'utiliser $\boldsymbol{X}$. Le programme suivant fait tout ça, il applique cette méthode sur plusieurs bases de données :

   [« **regression.do** »]

2) La régression polynomiale (**section 7.2**)
3) La régression en discontinuité (**section 7.3**) qui est appropriée dans le cas où il n'y a pas recouvrement pour au moins une variable du MAT.

Pour bien comprendre la section 7.2, il faut voir ce que l'on peut appeler le Potential Outcome Regression Model. C'est un cadre d'analyse qui repose sur une intégration des RP de l'équation de Neyman-Rubin au modèle de régression linéaire dans lequel intervient le fameux terme d'erreur. Les questions à se poser sont les suivantes :

- Faut-il une nouvelle CIA, plus générale, combinée avec le terme d'erreur ?
- Peut-on combiner la régression avec au moins une des méthodes vues avant (méthode d'appariement ? méthode par le score de propension ?)
- Faut-il poser des hypothèses de Gauss-Markov ?

Il est clair que ce mariage n'est pas trivial, à en croire l'assertion suivante d'Imbens et Angrist (1994, 467) :

> « The framework we use […] defines causal effects in terms of potential outcomes or counterfactuals rather than in terms of the parameters of a regression model ».

7.1. Modèle de régression résultat observé-traitement

7.1.1. Quelques rappels sur la régression

Nous allons travailler avec le modèle

$$Y = E(Y|D) + \epsilon, \tag{7.1}$$

sans autre variable explicative, et
sans supposer la linéarité.

Vous avez utilisé cette approche en économétrie tout le temps. Angrist et Pischke (2009) passent quelques pages à justifier l'intérêt de l'approche, dans l'optique de l'inférence causale. En résumé :

- Le membre de droite de (7.1) résulte de la propriété de décomposition de Y en une fonction d'espérance conditionnelle (FEC) et un terme d'erreur ϵ. Par définition, $E(\epsilon|D) = 0$ (p. 32).
- La décomposition (7.1) modélise l'influence de prédicteurs (facteurs de premier intérêt), ici D, sur la variable de résultat Y. Par exemple, la relation salaire-études supérieures (p. 29), avec D une variable dichotomique indiquant si l'individu a fait des études supérieures dans sa vie. C'est une variable aléatoire (p. 30).

[rappeler la forme intégrale de $E(Y|D = d)$]

- $E(Y|D)$ est la meilleure fonction de prédiction de l'effet de D sur Y, au sens où elle minimise l'espérance du carré de l'erreur (*minimum mean squared error*). Supposons une fonction $m(D)$ de la variable aléatoire D, telle que $Y = m(D) + \epsilon$. Pour n'importe quelle valeur de D,

$$\min_{m(d)} E[\big(Y - m(d)\big)^2|D = d] \Rightarrow m(d) = E(Y|D = d).$$

C'est ce qu'Angrist et Pischke appellent la propriété de prédiction de la FEC.

- L'analyse de la variance de Y montre qu'elle dépend de la variance du 'modèle' (de la FEC), et de celle de l'erreur (démonstration plus détaillée dans l'**annexe D**).

Partons de l'identité $V(Y|D) = \int \big(y - E(Y|D)\big)^2 C(y|D) = E(Y^2|D) - E^2(Y|D)$.
Quant à la variance de Y, elle vaut $V(Y) = \int y^2 m(y) - \Big(\int y m(y)\Big)^2$. Introduit D :

$$\begin{aligned} V(Y) &= \int y^2 m(y) - \Big(\int y m(y)\Big)^2 \\ &= \int y^2 \int j(y,d) - \Big(\int y \int j(y,d)\Big)^2 \\ &= \int\int y^2 c(y|d) m(d) - \Big(\int\int y c(y|d) m(d)\Big)^2 \\ &= \boxed{E\big(E(Y^2|D)\big)} - E^2(E(Y|D)) \end{aligned}$$

Or, si on intègre $V(Y|D)$ sur D, on obtient $E\big(V(Y|D)\big) = \boxed{E\big(E(Y^2|D)\big)} - E(E^2(Y|D))$.
Par conséquent,

$$\begin{aligned} V(Y) &= E\big(V(Y|D)\big) + E\big(E^2(Y|D)\big) - E^2\big(E(Y|D)\big). \\ &= E\big(V(Y|D)\big) + V\big(E(Y|D)\big). \end{aligned}$$

■

- Si on randomise D, son coefficient dans la régression a une interprétation causale. Sinon, ça dépend des MSE que l'on utilise (pp. 28-29).

7.1.2. Randomisation et exogénéité de D

On suppose donc le modèle linéaire :

$$Y = \beta_1 + D\beta_2 + \epsilon. \quad (7.1)$$

Le fait d'écrire cette équation pose immédiatement la question de la relation entre le traitement et l'erreur ϵ.

Si le traitement est randomisé, cette spécification de la relation entre le RO Y et le traitement D ne change pas la nature du traitement qui se trouve à l'intérieur. Statistiquement, $Cov_{D,\epsilon}(D, \epsilon) = 0$.

Si, à la place de la randomisation de D on suppose $E_{\epsilon|D}(\epsilon|D) = 0$, la fameuse hypothèse d'exogénéité, on obtient le résultat, $Cov_{D,\epsilon}(D, \epsilon) = 0$. En résumé, randomisation et $E_{\epsilon|D}(\epsilon|D) = 0$ impliquent chacune l'exogénéité de D, $Cov_{D,\epsilon}(D, \epsilon) = 0$. **Exercice 7.4.1**

7.1.3. Relation entre ϵ et les RP

La question plus difficile est celle de savoir comment introduire les résultats potentiels dans le modèle (7.1), et est-ce que le fait de les introduire nous oblige à faire des suppositions d'exogénéité supplémentaires. Par exemple, faut-il que D soit indépendant de $Y(0)$? S'il y a randomisation, D est **indépendant des RP** (supposition *strong ignorability*). Mais alors, quelle relation ϵ entretient avec les RP ?

Mettons l'équation de Neyman-Rubin dans (7.1) :

$$DY(1) + (1 - D)Y(0) = \beta_1 + D\beta_2 + \epsilon,$$

arrangeons un peu :

$$Y(0) + D(Y(1) - Y(0)) = \beta_1 + D\beta_2 + \epsilon,$$

et procédons par identification :

$$\begin{aligned} Y(0) &= \beta_1, \\ Y(1) - Y(0) &= \beta_2. \\ ? &= \epsilon \end{aligned}$$

Malheureusement, ϵ n'est identifié à rien ! Comment rendre le modèle de régression et le modèle avec RP compatibles ? Un approche consiste à partir du système suivant, où nous avons le modèle 7.1 pour chaque RP (un modèle de régression dans chaque groupe de traitement), combiné à l'équation de Neyman-Rubin :

$$Y(0) = E(Y(0)) + \epsilon_0 \quad (7.2)$$
$$Y(1) = E(Y(1)) + \epsilon_1 \quad (7.3)$$
$$Y = DY(1) + (1 - D)Y(0) \quad (7.4)$$

C'est l'approche d'Heckman (1990),[35] reprise également dans Wooldridge (2003), p. 611. Alors,

$$\begin{aligned} Y &= DY(1) + (1 - D)Y(0) \\ &= D(E(Y(1)) + \epsilon_1) + (1 - D)(E(Y(0)) + \epsilon_0) \end{aligned}$$

$$Y = E(Y(0)) + D(E(Y(1)) - E(Y(0))) + D\epsilon_1 + (1 - D)\epsilon_0 \quad (7.5)$$

[35] On peut montrer que même si on part de $Y(0) = \beta_{10} + D\beta_{20} + \epsilon_0$, β_{20} est éliminé dans la simplification.

Il s'agit en fait d'un *switching regression model* (Wooldridge, 2003, 611), où selon le statut d'une unité d'observation, on a l'équation (7.2) ou bien (7.3).

Egalisons (7.1) et (7.5) :

$$E(Y(0)) + D(E(Y(1)) - E(Y(0))) + D\epsilon_1 + (1-D)\epsilon_0 = \beta_1 + D\beta_2 + \epsilon,$$

et identification membre à membre :

$$\begin{aligned} E(Y(0)) &= \beta_1, \\ E(Y(1)) - E(Y(0)) &= \beta_2, \\ D\epsilon_1 + (1-D)\epsilon_0 &= \epsilon. \end{aligned}$$

■

ϵ est désormais égale à une erreur composite. On peut remarquer que le coefficient devant D est l'effet causal moyen dans la population, $E(Y(1)) - E(Y(0))$.

À quelle condition D est exogène dans cette régression ?
Plus formellement, à quelle(s) condition(s) $Cov(\epsilon, D) = 0$?

$$\begin{aligned} Cov(\epsilon, D) &= Cov(D\epsilon_1 + (1-D)\epsilon_0, D) \\ &= Cov(D\epsilon_1, D) + Cov\big((1-D)\epsilon_0, D\big) \\ &= E(D^2\epsilon_1) - E(D\epsilon_1)E(D) + E((1-D)D\epsilon_0) - E\big((1-D)\epsilon_0\big)E(D) \end{aligned}$$

Or,

$$\begin{aligned} E(D^2\epsilon_1) - E(D\epsilon_1)E(D) &= E(D^2E(\epsilon_1|D)) - E(DE(\epsilon_1|D))E(D) \\ &= E\{[D(D - E(D))]E(\epsilon_1|D)\}, \end{aligned}$$

et,

$$\begin{aligned} E((1-D)D\epsilon_0) - E\big((1-D)\epsilon_0\big)E(D) &= E((1-D)DE(\epsilon_0|D)) - E((1-D)E(\epsilon_0|D))E(D) \\ &= E\{\big[(1-D)\big(D - E(D)\big)\big]E(\epsilon_0|D)\}. \end{aligned}$$

Si l'on somme ces deux expression on a $E\big\{\big(D - E(D)\big)(DE(\epsilon_1|D) + (1-D)E(\epsilon_0|D)\big\}$, d'où la condition triviale : $E(\epsilon_1|D) = E(\epsilon_0|D) = 0$.[36] Alors $Cov(\epsilon, D) = 0$. Or, $E(\epsilon_d|D) = E\big(Y(d) - E\big(Y(d)\big)\big|D\big) = E(Y(d)|D) - E(Y(d))$, $d \in \{0; 1\}$. Par conséquent, $E(\epsilon_d|D) = 0 \Leftrightarrow E(Y(d)|D) - E\big(Y(d)\big) = 0 \Leftrightarrow E(Y(d)|D) = E\big(Y(d)\big)$, $d \in \{0; 1\}$. La supposition de *strong ignorability* suffit, sous la forme d'indépendance en moyenne, pas en distribution.

Si on a la *strong ignorability*, alors β_2 a une interprétation causale.

Chez Imbens (2004), c'est un petit peu plus simple car on ne suppose un modèle de régression que pour le groupe de contrôle, plus une hypothèse pour l'effet causal, avec toujours l'équation de Neyman-Rubin. Il n'y a donc qu'une équation qui change. On peut

[36] Plus tard : Wooldridge (2003, 606-607) obtient cette condition en soustrayant (7.2) à (7.3). Puis il déduit que ECMT=ECM si et seulement si on a $E(\epsilon_1 - \epsilon_0|D = 1) = 0$; voir aussi Heckman (1990, 314). On voit que si on développe $E\{[D(D - E(D))]E(\epsilon_1|D)$ et $E\{\big[(1-D)\big(D - E(D)\big)\big]E(\epsilon_0|D)\}$, on obtient $V(D)[E(\epsilon_1|1) - E(\epsilon_0|0)]$. La condition de nullité est donc en réalité plus faible : $E(\epsilon_1|1) = E(\epsilon_0|0)$.

déjà anticiper que la condition unique sera $E(Y(0)|D) = E\big(Y(0)\big)$. On a le modèle suivant :

> Wooldridge (2003), p. 604, introduit le vecteur aléatoire $(Y_i(0), Y_i(1), D_i)$, sous-entendu, un échantillon aléatoire, mais il revient à la population. C'est généralement le choix de Angrist et Pischke (2009).

Revenons au modèle :

$$Y(0) = E(Y(0)) + \epsilon_0 \quad (7.6)$$
$$Y(1) = Y(0) + \tau \quad (7.7)$$
$$Y = DY(1) + (1 - D)Y(0) \quad (7.8)$$

Alors,

$$Y = DY(1) + (1 - D)Y(0) = D(E(Y(0)) + \epsilon_0 + \tau) + (1 - D)(E(Y(0)) + \epsilon_0)$$
$$Y = E(Y(0)) + D\tau + \epsilon_0$$

Ici, D est exogène, ssi D est indépendant de ϵ_0. τ pourra alors être estimé de manière consistante. Etant donné (7.6), cela revient à supposer l'indépendance en moyenne entre $Y(0)$ et D, i.e. la condition $Cov(Y(0), D) = 0$; voir Wooldridge (2003, p. 606).

> Si on développe cette condition, étant donné que D prend deux valeurs, 0 et 1, on aboutit à $V(D)[E(Y(0)|D = 1) - E(Y(0)|D = 0)] = 0$.[37] Autrement dit, l'absence de BS, ou aussi $E(Y(0)|D) = E(Y(0))$. On a donc pas besoin de la *strong ignorability* complète (voir la note de bas de page précédente) qui poserait en plus comme condition que $E(Y(1)|D) = E(Y(1))$. Tant mieux quelque part, puisque la formule du BS n'implique pas $Y(1)$ justement. Ici aussi, une randomisation, ou bien une méthode d'appariement bien utilisée devrait aboutir à estimer τ de manière consistante.

On peut conclure cette sous-section en disant que les condition qui font que D est exogène, sont aussi fortes que dans le modèle de régression linéaire classique. Au lieu de l'exogénéité, $Cov(D, \epsilon)$, on a, dans le cas le plus faible, $E(Y(0)|D) = E(Y(0))$. On ne voit pas comment cette condition pourrait être satisfaite en dehors d'un MATAC. Ou bien, de poser la CIA que pour $Y(0)$: $E(Y(0)|D, X) = E(Y(0)|X)$ Wooldridge (2003, 606-707), où X est observable (cf. **infra**).

[37] À condition qu'il n'y ait pas d'autres variables non-observées en jeu. On a une distribution de probabilité qui porte sur $(Y(0), Y(1), D, Y, \epsilon_0)$.

7.1.4. L'estimateur des MC identifie l'ECM (cas bivarié)

On continue de raisonner dans la population. On sait que sous certaines suppositions, β_2 est égal à l'ECM, et que β_1 est égal au RP des non-traités. Ce que l'on veut savoir, maintenant c'est si les MC permettent de calculer un effet causal.

$$Y = \beta_1 + D\beta_2 + \epsilon \tag{7.9}$$

Rappelons qu'il s'agit d'une forme réduite de l'équation (7.5), que nous recopions ici :

$$Y = E(Y(0)) + D(E(Y(1)) - E(Y(0))) + D\epsilon_1 + (1 - D)\epsilon_0 \tag{7.5}$$

Notons $p \coloneqq E(D)$. Les estimateurs des MCO de β_2 et β_1 sont :

$$\beta_2^{MC} \coloneqq \frac{Cov(Y,D)}{V(D)}, \beta_1^{MC} \coloneqq E(Y) - E(D)\beta_2^{MC}$$

Or,

$$Cov(Y, D) = E(YD) - E(Y)E(D)$$
$$E(YD) = E\big(DE(Y|D)\big) = 1E(Y|1)p + 0E(Y|0)(1-p) = E(Y|1)p,$$
$$E(Y) = E\big(E(Y|D)\big) = E(Y|1)p + E(Y|0)(1-p),$$
$$E(Y)E(D) = E(Y|1)p^2 + E(Y|0)p(1-p),$$

Donc, $Cov(Y, D) = E(Y|1)p - E(Y|1)p^2 + E(Y|0)p(1-p)$
$= [E(Y|1) - E(Y|0)]p(1-p)$

Et $V(D) = p(1-p)$

Par conséquent,

$$\frac{Cov(Y,D)}{V(D)} = E(Y|1) - E(Y|0).$$

Si on met l'équation de Neyman-Rubin à l'intérieur de ces espérances, on a :

$$E(Y|1) - E(Y|0) = ECMT + BS,$$

où $BS \coloneqq E(Y(0)|D = 1) - E(Y(0)|D = 0)$; cf. **chapitre 4**.

Il n'y a pas 36 solution ici, le seul moyen d'éliminer le BS c'est la randomisation ou une hypothèse sur BS. Et dans ce cas, l'ECM est égal à l'ECMT. Comme dans la section précédente, on estime l'ECMT plutôt que l'ECM. L'un et l'autre peuvent se défendre, mais la condition d'exogénéité dans le premier cas est plus faible ($BS = 0$) puisqu'elle ne porte que sur $Y(0)$.

7.2. Estimateur paramétrique polynomial (on introduit X)

Comme on l'a vu précédemment, sous (CIA + recouvrement) + équilibre (la version en termes d'espérances), il suffit de régresser Y sur D et X pour obtenir une estimateur MCO sans biais de β_2. Nous avons toujours :

$$Y = E(Y(0)) + D(E(Y(1)) - E(Y(0))) + D\epsilon_1 + (1 - D)\epsilon_0 \qquad (7.5)$$

Une autre approche, est d'utiliser un modèle paramétrique mais plus flexible, de type polynomial, avec des termes d'interaction. Cette approche permet d'obtenir des estimateurs asymptotiques des variances des coefficients, notamment de l'ECM ou de l'ECMT ; Wooldridge (2003), p. 609, 611-614.

On a deux résultats intéressants. On reprend 7.2 et 7.3, qui sous CIA impliquent :

$E(Y(1)|D,X) = E(Y(1)) + E(\epsilon_1|X)$, et
$E(Y(0)|D,X) = E(Y(0)) + E(\epsilon_0|X)$

On peut définir $E(\epsilon_1 - \epsilon_0|X)$ comme l'effet propre (non-observable) à l'individu, que l'on fait dépendre d'observables. C'est ici la source du biais de sélection.

Wooldridge (2003, 611) pose la condition $E(\epsilon_1|X) = E(\epsilon_0|X)$, d'effet propre nul, ce qui implique que l'ECM est le même pour tous les individus. On prend la différence entre les deux équations précédentes :

$E(Y(1)|D,X) - E(Y(0)|D,X) =$ECM.

Or, d'après la loi des espérances itérée :

$E(E(Y(1)|D,X) - E(Y(0)|D,X)|D) = E(Y(1)|D) - E(Y(0)|D) =$ECM.

Ce qui est aussi vrai pour $D = 1$, donc ECMT=ECM.

L'autre résultat est que, sous la supposition supplémentaire que $E(\epsilon_0|X)$ est une fonction des X, $h_0(\boldsymbol{X})$, qu'on appelle fonction de contrôle (du biais de sélection), alors $E(Y|D,X)$ est de la forme

$$\beta_{01} + \eta_0 + D\beta_{02} + h_0(\boldsymbol{X})\beta_0.$$

Régresser Y sur D, $h_0(\boldsymbol{X})$ permet d'obtenir un estimateur consistant de l'ECM.

C'est-à-dire, on est toujours dans le cadre du SRM, avec les suppositions que β_2 est l'ECM, et β_1 le RP moyen du groupe de contrôle. Comme son indice l'indique, la fonction $g_0(\boldsymbol{X})$ est pour le groupe de contrôle.

La régression en discontinuité utilise ce type de fonction.

7.3. Régression en discontinuité

On a ce protocole (*regression discontinuity design*) quand la sélection des individus, D, dépend du franchissement d'un seuil z_0 (*cutoff value*) par une variable Z (*forcing variable*). Y est alors discontinue au point z_0. Les individus autour de z_0 sont suffisamment similaires pour supposer que c'est comme s'ils avaient été randomisés.

Pour Angrist et Pischke (2009), cela se produit par exemple quand une PP repose sur une législation fixant un seuil d'éligibilité. Par exemple,

Pour bénéficier d'un taux d'impôt sur les sociétés (IS) réduit (15 %) quand on est une TPE, le bénéfice annuel ne doit pas dépasser 38120 €

Dans l'expérimentation de terrain TICELEC, la distance entre le compteur électrique et la passerelle domestique devait être inférieure à 20 m
(l'organisation des pièces, l'épaisseur des murs, … comptaient aussi)

Les protocoles courants sont le *sharp design* et *fuzzy design*. Nous verrons le premier.

7.3.1. Modèle avec protocole *sharp*

La formalisation de la discontinuité passe par une *step function* (Wooldridge, 2003, 614). Parmi les variables explicatives, $\boldsymbol{X}$, il existe un Z tel que

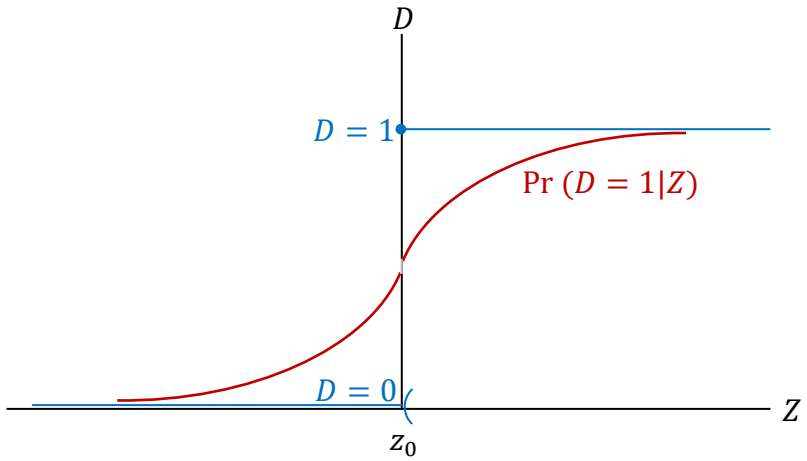
$$D = \begin{cases} 1 \text{ si } Z \geq z_0 \\ 0 \text{ sinon} \end{cases} \Leftrightarrow \underbrace{\mathbb{I}_{[z_0;+\infty[}(Z)}_{\substack{\textit{step} \\ \textit{function}}} \tag{7.13}$$

Le critère de sélection dans les groupes de traitement est exogène et observable, ce qui revient à dire que le MAT est déterministe.

La supposition de recouvrement $\forall\, i, d \in \{0;1\}, x, 0 < \Pr(D_i = d|X_i = x) < 1$ ne peut être faite dans la mesure où $\exists\, i, j : \Pr(D_i = 1|Z_i = z_0) = 1$ et $\Pr(D_j = 0|Z_j = z_0) = 0$.

En théorie, au franchissement de la valeur seuil z_0, le résultat Y devrait être décalé pour les individus exposés au traitement, comme la probabilité $\Pr(D = 1|Z)$. C'est ce décalage qui mesure l'effet causal de D (pas l'effet de Z, qui est continue).

Figure 7.2 : saut de D (sharp) et $\Pr(D = 1|Z)$ (fuzzy) au point de discontinuité $Z = z_0$.



Contrairement au protocole *sharp*, selon lequel la position d'un individu à gauche ou à droite de z_0 détermine son traitement (sélection sur observables), le protocole *fuzzy* fait intervenir une ou plusieurs variables instrumentales.

L'avantage du protocole fuzzy, qui fait penser à l'estimateur LATE que nous verrons dans le **chapitre 9**, est qu'il est plus réaliste car il tient compte du fait que des individus éligibles (une petite fraction) ne participent pas, pendant que des non-éligibles participent. Alors qu'avec le protocole *sharp*, d'où le nom, la relation entre D et Z est déterministe (« *expected take up* » = « *actual take-up* »).

On reprend le modèle de régression avec RP de Imbens (2004), dans lequel on insère la discontinuité (Wooldridge, 2010, 955). Nous avons :

$$\begin{aligned} Y(0) &= E\big(Y(0)\big) + \epsilon_0, \\ Y(1) &= Y(0) + \beta_2, \text{ et} \\ Y &= DY(1) + (1-D)Y(0) = \mathbb{I}_{[z_0;+\infty[}(Z)Y(1) + \mathbb{I}_{]-\infty;\, z_0[}(Z)Y(0). \end{aligned}$$

Vu que $\mathbb{I}_{]-\infty;\, z_0[}(Z) = 1 - \mathbb{I}_{[z_0;+\infty[}(Z)$, alors après avoir noté $\beta_1 \equiv E\big(Y(0)\big)$, on obtient :

$$Y = \beta_1 + \mathbb{I}_{[z_0;+\infty[}(Z)\beta_2 + \epsilon.$$

Supposons que la quantité qui nous intéresse autour du point z_0 soit l'ECM en ce point :

$$E(\beta_2|Z = z_0) = E(Y(1)|Z = z_0) - E(Y(0)|Z = z_0).$$

Le dernier terme n'existant pas par définition, on va le remplacer par $\lim_{Z\uparrow z_0} E(Y|Z)$.

Ce n'est pas la peine d'utiliser une méthode s'appuyant par le score de propension à cause du non-respect de la supposition de recouvrement au point z_0. À gauche de ce point, et en supposant $\boldsymbol{X} \coloneqq (X, Z)$, $\Pr(D = 1|\boldsymbol{X}) = 0 \;\forall\; Z < z_0$. Vu qu'on ne peut diviser par 0, la méthode de pondération à la Horvitz-Thompson est exclue.

CIA est nécessairement vérifiée puisque D est une fonction déterministe de Z. Si $Z \geq z_0$, alors $E(Y(d)|D = d', Z = z)$ ne dépend pas de D, mais de Z puisque D est relié déterministiquement à Z.

Il y a deux hypothèses à vérifier avant d'utiliser cette méthode. Nous en donnerons un aperçu lors de l'application à l'évaluation de Khandker *et alii* (2010) ; voir aussi Nichols (2007, 528), le chapitre 6 d'Angrist et Pischke (2009, 253-) :

ϵ et β_2 ont la même limite au point z_0 donc sont continues en ce point

$$\begin{aligned} \lim_{Z\uparrow z_0} E(\epsilon|Z) &= \lim_{Z\downarrow z_0} E(\epsilon|Z), \\ \lim_{Z\uparrow z_0} E(\beta_2|Z) &= \lim_{Z\downarrow z_0} E(\beta_2|Z). \end{aligned}$$

Considérons l'estimateur des moindres carrés suivant :

$$\begin{aligned} \beta_2^{MC} &\equiv \lim_{Z\downarrow z_0} E(Y|Z) - \lim_{Z\uparrow z_0} E(Y|Z) \\ &= \lim_{Z\downarrow z_0} E(\beta_1 + \mathbb{I}_{[z_0;+\infty[}(Z)\beta_2 + \epsilon|Z) - \lim_{Z\uparrow z_0} E(\beta_1 + \mathbb{I}_{[z_0;+\infty[}(Z)\beta_2 + \epsilon|Z) \\ &= \lim_{Z\downarrow z_0} 1E(\beta_2\epsilon|Z) - 0 \lim_{Z\uparrow z_0} E(\beta_2|Z) + \lim_{Z\downarrow z_0} E(\epsilon|Z) - \lim_{Z\uparrow z_0} E(\epsilon|Z) \\ &= \lim_{Z\downarrow z_0} E(\beta_2|Z). \end{aligned}$$

Cette quantité est environ égale à celle qui nous intéresse, l'ECM au point z_0 (Imbens et Wooldridge, 2009, 62, 65).

On ne va pas calculer une différence de limites, $\lim_{Z\downarrow z_0} E(Y|Z) - \lim_{Z\uparrow z_0} E(Y|Z)$.

Ni faire passer une fonction 'globale' pour estimer un effet causal autour de z_0. Une possibilité serait de partir d'un modèle de régression pour chaque RP contrefactuel, avec polynôme de degré un en $Z - z_0$: $Y(0) = E(Y(0)|Z = z_0) + \beta_0(Z - z_0) + \epsilon_0$, et $Y(1) = E(Y(1)|Z = z_0) + \beta_1(Z - z_0) + \epsilon_1$.

Le centrage de Z est une normalisation assurant que l'effet du traitement au point z_0 peut-être mesuré par le coefficient devant D.

En effet, notons $\beta_1 \equiv E(Y(0)|Z = z_0)$, $\beta_2 \equiv E(Y(1) - Y(0)|Z = z_0)$, $\delta \equiv \beta_1 - \beta_0$ et $\epsilon = D\epsilon_1 + (1 - D)\epsilon_0$. En remplaçant $Y(0)$ et $Y(1)$ par ces deux fonctions dans l'équation de Neyman-Rubin, on obtient la régression à estimer : $E(Y|D, Z) = \beta_1 + D\beta_2 + (Z - z_0)\beta_0 + D(Z - z_0)\delta + \epsilon$. Ce serait problématique car nous obtiendrions une régression 'globale' (éloignée de z_0), et non locale, pour estimer un effet moyen (β_2 dans la régression) ; Wooldridge (2010, 956). Au point z_0, il reste $\beta_1 + D\beta_2$, de sorte que l'estimation de β_2 mesure le saut que fait Y en z_0.

On va faire passer localement (autour du point de discontinuité) deux fonctions de régression (régression linéaire locale), dans le sous-échantillon S_h, $h > 0$ tel que $Z \in [z_0 - h; z_0 + h]$ (*bandwidth*). En 'resserrant' l'échantillon autour de z_0, on comprend qu'il va falloir estimer deux régressions :

1) Pour $\boldsymbol{Z \in]z_0 - h; z_0[}$: régresser $\boldsymbol{Y}$ sur une constante et $\boldsymbol{Z - z_0}$. Dans ce cas, $\boldsymbol{Y = E(Y(0)|Z = z_0) + \beta_0(Z - z_0) + \epsilon_0}$.
2) Pour $\boldsymbol{Z \in [z_0; z_0 + h[}$: *idem* et $\boldsymbol{Y = E(Y(1)|Z = z_0) + \beta_1(Z - z_0) + \epsilon_1}$.

On peut calculer les ordonnées à l'origine et coefficient de pente de chaque modèle $E(Y(0)|Z = z_0)$, $E(Y(1)|Z = z_0)$, β_0 et β_1 par les moindre carrés. La différence des ordonnées à l'origine, i.e., $E(Y(1)|Z = z_0) - E(Y(0)|Z = z_0)$ est l'ECM au point z_0. En notant $z \coloneqq Z - z_0$, on peut élever z au carré, cube, en dehors des termes d'interaction et dans les termes d'interaction, que l'on place dans la régression suivante (une seule régression, selon une approche semi-paramétrique) :

$$Y = \alpha_1 + \alpha_2 z + \alpha_3 z^2 + \beta D + \gamma z D + \cdots + \epsilon.$$

L'intérêt d'une seule régression est de ne pas à avoir à calculer les erreurs standard à part.

Le calcul du h optimal dépend du noyau considéré le critère des moindres carrés ; voir Imbens et Wooldridge (2009, 65).

7.3.2. Application (Khandker, 2005)

Dans l'application que nous allons voir, un ménage n'est éligible à un programme de microcrédit que s'il possède des terres d'une surface totale < 50 decimals $= 0{,}5$ acre (1 decimal $= 40{,}46$ m^2). Alors, $z_0 = 50$, et $D = 1$ si $Z < z_0$.

On note que la condition d'être traité repose sur l'inégalité « $<$ », qui est opposée à celle de l'équation (7.13). Ce n'est pas grave, on permute les flèches « ↓ » et « ↑ ».

Le microcrédit de la banque Grameen est ciblé sur les petits propriétaires terriens (le seuil de 50 ci-dessus). Une bourse est disponible pour les élèves ayant des résultats suffisants et dont les parents ont des petits revenus.

[**Encadré** Khandker (2005), Khandker, Koolwal et Hussain (2010)]

Encadré : Microfinance et pauvreté

Khandker (2005), Khandker, Koolwal et Hussain (2010)

Les ménages pauvres qui ont accès à la microfinance sont-ils moins pauvres ? C'est la question (causale) à laquelle répondent les travaux menés par Shahidur Kahndker au Bangladesh en zones rurales. Son travail est dans le domaine de l'économie des pays en développement, domaines dans lequel on trouve des prédécesseurs, Angus Deaton, Alain de Janvry, Elisabeth Sadoulet, et plus récemment Esther Duflo. Il existe des travaux sur l'impact de la participation d'hommes et de femmes à des programmes de microfinance sur les dépenses du ménage, la scolarisation des enfants, etc. Un concept théorique important, développé par Amartya Sen, est celui des capabilités, *i.e*, la liberté des individus de choisir leur mode de vie.

L'application comporte deux bases de données

- Période 1991-1992 : 1800 ménages piochés aléatoirement dans un échantillon de 29 thanas : environ 1540 ménages de 24 thanas sont exposés au programme de microcrédit, et le reste des ménages (5 thanas) non-exposé.
- Période 1998-1999 : 1129 ménages piochés aléatoirement, avec une répartition en termes de thanas différente

Parmi les ménages exposés, 40 % environ n'ont pas participé au programme de microcrédit. Il est donc pertinent de distinguer éligibilité et participation. Cette distinction peut être utilisée comme source d'identification de l'effet causal, à condition que l'éligibilité (un instrument) soit exogène (cf. **chapitre 9**). Ici, la variable d'éligibilité (*forcing variable*), Z, est la surface agricole détenue par le ménage. Un ménage est éligibilité si $Z < 0{,}5$ acre ($10000\ m^2 = 1$ ha $= 2{,}47$ acre, donc 0,5 acre $\cong 2023\ m^2$). La valeur de Z d'un ménage détermine sa participation du ménage (le traitement), D.

L'évaluation considère le genre (femme ou homme dans l'article) comme variable de stratification. Il y a plusieurs variables de résultat : dépenses de consommation par tête, scolarisation des enfants, etc. Le fait que la randomisation ait eu lieu au niveau du genre (strate), chaque genre décidant de participer ou pas à l'intérieur de son groupe. Le genre joue un rôle important car certains villages sont genrés. Du coup, l'insertion d'un effet fixe (ou aléatoire) « village » est importante pour neutraliser l'influence potentiellement confondante du genre.

Les ménages tels que Z est dans le voisinage de 0,5, sont considérés suffisamment proches pour pouvoir supposer que leur participation est aléatoire. Les auteurs trouvent que le microcrédit a plus d'impact quand la population des bénéficiaires est féminine. En comparaison, l'éligibilité selon la surface agricole est moins fiable, sans doute à cause du problème de *nonenforceability*).

La mise à disposition de deux enquêtes successives permet :

- d'étudier la sensibilité des résultats de la première enquête, en les comparant à ceux issus de la deuxième enquête ;
- d'estimer un modèle en double-différence, méthode que nous verrons dans le **chapitre 8**.

Sur les conversions, voir http://www.fao.org/3/AC374E/AC374E19.htm

7.4. Exercice

7.4.1) Montrer que la randomisation du traitement D dans le modèle $Y = \beta_1 + D\beta_2 + \epsilon$, implique l'exogénéité de D par rapport à ϵ, c'est-à-dire, $Cov(\epsilon, D) = 0$. Montrer ensuite que si D n'est pas randomisée, il faut supposer l'exogénéité : $E(\epsilon|D) = 0$.

Correction de l'exercice du chapitre 7

7.4.1) Tout d'abord, la randomisation de D équivaut intuitivement à l'indépendance entre D et ϵ, que ϵ contienne des variables omises (VO) ou pas. Donc $Cov_{D,\epsilon}(D,\epsilon) = E_{D,\epsilon}(D\epsilon) - E_D(D)E_\epsilon(\epsilon) = E_D(D)E_\epsilon(\epsilon) - E_D(D)E_\epsilon(\epsilon) = 0$ (indépendance $\Rightarrow$ non-corrélation ; cf. **chapitre 2**).

En l'absence de randomisation, soit il n'y a pas de VO, alors $\epsilon \equiv Y - E(Y|D)$ où ϵ est une variable aléatoire complètement déterminée par Y et D. D est exogène par définition ! $E_{\epsilon|D}(\epsilon|D) = E_{Y|D}(Y - E(Y|D)|D) = E(Y|D) - E(Y|D) = 0$. Une autre conséquence est que ϵ est centrée : $E_\epsilon(\epsilon) = E_D(E_{\epsilon|D}(\epsilon|D)) = E_D(0) = 0$).

Supposons au contraire que dans $Y = \beta_1 + D\beta_2 + \epsilon$, la variable ϵ symbolise les VO dans le modèle $E(Y|D) = \beta_1 + D\beta_2$. Alors l'exogénéité (les VO ne sont pas corrélées avec D) $Cov_{D,\epsilon}(D,\epsilon) = 0$, ne découle plus (trivialement) du modèle, mais doit être supposée. En fait, il suffit de supposer $E_{\epsilon|D}(\epsilon|D) = 0$. En effet, $Cov_{D,\epsilon}(D,\epsilon) = E(D\epsilon) - E(D)E(\epsilon) = E\left(DE_{\epsilon|D}(\epsilon|D)\right) - E(D)E\big(E(\epsilon|D)\big) = E(D0) - E(D)E(0) = 0$.

8. Différence de différences et contrôle synthétique

On rencontre la méthode de différence de différences (DiD par la suite, aussi appelée double différence) dans les EXN (cf. **sous-section 2.3.4**) quand un événement naturel, exogène pour les individus (pas anticipé), sépare ces derniers en traités et non-traités. Selon le problème économique, nous devrons plutôt parler de séparation entre individus « éligibles » qui participent et « non-éligibles » qui ne participent pas (il y a toujours une part d'auto-sélection (un éligible qui ne participe pas, etc.). Voyons des exemples :

> L'étude de Card et Krueger (1994), reprise dans Imbens et Rubin (2015) pour illustrer l'estimateur d'appariement (cf. **section 6.3**), utilise un protocole DiD : la hausse du « SMIC » dans l'État du New Jersey sépare les entreprises entre celles de cet État (éligibles) et de Pennsylvanie (non-éligibles). Mais ne « participent » (sont traitées) que celles dont des salariés ont moins que le SMIC.
>
> **[Card et Krueger, 1994, p. 780]**
>
> L'étude de Vandenbussche et Viegelahn (2015) des effets de la politique antidumping de l'UE en 2013, en réponse au dumping d'équipementiers chinois en cellules photovoltaïques et panneaux solaires, pousse des importateurs à substituer des intrants non-taxés aux intrants de Chine taxés où à sortir du marché.
>
> Le soutien public en faveur de la R&D, *via* le CIR introduit en 1983, suivi des réformes de 1999, 2004 , 2006 et 2008 (Bozio *et alii*, 2019). L'effet de la réforme de 2008 peut être étudié en exploitant le fait que le recours au CIR n'est pas obligatoire. Si cet effet n'est pas anticipé, on peut parler d'EXN.
>
> Certains événements inattendus créent des situations d'EXN, qui peuvent être facilement analysées à partir d'un protocole en différence de différences : la guerre en Irak, la « crise de la vache folle », la Covid-19 (interdiction de voyager …), etc.

Deux points importants à noter.

> Le **passage du temps** est modélisé dans le protocole, contrairement aux modèles vus dans les chapitres précédents. Par ex., dans un DiD à deux périodes, le temps intervient sous la forme d'une variable dichotomique indiquant ces périodes.
>
> L'autre point, sur lequel nous reviendrons, est que la méthode DiD permet de contrôler des variables non-observées conditionnant la sélection dans les groupes de traitements. La plupart du temps, le contrôle se fait avec des effets fixes.

Le protocole DiD est introduit en économie par O. Ashenfelter en 1978, mais sans être nommé comme tel. D. Card (Prix Nobel d'Economie en 2021) et O. Ashenfelter utilisent *difference-in-difference* pour la première fois en 1985. Deux articles très cités sont ceux de Card et Krueger (1994), bien exposé dans Angrist et Pischke (2009), dont nous avons déjà parlé, et Bertrand *et alii* (2004). Un article de référence pour comprendre la différence entre un DiD et un protocole avant-après est Meyer (1995).

La **section 1** sera consacrée à des applications de la méthode DiD dans le cas dit 2 × 2 (deux groupes sur deux périodes de traitement). Nous ferons une analyse un peu plus théorique des résultats dans la **section 2**. Dans la **section 3**, nous verrons la méthode du contrôle synthétique d'Abadie *et alii* (2010), qui est une des généralisation possibles de la méthode DiD ; nous répliquerons cet article.

Dès lors que l'on a plus de deux périodes se pose notamment deux question :

- **Une fois un individu traité, l'est-il jusqu'à la dernière date d'observation ?**
- **Les individus sont-ils traités à la même date ?**

Les protocoles les plus en vogue pour traiter de ces questions sont appelés *event study design* et *DiD with staggered rollout* traduction ?

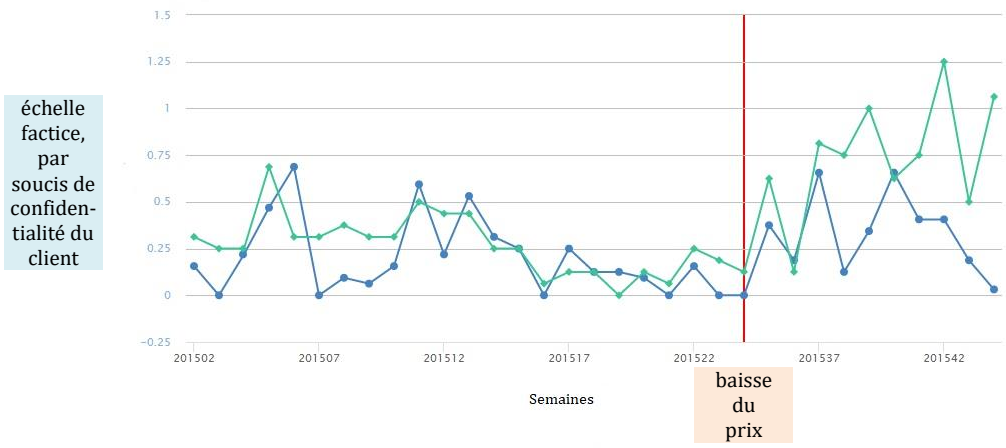
8.1. Exemples d'applications

Nous allons voir deux exemples d'applications de la méthode DiD. Un bref exemple d'application dans le privé, puis une évaluation bien connue de Kiel et McClain (1995), reprise dans les ouvrages de Wooldridge (2003, 2010) et Wooldridge (2009).

8.1.1. Politique de prix dans la vente au détail

Dans le secteur des services aux Pays-Bas, le cabinet de conseil Veneficus utilise la méthode DiD pour aider des entreprises à optimiser leur offre : quel est l'effet de la baisse du prix de certains produits sur les ventes ? sur le chiffre d'affaires ?

Figure 8.1 : évolution hebdomadaire du chiffre d'affaires dans les groupes de traitement avant et après le changement de prix (01-10/2015)



Dans cette évaluation, l'intervention a consisté à changer les prix des biens d'un groupe de magasins et de comparer l'effet à celui de magasins dans lesquels le changement n'a pas lieu. L'entreprise peut ainsi se faire une idée du résultat contrefactuel, i.e. du résultat si le changement de prix n'avait pas eu lieu (la courbe bleue dans la **figure 8.1**). Cette méthode aurait permis d'améliorer de 20 % la marge bénéficiaire du client.

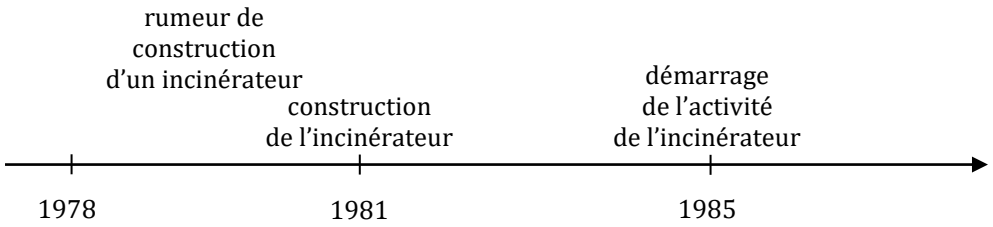
8.1.2. Politique locale d'urbanisme

Wooldridge (2009, 450-) applique la méthode DiD à une EXN étudiée d'abord par Kiel et McClain (1995). Il s'agit d'un problème d'économie urbaine touchant au marché

du logement résidentiel (JEL Code R3). L'étude porte sur l'effet de l'installation d'une déchetterie sur le prix de vente de maisons à North Andover (ville du comté d'Essex dans l'Etat du Massachussets).

[Lancer « **did_incinerator.R** »]

Figure 8.2 : découpage des événements



La variable de résultat $Y_{i,t}$ (« rprice », V24) est le prix de vente d'une maison i en $t = 1978$ ou $t = 1981$, mesuré en $ constant de l'année 1978. C'est un pseudo-panel, mais on fait comme si c'était un panel dans cette présentation.

Le traitement est une externalité négative de l'incinérateur de déchets, d'autant plus élevée que i est proche de ce dernier. Le traitement expérimental est la variable continue (cf. **chapitre 6**) « dist » (V13), ou discrétisée D_i (« nearinc », V22), qui prend la valeur 1 si la maison se situe à moins de 4,82 km de l'incinérateur (15840 *feet* = « three miles » dans Wooldridge (2009, 450)) et 0 sinon.

Testons l'hypothèse nulle (H_0) que le traitement n'a pas d'effet à partir de la différence des prix moyens en 1981 :

Figure 8.3 : estimateur de différence des moyennes de l'effet de l'incinérateur sur le prix des maisons *

```
        Two Sample t-test

data:  RPRICE by NEARINC
t = 5.2659, df = 140, p-value = 5.139e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 19166.58 42209.97
sample estimates:
mean in group 0 mean in group 1
      101307.51        70619.24
```

* Obtenu sous R avec la commande `t.test(RPRICE~NEARINC, subset(DATA, YEAR==1981), mu=0, paired=F, var.equal=T)`.

On rejette H_0 au seuil de 5 %. L'estimation *après* de l'effet, $\bar{Y}_{1,1981} - \bar{Y}_{0,1981} = 70619\$ - 101307\$ = -30688\$$ suggère qu'une maison proche de l'incinérateur coûte 30 000$ de moins en moyenne. L'erreur standard est environ égale à 5828, ce qui conduit à la statistique de Student de 5,3 (la valeur est positive car R prend $D = 0$ comme diminuende).

Malgré le rejet de H_0, on ne peut parler d'effet causal. En effet, le même test, mais pour 1978, avant la diffusion de la rumeur de construction d'un incinérateur (voir la **Figure 8.2**), est aussi significativement négatif : $\bar{Y}_{1,1978} - \bar{Y}_{0,1978} = 63692\$ - 82517\$ = -18825\$$ (p-valeur < 0,01). L'erreur standard vaut environ 4745.

> Ce dernier résultat s'explique de deux manières : la possibilité qu'un incinérateur s'installe apparu dans la presse locale fin 1978. La zone accueillant le futur incinérateur devait avoir une double classification, i.e. permettant un usage résidentiel et industriel, ce qui décroit la valeur des terrains concernés.

Il est alors clair que l'effet mesuré pour 1981 combine celui de l'incinérateur et de la nature des terrains indépendamment de la construction de l'incinérateur. Retirons cet effet « nature des terrains » ou effet groupe en 1978,de lui-même en 1981 :

$$\bar{Y}_{1,1981} - \bar{Y}_{0,1981} - \left(\bar{Y}_{1,1978} - \bar{Y}_{0,1978}\right) = -30688{,}27 - (-18824{,}37) = -11863{,}9,$$

Il y a aussi un effet « passage du temps » (effet temporel), qui est contrôlé. En effet, le prix réel des maisons a augmenté (hors de tout effet de l'incinérateur) : $\bar{Y}_{0,1981} = 101307{,}51 > \bar{Y}_{0,1978} = 82517{,}23$. La différence $\bar{Y}_{0,1981} - \bar{Y}_{0,1978}$ reflète l'appréciation des maisons, déterminée par l'attractivité du territoire, la santé de l'économie locale, de meilleures infrastructures, etc. (des facteurs non-observés, additifs, propres à chaque groupe de maisons). Pour le percevoir, il suffit de permuter les 2e et 3e termes ci-dessus :

$$\bar{Y}_{1,1981} - \bar{Y}_{1,1978} - \left(\bar{Y}_{0,1981} - \bar{Y}_{0,1978}\right) = 6926{,}38 - 18790{,}29 = -11863{,}9.$$

On a contrôlé pour une tendance commune aux deux groupes de maisons. La différence $\bar{Y}_{1,1981} - \bar{Y}_{1,1978} = 6926{,}38 = -11863{,}9$ est l'estimateur avant-après de l'effet de l'incinérateur. Cet estimateur est biaisé car il inclut l'effet de l'incinérateur sur les maisons proches plus l'effet temporel, qui est neutralisé grâce à $\bar{Y}_{0,1981} - \bar{Y}_{0,1978}$.

On lui préfère l'estimateur DiD de cet effet, dont on peut tester la significativité en estimant les coefficients du modèle linéaire multiple suivant par les moindres carrés ordinaires (nous reviendrons dessus plus loin, dans la **section 2**) :

$$Y_{i,t} = \beta_1 + \beta_2 D_i + \beta_3 D_t + \beta_4 D_i D_t + \beta_5 X_{i,t} + \epsilon_{i,t}$$

où $D_i = 1$ si i est près de l'incinérateur, 0 sinon. Et $D_t = 1$ si $t = 1981$ et 0 sinon (1978).

Aucune autre variable que D_i et D_t ne figurant dans le modèle, les coefficients s'interprètent facilement. Les estimations de ces coefficients sont exactement fonctions des moyennes avant $(\bar{Y}_{1,1978}, \bar{Y}_{0,1978})$ et après $(\bar{Y}_{1,1981}, \bar{Y}_{0,1981})$:

$\hat{\beta}_1 = \bar{Y}_{0,1978} = 82517{,}23$: le prix des maisons loin du futur incinérateur
$\hat{\beta}_2 = \bar{Y}_{1,1978} - \bar{Y}_{0,1978} = -18824$: effet de la proximité au futur incinérateur
$\hat{\beta}_3 = \bar{Y}_{0,1981} - \bar{Y}_{0,1978} = 18790$: variation du prix, hors effet de l'incinérateur
$\hat{\beta}_4 = \bar{Y}_{1,1981} - \bar{Y}_{0,1981} - \left(\bar{Y}_{1,1978} - \bar{Y}_{0,1978}\right) = -11863{,}9$: effet de l'incinérateur, hors effet « terrain »

Figure 8.4 : estimations DD (obtenues sous R)

```
> summary(lm(RPRICE ~ NEARINC + Y81 + NEARINC*Y81))

Call:
lm(formula = RPRICE ~ NEARINC + Y81 + NEARINC * Y81)

Residuals:
   Min     1Q Median     3Q    Max
-60678 -17693  -3031  12483 236307

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     82517       2727  30.260  < 2e-16 ***
NEARINC        -18824       4875  -3.861 0.000137 ***
Y81             18790       4050   4.640 5.12e-06 ***
NEARINC:Y81    -11864       7457  -1.591 0.112595
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30240 on 317 degrees of freedom
Multiple R-squared:  0.1739,    Adjusted R-squared:  0.1661
F-statistic: 22.25 on 3 and 317 DF,  p-value: 4.224e-13
```

b_1
b_2
b_3
b_4

Les résultats dans R de **`summary(lm(RPRICE ~ NEARINC + Y81 + NEARINC*Y81))`** montrent que le test bilatéral pour l'hypothèse H_0 d'absence d'effet de l'incinérateur (la variable **`NEARINC:Y81`**) conduit à ne pas rejeter H_0 (le Student vaut $-1{,}59$).

> Ce serait dommage d'accepter à tort (**risque de deuxième espèce**). Le non-rejet de H_0 n'est pas vraiment attribuable à la taille de l'échantillon (env. 300 obs.), mais plutôt à $H_1 : \beta_4 \neq 0$, qui n'est pas une hypothèse alternative très pertinente. Nous savons bien qu'une fois un incinérateur installé, les maisons qui sont près perdront de la valeur. Le test unilatéral suivant : $H_0 : \beta_4 = 0$, $H_1 : \beta_4 < 0$, est préférable, toujours au seuil de 5 %. Quelle est la valeur critique tabulée c : $\Pr(T(317) > c) = 0{,}95$? Calculons **`qt(0.95, 317, lower.tail = TRUE, log.p = FALSE)`**. On trouve 1,649. Or $-1{,}649 < -1{,}591$, donc on ne rejette pas.

On réestime le modèle (code **`did_incinerator.R`**) en contrôlant pour des facteurs supplémentaires (âge, nombre de pièces de la maison, etc.) L'âge compte car moins de nouvelles constructions près de l'incinérateur ⇒ le parc immobilier vieillit plus vite près de l'incinérateur ⇒ prix plus bas. La différence normalisée pour l'âge en 1981 vaut 0,77 (une différence d'âge environ égale à 0,8 écart-type, ce qui est proche de 1). L'effet passe à -21920, c'est une grosse différence.

[Lancer « **`did_incinerator.do`** »]

8.2. Discussion théorique

L'estimateur $\widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}$ est celui de l'effet du traitement sur les traités. C'est ce que nous allons voir avant de rentrer dans des questions de spécification du modèle. Puis, nous schématiserons la méthode de la différence de différences (DiD). La distinction entre panel et pseudo-panel peut compter dans la discussion.

8.2.1. Différence de différences et identification de l'ECMT

Nous pouvons nous appuyer sur l'écriture du modèle de régression avec RP pour montrer que $\boldsymbol{\beta}_{\boldsymbol{4}}$ est bien l'effet du traitement sur les traités. Pour écrire un modèle de régression avec RP, nous nous inspirons du modèle standard d'Imbens et Wooldridge (2009) sans variables de contrôle pour ne pas compliquer la discussion (*ibidem*, p. 67). Mais surtout, nous généralisons l'idée de Wooldridge (2003) développée dans le chapitre 7 au cas de deux périodes, ce qui rend l'équation de Rubin, qui relie le résultat observé aux résultats potentiels, plus compliquée. Soit :

$$\boldsymbol{Y_{i,2}(0) = \beta_1 + \epsilon_{i,t}},$$
$$\boldsymbol{Y_{i,2}(1) = \beta_1 + \beta_2 + \epsilon_{i,t}},$$
$$\boldsymbol{Y_{i,3}(0) = \beta_1 + \beta_3 + \epsilon_{i,t}},$$
$$\boldsymbol{Y_{i,3}(1) = \beta_1 + \beta_2 + \beta_3 + \beta_4 + \epsilon_{i,t}},$$
$$\boldsymbol{Y_{i,t} = \left[Y_{i,3}(1)D_i + Y_{i,3}(0)(1 - D_i)\right]D_t + \left[Y_{i,2}(1)D_i + Y_{i,2}(0)(1 - D_i)\right](1 - D_t)},$$
$$\boldsymbol{i = 1, \dots, N, t = 2, 3}$$

Combinons ces équations :

$$\boldsymbol{Y_{i,t} = \beta_1 + \beta_2 D_i + \beta_3 D_t + \beta_4 D_i D_t + \epsilon_{i,t}}, \qquad \textbf{(8.1)}$$

Nous retombons sur le même résultat qu'Imbens et Wooldridge (2009), à la différence près que ces auteurs ne font pas apparaître l'indice temporel $\boldsymbol{t}$. De plus, ils n'écrivent pas l'équation de Rubin (est-elle implicite dans leur modèle ?). Du coup, la variable $\boldsymbol{D_{i,t} \equiv D_i D_t}$ est introduite dans (8.1) chez eux, alors que chez nous, elle apparaît dans l'équation de Rubin, de la même manière que $\boldsymbol{D_i}$ apparaissait dans $\boldsymbol{Y_i(1)D_i + Y_i(0)(1 - D_i)}$.

L'effet temporel, $\boldsymbol{\beta_3}$, est commun aux deux groupes ; il est dans $\boldsymbol{Y_{i,3}(0)}$ et $\boldsymbol{Y_{i,3}(1)}$. Supposons que $\boldsymbol{E\left(\epsilon_{i,t} \middle| D_i, D_t\right) = 0}$, $\forall\, \boldsymbol{i, t}$ (exogénéité stricte). Ecrivons la différence de différences de moyennes suivante :

$$\begin{aligned}
&\boldsymbol{E\left(Y_{i,t} \middle| D_i = 1, D_t = 1\right) - E\left(Y_{i,t} \middle| D_i = 0, D_t = 1\right) - \left[E\left(Y_{i,t} \middle| D_i = 1, D_t = 0\right) - E\left(Y_{i,t} \middle| D_i = 0, D_t = 0\right)\right]} \\
&\boldsymbol{= \beta_1 + \beta_2 + \beta_3 + \beta_4 - (\beta_1 + \beta_3) - [\beta_1 + \beta_2 - \beta_1]} \\
&\boldsymbol{= \beta_2 + \beta_4 - [\beta_2]} \\
&\boldsymbol{= \beta_4.}
\end{aligned}$$

Or, $\boldsymbol{\beta_4}$ est le coefficient de $\boldsymbol{D_{i,t}}$, qui ne s'active que pour les individus traités en $\boldsymbol{t = 3}$. Par conséquent, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}$ est un estimateur de l'ECMT ; voir également Angrist et Pischke (2009, 234). On peut obtenir cette estimation de différentes manières : en prenant les différences de moyennes de $\boldsymbol{Y}$ calculées dans chaque sous-population. Nous pouvons aussi passer par les MCO.

8.2.2. Questions de spécification du modèle

La spécification dépend de la nature des données : soit un panel (*panel/longitudinal data*) ou un pseudo-panel (coupes instantanées empilées, *repeated/pooled cross sections*)

- Panel : intervention sur des prix d'un chaîne de magasins (mêmes magasins avant et après la manipulation des prix).
- Pseudo-panel : évaluation d'une externalité sur les valeurs immobilières (il a pu y avoir de nouvelles constructions entre 1978 et 1981).

Le protocole DiD change peu selon que l'on affaire à l'un ou l'autre type de données ; voir Lee (2016, 135) et Cerulli (2015). Il y a un hypothèse d'identification cruciale de l'effet causal : contrôler la tendance commune, ce que fait la différence $\boldsymbol{E}\left(\boldsymbol{Y_{i,t}}\middle|\boldsymbol{D_i}=\boldsymbol{0},\boldsymbol{D_t}=\boldsymbol{1}\right)-\boldsymbol{E}\left(\boldsymbol{Y_{i,t}}\middle|\boldsymbol{D_i}=\boldsymbol{0},\boldsymbol{D_t}=\boldsymbol{0}\right)$. Mais dans le panel, il existe probablement une corrélation sérielle. Dans un pseudo panel, $\boldsymbol{Cov}\left(\boldsymbol{\epsilon_{i,t}},\boldsymbol{\epsilon_{i,t-1}}\right)=\boldsymbol{0}$ est plausible. Dans un panel, on a $\neq \boldsymbol{0}$, mais on ne peut pas correctement éliminer cette corrélation avec deux dates.

Contrôle des effets (fixes) individuels : Dans un panel avec constante, l'effet individuel peut être absorbé à l'aide de la **transformation within**. Avec le protocole DiD $\boldsymbol{2\times 2}$, les effets individuels sont absorbés par les différences de moyennes. Dans l'exemple précédent, on a la différence $\overline{\boldsymbol{Y}}_{\boldsymbol{g,1981}}-\overline{\boldsymbol{Y}}_{\boldsymbol{g,1978}}$, $\boldsymbol{g}\in\{\boldsymbol{0};\boldsymbol{1}\}$. Dans un panel, on peut régresser $\boldsymbol{\Delta Y_{i,t}}$ sur $\boldsymbol{1},\boldsymbol{D_i}$ (Baltagi, 2021, 20). Le coefficient devant $\boldsymbol{D_i}$ sera $\overline{\boldsymbol{\Delta Y_{1,t}}}-\overline{\boldsymbol{\Delta Y_{0,t}}}=\overline{\boldsymbol{Y}}_{\boldsymbol{1,t}}-\overline{\boldsymbol{Y}}_{\boldsymbol{1,t-1}}-(\overline{\boldsymbol{Y}}_{\boldsymbol{0,t}}-\overline{\boldsymbol{Y}}_{\boldsymbol{0,t-1}})$.

On ne peut pas estimer les effets individuels à cause de la multicolinéarité entre ces effets et le vecteur de traitement.

Soit $\boldsymbol{e}_i$ l'effet individuel associé à i (un vecteur $N \boxtimes 1$ dont la $i$ème valeur vaut 1), et $\boldsymbol{D}$ est le vecteur de traitement (vecteur $N \boxtimes 1$ qui vaut 1 pour tous les $i: D_i = 1$). Le vecteur $\boldsymbol{D} = (D_1, \dots, D_N)'$ peut se décomposer dans la base canonique $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_N)$: $\boldsymbol{D} = \sum_i D_i \boldsymbol{e}_i$. Or, $D_i \in \{0; 1\}$, donc on a $\sum_i D_i \boldsymbol{e}_i = \sum_{i:D_i=1} D_i \boldsymbol{e}_i$. Par conséquent $\boldsymbol{D}$ et $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{N_1})$ sont liés. ■

L'estimateur des MCO est un tout petit peu difficile à obtenir, car il y a quatre coefficients à estimer. Cependant, comme les variables ne prennent que des valeurs égales à 0 ou à 1, de nombreuses simplifications apparaissent lors de la résolution des équations normales.

[Exercice 8.4.2]

8.2.3. Schématisation de la différence des différences

Les trois tableaux suivants montrent, de manière schématique, comment la différence des différences (de moyennes) contrôle les effets groupe ET individuel et temporel. Afin que ce ne soit pas trop abstrait, on applique le schéma à l'EX de l'incinérateur :

Soient :

- β_2 : l'effet groupe distinguant le groupe des maisons proches de celles éloignées, indépendamment de l'effet de la construction de l'incinérateur ;
- β_3 : l'effet temporel distinguant les années 1978 et 1981. C'est l'appréciation/dépréciation des maisons indépendamment de l'inflation ;
- β_1 : le prix (moyen) des maisons loin de l'incinérateur, avant (en 1978);
- β_4 : l'effet de l'incinérateur sur les maisons proches en 1981 (ECMT).

Lee (2016, 137) appelle β_2 et β_3 des effets marginaux, que l'on retire en séquence par chacune des différences. La question est : commet obtenir β_4 ?

L'estimateur DiD est pour la situation compliquée où il faudrait contrôler pour des facteurs non observables β_2 et β_3. Il marche dans les différents cas : un effet groupe-pas d'effet temporel, pas d'effet groupe-un effet temporel, les deux.

Il y a donc trois cas ; on commence toutefois par le tableau « aucun effet ».

1) Pas d'effet groupe, ni temporel.

		D_i	
		0 (témoin)	1 (test)
D_t	0 (1978)	β_1	β_1
	1 (1981)	β_1	$\beta_1 + \beta_4$

Deux solutions simples pour identifier $\boldsymbol{\beta_4}$: $\boldsymbol{\beta_1 + \beta_4 - \beta_1}$. Pour l'année 1981, on soustrait le prix des maisons éloignées de celui des maisons proches (par différence de moyennes en 1981 ou régression de $\boldsymbol{Y_{i1981}}$ sur 1,$\boldsymbol{D_i}$). Ou, pour les maisons proches, on retire le prix de 1978 à celui de 1981 (estimateur avant-après, régression de $\boldsymbol{Y_{i:D_i=1,t}}$ sur 1,$\boldsymbol{D_t}$). Plus compliquée : régresser $\boldsymbol{Y_{it}}$ sur 1,$\boldsymbol{D_i D_t}$.

2) Un effet groupe, pas d'effet temporel.
(différences permanentes entre les groupes), adoption : Imbens et Wooldridge (2009, 67)

		D_i	
		0	1
D_t	0	β_1	$\beta_1 + \beta_2$
	1	β_1	$\beta_1 + \beta_2 + \beta_4$

Solution pour obtenir $\boldsymbol{\beta_4}$: $\boldsymbol{\beta_1 + \beta_2 + \beta_4 - (\beta_1 + \beta_2)}$. Pour les maisons proches, on soustrait le prix de 1978 de celui de 1981. Comme précédemment, on peut faire une régression de $\boldsymbol{Y_{i:D_i=1,t}}$ sur 1,$\boldsymbol{D_t}$ (une dummy temporelle élimine l'effet groupe). On retrouve l'estimateur avant-après.

On peut retenir que l'estimateur avant-après ne s'applique qu'au groupe test

3) Pas d'effet groupe, un effet temporel.

		D_i	
		0	1
D_t	0	β_1	β_1
	1	$\beta_1 + \beta_3$	$\beta_1 + \beta_3 + \beta_4$

La solution est : $\boldsymbol{\beta_1 + \beta_3 + \beta_4 - (\beta_1 + \beta_3)}$. On régresse $\boldsymbol{Y_{i1981}}$ sur 1,$\boldsymbol{D_i}$ (une dummy groupe élimine l'effet temporel)

4) Effet groupe, effet temporel (indépendant du traitement).

		D_i	
		0	1
D_t	0	β_1	$\beta_1 + \beta_2$
	1	$\beta_1 + \beta_3$	$\beta_1 + \beta_2 + \beta_3 + \beta_4$

Deux solutions équivalentes pour estimer $\boldsymbol{\beta_4}$:

(i) $\boldsymbol{\beta_1 + \beta_2 + \beta_3 + \beta_4 - (\beta_1 + \beta_3) - [(\beta_1 + \beta_2) - \beta_1]}$: pour l'année 1981, on soustrait le prix des maisons éloignées de celui des maisons proches. On fait pareil pour 1978, puis on retire cette deuxième différence de la première ;

(ii) $\boldsymbol{\beta_1 + \beta_2 + \beta_3 + \beta_4 - (\beta_1 + \beta_2) - [(\beta_1 + \beta_3) - \beta_1]}$: pour les maisons proches, on soustrait le prix des maisons de 1978 de celui de 1981. On fait pareil pour les maisons éloignées puis on soustrait cette différence de la première.

8.2.4. Limites du protocole avant-après

Le lien avec le MCR, afin de faire apparaître l'effet causal, n'est pas compliqué. Faisons un petit détour par l'estimateur avant-après, que nous venons de voir de manière schématique. Comme dans Lee (2016), on note la période « avant » $t = 2$, et celle après, $t = 3$ (évite de confondre les indices temporels et de groupe i: $D_i \in \{0\,;1\}$).

> L'auteur distingue l'éligibilité (*qualification*) au groupe test, de la sélection dans ce groupe, ce que nous ne ferons pas ici mais dans le **chapitre 9**.

Meyer (1995) suppose le modèle $Y_{it} = \beta_1' + D_t\beta_4' + U_{it}$, avec $D_t = 1$ si $t = 3$, la date où un événement conduit tous les i à être traités. U_{it} est un terme d'erreur. Il s'intéresse aux conditions auxquelles la mesure de β_4', obtenue par les MC par ex. $(E(Y_{it}|D_t = 1) - E(Y_{it}|D_t = 0))$, a une interprétation causale. Pour Meyer (1995), cette condition est $E(U_{it}|D_t) = 0$ (4 conditions), et qu'il n'y ait pas d'effet en l'absence de traitement (pas d'effet en $t = 2$).

C'est une condition forte, que l'on trouve facilement en combinant ce modèle avec le MCR. Supposons que $\forall i$, $Y_{i2}(0) = \beta_1' + U_{i2}$, $Y_{i3}(1) = \beta_1' + \beta_4' + U_{i3}$, et $Y_{it} = D_tY_{i3}(1) + (1 - D_t)Y_{i2}(0)$. On remarque déjà que l'effet causal $Y_{i3}(1) - Y_{i3}(0)$ n'est défini que sous l'hypothèse $Y_{i3}(0) = Y_{i2}(0)$ car tout le monde est traité en $t = 3$ (pas de groupe témoin). Cette hypothèse a pour implication immédiate (cf. **8.2.1**) qu'il n'y a pas d'effet temporel ! Ecrivons l'effet causal : $Y_{i3}(1) - Y_{i2}(0) = \beta_4' + U_{i3} - U_{i2}$.

En mettant les première et deuxième équations dans la troisième, on obtient :

$$\begin{aligned} Y_{it} &= D_t[\beta_1' + \beta_4' + U_{i3}] + (1 - D_t)[\beta_1' + U_{i2}] \\ &= \beta_1' + D_t\beta_4' + U_{it}, \text{ avec } U_{it} \coloneqq D_tU_{i3} + (1 - D_t)U_{i2}. \end{aligned}$$

Alors, l'EMT $E(Y_{it}|D_t = 1) - E(Y_{it}|D_t = 0)$ est égal à :

$$\begin{aligned} &= E(Y_{i3}(1) - Y_{i3}(0)|D_t = 1) + E(Y_{i3}(0)|D_t = 1) - E(Y_{i2}(0)|D_t = 0) \\ &= E(Y_{i3}(1) - Y_{i2}(0)|D_t = 1) + E(Y_{i2}(0)|D_t = 1) - E(Y_{i2}(0)|D_t = 0) \\ &= \beta_4' + E(U_{i3} - U_{i2}|D_t = 1) + E(U_{i2}|D_t = 1) - E(U_{i2}|D_t = 0) \\ &= \beta_4' + E(U_{i3}|D_t = 1) - E(U_{i2}|D_t = 0), \end{aligned}$$

qui est l'effet causal plus la différence avant-après des erreurs espérées. S'il y a un effet groupe (β_2) et individuel (μ_i), ils seront neutralisés. On a 2 conditions sur les erreurs, pas 4 ! Une solution est de trouver un groupe témoin jamais traité.

8.2.5. Hypothèses d'identification : « ignorabilité », « tendance commune »

Toujours avec $t = 2,3$, et $D = 1$ si l'individu est éligible et dans le groupe test. La dummy temporelle est notée D_3, qui vaut 1 en $t = 3$ et 0 en $t = 2$. Sans perte de généralité, on n'ajoute qu'une variable de confusion X au modèle. Alors, la différence

$$E(Y|D = 1, D_3 = 1, X) - E(Y|D = 1, D_3 = 0, X) - [E(Y|D = 0, D_3 = 1, X) - E(Y|D = 0, D_3 = 0, X)], \quad (8.2)$$

qui soustrait la « tendance » du groupe témoin à celle du groupe test, est égale à

$$E(Y(1)|D = 1, D_3 = 1, X) - E(Y(0)|D = 1, D_3 = 0, X) - [E(Y(0)|D = 0, D_3 = 1, X) - E(Y(0)|D = 0, D_3 = 0, X)] \quad (8.3)$$

Les individus ne sont éligibles qu'en $t = 3$

On n'a pas encore fait apparaître l'ECMT, car D_3 change. C'est pour cela que l'on ajoute et soustrait le résultat contrefactuel $E(Y(0)|D = 1, D_3 = 1, X)$. On obtient :

$$E(Y(1)|D = 1, D_3 = 1, X) - E(Y(0)|D = 1, D_3 = 1, X) + E(Y(0)|D = 1, D_3 = 1, X) - E(Y(0)|D = 1, D_3 = 0, X) - [E(Y(0)|D = 0, D_3 = 1, X) - E(Y(0)|D = 0, D_3 = 0, X)].$$

Il s'agit de l'ECMT en $t = 3$, +
l'effet temporel sur les individus du groupe test s'ils n'avaient pas été traités −
l'effet temporel sur les individus témoins (une quantité observable).

La condition d'identification est simplement que la différence entre les effets temporels soit égale à zéro. C'est la supposition de tendance commune (CTA) !

Nous constatons que nous n'avons pas besoin de CIA, ce qui montre l'efficacité du protocole DiD. Dans l'**exercice 8.4.1**, vous êtes invités à montrer que sans CTA, mais sous CIA, $E(Y(d)|D, D_3, X) = E(Y(d)|D_3, X)$, la quantité (**8.2**) mesure l'ECMT.

CTA peut être vérifiée. Pour cela, il nous faut un modèle pour $Y(0)$. Supposons un modèle de régression $Y_{it}(0) = \beta_{10} + \beta_{30}D_3 + \beta_{50}X_{it} + \varepsilon_{it}$. Alors CTA implique :

$$\beta_{10} + \beta_{30} + \beta_{50}E(X_{i3}|D = 1) - \left(\beta_{10} + \beta_{50}E(X_{i2}|D = 1)\right) = [\beta_{10} + \beta_{30} + \beta_{50}E(X_{i3}|D = 0) - (\beta_{10} + \beta_{50}E(X_{i2}|D = 0)],$$

ce qui revient à $E(X_{i3}|D = 1) - E(X_{i2}|D = 1) - [E(X_{i3}|D = 0) - E(X_{i2}|D = 0)] = 0$.

C'est un DiD sur X. Wing *et alii* (2018, 460) suggèrent de vérifier CTA en testant la nullité de β_4^X dans :

$$X_{it} = \beta_1^X + \beta_2^X D_i + \beta_3^X D_t + \beta_4^X D_i D_t + \varepsilon_{it}^X.$$

Pour estimer l'ECMT, on insère X dans (8.1), ce qui donne le modèle suivant :

$$Y_{it} = \beta_1 + \beta_2 D_i + \beta_3 D_t + \beta_4 D_i D_t + \beta_5 X_{it} + \varepsilon_{it}. \quad (8.4)$$

Ces deux modèles sont estimables par la méthode des MCO. Voir l'**exercice 8.4.3** pour l'estimateur de (**8.4**).

8.3. Contrôle synthétique

Abadie, Diamond et Hainmueller (2010), Abadie et Gardeazabal (2003) ont introduit la méthode du contrôle synthétique (MCS). À une date donnée, un événement sépare la population en deux groupes de traitement. Généralisation du protocole DiD par un modèle à facteurs (effets individuels variant dans le temps, effets temporels communs).

PFIC ⇒ on estimer un RP contrefactuel $E(Y(0)|D=1)$ dans le temps.

Le nombre d'individus dans la population, et *a fortiori* l'échantillon, n'est pas grand (quelques dizaines en pratique, rarement quelques centaines) :

Un individu traité (qui peut être un agrégat d'individus), $N_1 = 1$.

Plus généralement, l'événement a lieu à un niveau agrégé. Deux exemples :

(i) Des États (agrégation d'entreprises) : Card et Krueger (1994), qui étudient le salaire moyen des travailleurs de restaurants dans deux états américains, avec le New Jersey comme groupe test.

(ii) Des communes (agrégation de demandeurs d'emploi) : Gobillon et Magnac (2016) étudient le taux de chômage de communes françaises (le traitement c'est l'appartenance à une Zones franches urbaine).

Un petit réservoir d'individus témoins sélectionnés automatiquement

On écarte méticuleusement (sélection qualitative) des témoins, puis appariement de type NNM (**chapitre 6**), donc transparent et précis. Ce que Abadie, Diamond et Hainmueller (2010) et Abadie (2021) appellent *a comparative case study*. Problème : cela réduit la taille de l'échantillon, qui n'est déjà pas grand au départ.

⇒ méthode gourmande en temps de calcul !

Nombre élevé de périodes (ou pas, quelques dizaines en général)

Gobillon et Magnac (2016) considèrent des données mensuelles sur la période juillet 1989-juin 2003 ($\boldsymbol{T = 168}$). Ils appliquent la méthode à 148 communes françaises (après un pré-appariement), dont 13 traitées, réparties sur neuf Zones franches urbaines de la région parisienne.

Période 1955-2000 ($\boldsymbol{T = 46}$ années) chez Abadie et Gardeazabal (2003) qui étudient l'effet du conflit ETA au Pays-Basque sur le PIB réel par tête. Le groupe témoin comporte 16 régions autonomes d'Espagne.

Période 1970-2000 ($\boldsymbol{T = 31}$) dans l'étude d'Abadie, Diamond et Hainmueller (2010) sur la hausse des taxes sur la tabac en Californie (50+ États américains témoins potentiels, 38 retenus). Abadie *et alii* (2015) sur la réunification allemande ($\boldsymbol{T = 44}$).

Billmeier et Nannicini (2013) étudient l'effet de la libéralisation économique sur le PIB/tête pour la période 1963-2005 ($\boldsymbol{T = 43}$) dans un échantillon de pays, réduit à 127 après sélection. Le monde est découpé en régions, avec plusieurs pays considérés successivement comme individu test dans chacune. Ex. : en « Asie », l'Indonésie « libéralisée » en 1970 face à 8 témoins possibles (experiment A) ou une soixantaine (experiment B).

Il y a un arbitrage entre hypothèse de support commun plausible et puissance des tests

Campos *et alii* (2019) étudient l'effet de l'intégration européenne sur la croissance des pays membres pour la période 1970-2008 ($T = 39$). Les pays témoins sont piochés en dehors de l'UE. Par ex., pour l'Espagne, il y a 22 pays témoins candidats.

8.3.1. Protocole

Comme pour d'autres méthodes, ÉCMT dépend des RO de $I_0 \coloneqq \{i: D_i = 0\}$. Un algorithme *data-driven* sélectionne parmi I_0 et donne un poids w_i à chacun, l'idée étant qu'une moyenne pondérée d'individus témoins (un témoin composite, synthétique), une France synthétique par ex., offre une meilleure (*suitable*) comparaison qu'un seul témoin. Illustration perso d'une situation d'appariement intertemporel.

Transparence : on connait les poids relatifs de chaque individu témoin.

$\sum_{i:i\in I_0} w_i = 1 \rightarrow \sum_{i:i\in I_0'} w_i^* = 1$, avec $I_0' \subset I_0$, où $I_0' \coloneqq I \setminus \{i: w_i < \bar{c}\}$; $\bar{c} = 0$. On ne donne aucun poids aux individus écartés ($I_0 \backslash I_0'$), i.e. pas d'extrapolation ; Abadie (2021, 395), Abadie, Diamond et Hainmueller (2015, 498).

Protocole (notations-hypothèses) :

- Périodes : $1 \leq t \leq T_0$ (pré-traitement = *baseline period*, personne n'est traité), $T_0 + 1 \leq t \leq T$ (période de traitement, années d'anticipation incluses).
- $Y_{i,t}(0)$: RP d'une unité i non-exposée $\forall\, i, \forall\, t$.
- $Y_{i,t}(1)$: RP d'une unité exposée $\forall\, i, \forall\, t \geq T_0 + 1$.
- (•) Le traitement post-intervention n'a pas d'effet pré-intervention : $Y_{i,t}(1) = Y_{i,t}(0)$ $\forall\, t \leq T_0$. Empiriquement, il y a appariment pré-intervention.
- L'unité exposée est $i = 1$. $N_1 = 1, N_0 > 1$ (J dans Abadie et *alii*), $N_1 + N_0 \coloneqq N$.
- SUTVA

Tableau 8.1 : résultats potentiels et observés

	$i = 1$ (Californie)				$i > 1$			
	$d = 0$	$d = 1$	D	Y	$d = 0$	$d = 1$	D	Y
1	$Y_{1,1}(0)$	$Y_{1,1}(0)$	0	$Y_{1,1}(0)$	$Y_{i,1}(0)$	$Y_{i,1}(0)$	0	$Y_{i,1}(0)$
			.				.	
T_0	$Y_{1,T_0}(0)$	$Y_{1,T_0}(0)$	0	$Y_{1,T_0}(0)$	$Y_{i,T_0}(0)$	$Y_{i,T_0}(0)$	.	$Y_{i,T_0}(0)$
$T_0 + 1$	$Y_{1,T_0+1}(0)$	$Y_{1,T_0+1}(1)$	1	$Y_{1,T_0+1}(1)$	$Y_{i,T_0+1}(0)$	$Y_{i,T_0+1}(1)$	.	$Y_{i,T_0+1}(0)$
			.				.	
T	$Y_{1,T}(0)$	$Y_{1,T}(1)$	1	$Y_{1,T}(1)$	$Y_{i,T}(0)$	$Y_{i,T}(1)$	0	$Y_{i,T}(0)$

Note : observé, pas observable, ne nous intéresse pas, pas observable et ne nous intéresse pas

On peut définir un effet causal $\forall\, i, t : Y_{i,t}(1) - Y_{i,t}(0) \coloneqq \alpha_{i,t}$. On estime les $T - (T_0 + 1) + 1 = T - T_0$ quantités (variables dans le temps) $\alpha_{1,t} = Y_{1,t}(1) - Y_{1,t}(0)\ \forall\, t \geq T_0 + 1$.
PFIC : $Y_{1,t}(0)$ est inconnu $\forall\, t \geq T_0 + 1$, comme le montre l'équation de Rubin de chaque période, (••) $Y_{i,t} = Y_{i,t}(0) + D_{i,t}\alpha_{i,t}$, qui permet de remplir les colonnes 4 et 8 :

(i) $\boldsymbol{i = 1}$: $\forall\, \boldsymbol{t \leq T_0} \Rightarrow \boldsymbol{D_{1,t} = 0 \Rightarrow Y_{1,t} = Y_{1,t}(0) = Y_{1,t}(1)}$ **sous** □□ **et** □
$\forall\, \boldsymbol{t \geq T_0 + 1 \Rightarrow D_{1,t} = 1 \Rightarrow Y_{1,t} = Y_{1,t}(1)}$ **sous** □□

(ii) $\boldsymbol{i > 1}$: $\forall\, \boldsymbol{t \leq T_0 \Rightarrow D_{i,t} = 0 \Rightarrow Y_{i,t} = Y_{i,t}(0)}$ **sous** □□
$\forall\, \boldsymbol{t \geq T_0 + 1 \Rightarrow D_{i,t} = 0 \Rightarrow Y_{i,t} = Y_{i,t}(0)}$ **sous** □□

Prédire $\alpha_{1,t}\ \forall\, t \geq T_0 + 1$ par $\hat{\alpha}_{1,t} = Y_{1,t} - \hat{Y}_{1,t}(0)$. Reste à calculer $\hat{Y}_{1,t}(0)$.

$\hat{Y}_{1,t}(0) = \sum_{i=2}^{N_0+1} w_i Y_{i,t}(0) = \sum_{i=2}^{N_0+1} w_i Y_{i,t}$. Contrôle synthétique $W' = (w_2, \dots, w_{N_0+1})$.

8.3.2. Estimation

Estimer un modèle pour $Y_{i,t\leq T_0}(0)$ servant à prédire $Y_{1,t\geq T_0+1}(0)$ à partir de $i \geq 2$.

$Y_{i,t}(0) := \delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_i + u_{it}$, où u_{it} possède la structure à facteurs : les chocs non-observées sont approchés par des effets fixes temporels et l'interaction de ces effets avec les effets fixes individuels

$u_{it} := \boldsymbol{\lambda}_t \boldsymbol{\mu}_i + \epsilon_{it}$; $\boldsymbol{\lambda}_t \boldsymbol{\mu}_i$ est une modélisation non-structurelle de la corrélation entre coupes instantanées (*cross-correlation*) à chaque période (Bai, 2009, 1234).

δ_t est un facteur commun non observé avec poids constant $\forall\, i$: il capture un choc commun ayant un effet homogène sur $Y_{i,t}(0)$.

$\boldsymbol{\theta}_t \boldsymbol{Z}_i$ est la combinaison de régresseurs classiques observés. $\boldsymbol{\theta}_t' \sim r \boxtimes 1$, $\boldsymbol{Z}_i \sim r \boxtimes 1$, et $\boldsymbol{\theta}_t \boldsymbol{Z}_i = \sum_{s=1}^{r} \theta_{st} Z_{is}$, avec Z_{is} aléatoire !

$\boldsymbol{\lambda}_t$ est un vecteur de facteurs communs, $\boldsymbol{\mu}_i$ est un vecteur de poids (*factor loadings*), les deux étant non observés. $\boldsymbol{\lambda}_t' \sim F \boxtimes 1$, $\boldsymbol{\mu}_i \sim F \boxtimes 1$, et $\boldsymbol{\lambda}_t \boldsymbol{\mu}_i = \sum_{s=1}^{F} \lambda_{st} \mu_{is}$, avec μ_{is} aléatoire ! Les chocs communs $\boldsymbol{\lambda}_t$ ont un effet hétérogène sur $Y_{i,t}(0)$ (l'hétérogénéité μ_i a pour coefficient λ_t qui varie dans le temps). Par exemple, deux manières d'obtenir l'équivalent d'un modèle *two-way fixed effect*

$$F = 1, \boldsymbol{\mu}_i = \mathrm{M}_i, \boldsymbol{\lambda}_t = 1 \Rightarrow \boldsymbol{\lambda}_t \boldsymbol{\mu}_i = \mathrm{M}_i, \qquad Y_{i,t}(0) = \theta_t Z_i + \mathrm{M}_i + \delta_t + \epsilon_{it}$$

$$\delta_t = 0, F = 2, \boldsymbol{\lambda}_t = (1\ \Lambda_t),\ \boldsymbol{\mu}_i' = (\mathrm{M}_i\ 1) \Rightarrow \boldsymbol{\lambda}_t \boldsymbol{\mu}_i = \mathrm{M}_i + \Lambda_t, Y_{i,t}(0) = \theta_t Z_i + \mathrm{M}_i + \Lambda_t + \epsilon_{it},$$

ϵ_{it} est un bruit blanc gaussien par exemple.

La spécification à facteurs communs :

- généralise la double différence (sections 2 et 3 de MSE) ; voir Bai (2009) pour une présentation des modèles à facteurs, et Barhoumi *et alii* (2012) pour une revue de la littérature des versions dynamiques de ce type de modèles.
- entraîne une cross-sectional dépendance double : la dépendance de $\boldsymbol{\lambda}_t$ au temps produit une corrélation (Baltagi, 2005, 247) entre les $\boldsymbol{u}_{i,t}$, dont l'ampleur dépend de $\boldsymbol{\mu}_i$, et entre les $\boldsymbol{Y}_{i,t}$ Wooldridge (2010, 554-555), Bai (2009, 1238) montre comment éliminer $\boldsymbol{\mu}_i$.

On considère ensuite un vecteur de poids $\boldsymbol{W} = (w_2, \dots, w_{N_0+1})' \in R_+^{N_0}$ qui représente le contrôle synthétique, avec $\sum_j w_j = 1$. On se place entre 1 et T_0, la période d'appariement. Notons $\boldsymbol{Y}_j(0) = \boldsymbol{\delta} + \boldsymbol{\theta}\boldsymbol{Z}_j + \boldsymbol{\lambda}\boldsymbol{\mu}_j + \boldsymbol{\epsilon}_j$, avec $\boldsymbol{Y}_j(0), \boldsymbol{\delta}, \boldsymbol{\epsilon}_i \sim T_0 \boxtimes 1$, $\boldsymbol{\theta} \sim T_0 \boxtimes r$ et $\boldsymbol{\lambda} \sim T_0 \boxtimes F$, et remarquons que $\sum_j w_j \boldsymbol{\delta} = \boldsymbol{\delta} \sum_j w_j = \boldsymbol{\delta} 1 = \boldsymbol{\delta}$. Sur les $j \geq 2$, on a :

$$\begin{aligned} \textstyle\sum_j w_j \boldsymbol{Y}_j &= \textstyle\sum_j w_j \boldsymbol{Y}_j(0) \\ &= \boldsymbol{\delta} + \boldsymbol{\theta} \textstyle\sum_j w_j \boldsymbol{Z}_j + \boldsymbol{\lambda} \sum_j w_j \boldsymbol{\mu}_j + \sum_j w_j \boldsymbol{\epsilon}_j, \text{ et} \\ \boldsymbol{Y}_1 &= \boldsymbol{Y}_1(0) \\ &= \boldsymbol{\delta} + \boldsymbol{\theta}\boldsymbol{Z}_1 + \boldsymbol{\lambda}\boldsymbol{\mu}_1 + \boldsymbol{\epsilon}_1. \end{aligned}$$

On prend la différence, après avoir noté que $\boldsymbol{\epsilon}_1 = 1\boldsymbol{\epsilon}_1 = \sum_j w_j\, \boldsymbol{\epsilon}_1$,

$$\begin{aligned} \boldsymbol{Y}_1 - \textstyle\sum_j w_j \boldsymbol{Y}_j &= \boldsymbol{Y}_1(0) - \textstyle\sum_j w_j \boldsymbol{Y}_j(0) \\ &= \boldsymbol{\delta} + \boldsymbol{\theta}\boldsymbol{Z}_1 + \boldsymbol{\lambda}\boldsymbol{\mu}_1 + \boldsymbol{\epsilon}_1 - (\boldsymbol{\delta} + \boldsymbol{\theta} \textstyle\sum_j w_j \boldsymbol{Z}_j + \boldsymbol{\lambda} \sum_j w_j \boldsymbol{\mu}_j + \sum_j w_j \boldsymbol{\epsilon}_j) \\ \textbf{(A)} \qquad &= \boldsymbol{\theta}(\boldsymbol{Z}_1 - \textstyle\sum_j w_j \boldsymbol{Z}_j) + \boldsymbol{\lambda}(\boldsymbol{\mu}_1 - \textstyle\sum_j w_j \boldsymbol{\mu}_j) + \sum_j w_j (\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_j) \end{aligned}$$

Si on arrive à trouver $\boldsymbol{W}^*$: $\boldsymbol{Y}_1 - \sum_j w_j^* \boldsymbol{Y}_j$ soit proche de 0 entre 1 et T_0, on pourra utiliser $\sum_j w_j^* Y_{j,t}$ pour prédire $Y_{1,t}(0)$ en $t \geq T_0 + 1$. Problème : $\boldsymbol{\mu}_1 - \sum_j w_j \boldsymbol{\mu}_j$ n'est pas observable, $\boldsymbol{Z}_1 - \sum_j w_j \boldsymbol{Z}_j$ l'est pour une valeur de $\boldsymbol{W}^*$, mais $\boldsymbol{W}^*$ doit l'« annuler » aussi. Il

nous faut aussi des hypothèses concernant le dernier terme avec les erreurs $\boldsymbol{\epsilon}_1$, $\boldsymbol{\epsilon}_j$ (voir plus loin).
On suppose donc qu'il existe $\boldsymbol{W}^*$:

(B) $$\sum_j w_j^* \boldsymbol{Y}_j = \boldsymbol{Y}_1 \text{ et } \sum_j w_j^* \boldsymbol{Z}_j = \boldsymbol{Z}_1$$

(C) $\epsilon_{i,t}$~i.i.d. par définition, $\Rightarrow E\left(\epsilon_{i,t}\middle|\{Z_i,\mu_i\}_{i\geq 1}\right) = 0$. D'où $\mathrm{Cov}\left(u_{i,t}u_{j,t'}\middle|\{Z_i,\mu_i\}_{i\geq 1}\right) = 0$ si $i \neq j$ ou $t \neq t'$ ou les deux, mais $V\left(\epsilon_{i,t}\middle|\{Z_i,\mu_i\}_{i\geq 1}\right)$ si $i = j$ et $t = t'$. La démonstration dans le cas où $\boldsymbol{\lambda}_t\boldsymbol{\mu}_i$ est scalaire:

$$\begin{aligned} E\left(u_{i,t}u_{j,t'}\middle|\cdot\right) &= E\left((\boldsymbol{\lambda}_t\boldsymbol{\mu}_i + \epsilon_{i,t})(\boldsymbol{\lambda}_{t'}\boldsymbol{\mu}_j + \epsilon_{j,t'})\middle|\cdot\right) \\ &= E\left(\boldsymbol{\lambda}_t\boldsymbol{\mu}_i\boldsymbol{\lambda}_{t'}\boldsymbol{\mu}_j + \boldsymbol{\lambda}_t\boldsymbol{\mu}_i\epsilon_{j,t'} + \epsilon_{i,t}\boldsymbol{\lambda}_{t'}\boldsymbol{\mu}_j + \epsilon_{i,t}\epsilon_{j,t'}\middle|\cdot\right) \\ &= \boldsymbol{\lambda}_t\boldsymbol{\mu}_i\boldsymbol{\lambda}_{t'}\boldsymbol{\mu}_j + \boldsymbol{\lambda}_t\boldsymbol{\mu}_i E\left(\epsilon_{j,t'}\middle|\cdot\right) + \boldsymbol{\lambda}_{t'}\boldsymbol{\mu}_j E\left(\epsilon_{i,t}\middle|\cdot\right) + E\left(\epsilon_{i,t}\epsilon_{j,t'}\middle|\cdot\right) \\ &= \boldsymbol{\lambda}_t\boldsymbol{\mu}_i\boldsymbol{\lambda}_{t'}\boldsymbol{\mu}_j, \text{ et} \\ E\left(u_{i,t}\middle|\cdot\right)E\left(u_{j,t'}\middle|\cdot\right) &= \boldsymbol{\lambda}_t\boldsymbol{\mu}_i\boldsymbol{\lambda}_{t'}\boldsymbol{\mu}_j. \end{aligned}$$

■

Les calculs qui suivent font intervenir $\boldsymbol{\lambda}_t' = \begin{pmatrix} \lambda_{1t} \\ \vdots \\ \lambda_{Ft} \end{pmatrix}$ dans le produit $\boldsymbol{\lambda}_t'\boldsymbol{\lambda}_t = \begin{bmatrix} \lambda_{1t}^2 & \cdots & \lambda_{Ft}\lambda_{1t} \\ \vdots & \ddots & \vdots \\ \lambda_{Ft}\lambda_{1t} & \cdots & \lambda_{Ft}^2 \end{bmatrix}$. Et,

$$\sum \boldsymbol{\lambda}_t'\boldsymbol{\lambda}_t = \begin{bmatrix} \sum\lambda_{1t}^2 & \cdots & \sum\lambda_{Ft}\lambda_{1t} \\ \vdots & \ddots & \vdots \\ \sum\lambda_{Ft}\lambda_{1t} & \cdots & \sum\lambda_{Ft}^2 \end{bmatrix} = \boldsymbol{\lambda}'\boldsymbol{\lambda}, \text{ avec } \boldsymbol{\lambda}' \coloneqq \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1T_0} \\ \vdots & \ddots & \vdots \\ \lambda_{F1} & \cdots & \lambda_{FT_0} \end{bmatrix} \sim F \boxtimes T_0.$$

On définit ensuite pour tout $M : 1 \leq M \leq T_0$, $\xi(M) \coloneqq \min\left\{\xi \middle| \det\left(\frac{1}{M}\sum_{T_0-M+1}^{T_0} \boldsymbol{\lambda}_t'\boldsymbol{\lambda}_t\right) - \xi \boldsymbol{I}_F = 0\right\}$ (présentation de Ramis et *alii*, 2013, p. 35), la plus petite valeur propre.

(D) $\xi(M) \geq \underline{\xi} > 0$, $\Rightarrow \frac{1}{M}\sum_{T_0-M+1}^{t_0} \boldsymbol{\lambda}_t'\boldsymbol{\lambda}_t$ est définie positive ; théorème 8 de Magnus et Neudecker (1988, 14).

On a calculé $\xi(1)$ et $\xi(2)$ dans le cas $F = 2$, $\boldsymbol{\lambda}_t = (1\ \Lambda_t)$.

Notons que Λ_t étant dichotomique, alors $\Lambda_t\Lambda_{t'} = \Lambda_t^2 = \Lambda_t$ si $t = t'$ et 0 sinon (effets temporels mutuellement exclusifs).

(E) $\exists\, \bar{\lambda} : \left|\lambda_{ft}\right| \leq \bar{\lambda}\ \forall\, f, t.$

Les auteurs se débarrassent de $\boldsymbol{\mu}_1 - \sum_j w_j\boldsymbol{\mu}_j$ en passant par l'écriture matricielle suivante. On réduit le problème en projetant $\boldsymbol{Y}_1 - \sum_j w_j\boldsymbol{Y}_j - \boldsymbol{\theta}(\boldsymbol{Z}_1 - \sum_j w_j\boldsymbol{Z}_j)$ sur $\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{T_0}\}$ et considère $\boldsymbol{\mu}_1 - \sum_j w_j\boldsymbol{\mu}_j$ comme des paramètres à estimer, pour tout $\boldsymbol{W}$.

$$\underbrace{\boldsymbol{Y}_1 - \sum_j w_j\boldsymbol{Y}_j - \boldsymbol{\theta}\left(\boldsymbol{Z}_1 - \sum_j w_j\boldsymbol{Z}_j\right)}_{\boldsymbol{Y}\ T_0\boxtimes 1} = \underbrace{\boldsymbol{\lambda}}_{\boldsymbol{X}\ T_0\boxtimes F}\underbrace{\left(\boldsymbol{\mu}_1 - \sum_j w_j\boldsymbol{\mu}_j\right)}_{\boldsymbol{\beta}\ F\boxtimes 1} + \underbrace{\sum_j w_j(\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_j)}_{\boldsymbol{\varepsilon}\ T_0\boxtimes 1}.$$

si $T_0 > F$, c'est un problème d'approximation par les MC (Magnus et Neudecker, 1988, 259) du type « $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ », pour lequel « $\boldsymbol{X}$ » doit avoir au moins autant de lignes que de colonnes, $T_0 \geq F$, sinon « $\boldsymbol{X}'\boldsymbol{X}$ », i.e. $\boldsymbol{\lambda}'\boldsymbol{\lambda}$, n'est pas inversible inverse généralisée dans ce cas ? Notons $(\boldsymbol{\lambda}'\boldsymbol{\lambda})^{-1}\boldsymbol{\lambda}' \coloneqq \boldsymbol{Q}_{\boldsymbol{\lambda}}$. Remarque : $\boldsymbol{\theta}$ doit être estimé avant, sinon « $\boldsymbol{Y}$ » n'est pas observable. Donc :

(F) $$\left(\boldsymbol{\mu}_1 - \sum_j w_j\boldsymbol{\mu}_j\right)^{MC} = \boldsymbol{Q}_{\boldsymbol{\lambda}}\left(\boldsymbol{Y}_1 - \sum_j w_j\boldsymbol{Y}_j\right) - \boldsymbol{Q}_{\boldsymbol{\lambda}}\boldsymbol{\theta}\left(\boldsymbol{Z}_1 - \sum_j w_j\boldsymbol{Z}_j\right) - \boldsymbol{Q}_{\boldsymbol{\lambda}}\sum_j w_j(\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_j).$$

Alors, sous (**A**), (**B**), (**F**), et après avoir noté $\boldsymbol{\lambda Q_\lambda} \coloneqq \boldsymbol{P_\lambda}$

$$\boldsymbol{Y}_1 - \sum_j w_j^* \boldsymbol{Y}_j = \overbrace{\boldsymbol{\theta}\left(\boldsymbol{Z}_1 - \sum_j w_j^* \boldsymbol{Z}_j\right)}^{0} + \overbrace{\boldsymbol{P_\lambda}\left(\boldsymbol{Y}_1 - \sum_j w_j^* \boldsymbol{Y}_j\right)}^{0} - \overbrace{\boldsymbol{P_\lambda}\boldsymbol{\theta}\left(\boldsymbol{Z}_1 - \sum_j w_j^* \boldsymbol{Z}_j\right)}^{0} - \boldsymbol{P_\lambda}\sum_j w_j^*(\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_j) + \sum_j w_j^*(\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_j).$$

Et la prédiction de $\boldsymbol{\mu}_1 - \sum_j w_j^* \boldsymbol{\mu}_j$ est égale à $-\boldsymbol{Q_\lambda}\sum_j w_j^*(\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_j)$, qui est de moyenne nulle, comme nous allons le voir. En fait, les auteurs s'intéressent à la prédiction ci-dessus, de $\boldsymbol{Y}_1 - \sum_j w_j^* \boldsymbol{Y}_j$, pour chaque t entre $T_0 + 1$ et T:

$$\begin{aligned} Y_{1,t} - \sum_j w_j^* Y_{j,t} &= -\boldsymbol{\lambda}_t \boldsymbol{Q_\lambda}\sum_j w_j^*(\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_j) + \sum_j w_j^*(\epsilon_{1,t} - \epsilon_{j,t}), \\ &= \boldsymbol{\lambda}_t \boldsymbol{Q_\lambda}\sum_j w_j^* \boldsymbol{\epsilon}_j - \boldsymbol{\lambda}_t \boldsymbol{Q_\lambda}\boldsymbol{\epsilon}_1 + \sum_j w_j^*(\epsilon_{1,t} - \epsilon_{j,t}) \coloneqq R_{1,t} - R_{2,t} + R_{3,t}. \end{aligned}$$

De 1 à T_0, on peut montrer que cette quantité est proche de zéro sous les hypothèses précédentes. Ensuite, de $T_0 + 1$ à T, $\hat{\alpha}_{1,t} \coloneqq Y_{1,t} - \sum_j w_j^* Y_{j,t}$ et l'effet causal, et $R_{1,t} - R_{2,t} + R_{3,t}$ le biais.

Maintenant, les auteurs se placent dans $t \geq T_0 + 1$ et calculent sans dire qu'ils prennent l'espérance $E\left(R_{g,t}\middle|\{Z_i, \mu_i\}_{i\geq 1}\right)$, $g = 1,2,3$. $R_{2,t}$ et $R_{3,t}$ sont d'espérance nulle :

- $R_{2,t}$ ne dépend que de paramètres multipliés par $\boldsymbol{\epsilon}_1$ dont la moyenne est nulle ;
- $R_{3,t}$ dépend de w_j^* dont la valeur dépend de l'écart entre $\boldsymbol{Y}_1$ et $\sum_j w_j^* \boldsymbol{Y}_j$ dans (**A**), qui dépend lui-même de $\sum_j w_j^*(\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_j)$ où les erreurs $\epsilon_{i,t}$ vont de 1 à T_0 $\forall$ i. Alors que dans $R_{3,t}$, les erreurs vont de $T_0 + 1$ à T ;

$R_{1,t}$ est plus difficile car, comme pour $R_{3,t}$, les w_j^* dépendent des erreurs $\epsilon_{i,t}$ allant de 1 à T_0, mais les $\epsilon_{j,t}$ dans $\boldsymbol{\epsilon}_j$ vont également de 1 à T_0 (Botosaru et Ferman, 2019). On veut borner $E\left(R_{1,t}\middle|\cdot\right)$.

Après avoir réécrit $R_{1,t}$,

$$\boldsymbol{\lambda}_t \boldsymbol{Q_\lambda}\sum_j w_j^* \boldsymbol{\epsilon}_j = \sum_j w_j^* \boldsymbol{\lambda}_t \boldsymbol{Q_\lambda}\boldsymbol{\epsilon}_j = \sum_j w_j^* \boldsymbol{\lambda}_t(\boldsymbol{\lambda}'\boldsymbol{\lambda})^{-1}\boldsymbol{\lambda}'\boldsymbol{\epsilon}_j = \sum_j w_j^* \boldsymbol{\lambda}_t(\boldsymbol{\lambda}'\boldsymbol{\lambda})^{-1}\sum_s \boldsymbol{\lambda}_s'\epsilon_{js} =$$
$$\sum_j w_j^* \sum_s \boldsymbol{\lambda}_t\left(\sum_n \boldsymbol{\lambda}_n'\boldsymbol{\lambda}_n\right)^{-1}\boldsymbol{\lambda}_s'\epsilon_{js},$$

les auteurs bornent d'abord $\boldsymbol{\lambda}_t(\sum_n \boldsymbol{\lambda}_n'\boldsymbol{\lambda}_n)^{-1}\boldsymbol{\lambda}_s' \coloneqq m_{ts}$; ils montrent que :

$$m_{ts}^2 \leq \left(\frac{F}{T_0 \underline{\xi}}\bar{\lambda}^2\right)^2.$$

Puis, il s'agit d'utiliser différentes inégalités mathématiques afin de montrer que $E\left(|R_{1,t}|\middle|\cdot\right)$ est borné par une quantité qui a pour limite 0 quand $T_0 \to \infty$. Autrement dit, la période pré-traitement doit être assez grande.

Donc, $Y_{1,t} - \sum_j w_j^* Y_{j,t}$ est borné à droite. Cette différence fournit une estimation de l'effet causal sur la période $T_0 + 1, \dots, T$. Le vecteur $\boldsymbol{W}^*$, déterminé pour la période $1, \dots, T_0$, est appliqué à la période $T_0 + 1, \dots, T$.

8.3.3. Estimation de $\boldsymbol{W}^*$

Pour la période de pré-traitement $t \leq T_0$, on empile

- les variables explicatives $\boldsymbol{Z_1}$ et de pré-traitement $\boldsymbol{Y_1}$ dans $\boldsymbol{X_1}$.
- les variables explicatives $\boldsymbol{Z_0}$ et de pré-traitement $\boldsymbol{Y_0}$ des $\boldsymbol{N_0}$ témoins dans $\boldsymbol{X_0}$.

On cherche $\boldsymbol{W}$ qui minimise la distance suivante :

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}\|_{\boldsymbol{V}}$$

$(r + T_0) \boxtimes 1 \quad (r + T_0) \boxtimes N_0 \quad N_0 \boxtimes 1$

sous la contrainte que $\boldsymbol{W} \in \mathbb{R}_+$ et $\boldsymbol{e}'\boldsymbol{W} = 1$, où $\boldsymbol{e}$ est le vecteur unité de taille N_0. La matrice $\boldsymbol{V} \sim (r + T_0) \boxtimes (r + T_0)$ pondère comme pour une distance de Mahalanobis.

Le critère à minimiser est le Root Mean Square Prediction Error entre $\boldsymbol{Y}_1$ et $\boldsymbol{Y}_0\boldsymbol{W}$.

8.3.4. Application

Il existe une commande dans Stata, **`synth`**, mais aussi pour R et MATLAB. Nous allons l'appliquer pour répliquer intégralement l'article d'Abadie, Diamond et Hainmueller (2010). Le temps de calcul du test placébo est long !

Le traitement est la « Proposition 99 » en Californie (tobacco control program à grande échelle), implémentée en 1988. La variable de résultat est la consommation de tabac.

[« **sc_smoking.do** »]

8.4. Exercices

8.4.1) Supposez que l'indépendance conditionnelle entre $Y(d)$ et D : $E(Y(d)|D, X, D_3) = E(Y(d)|X, D_3)$. Montrer que la quantité (8.1) mesure l'ECMT.

8.4.2) On considère le modèle deux groupes-deux périodes suivant :

$$Y_{i,t} = \beta_1^Y + \beta_2^Y D_i + \beta_3^Y D_3 + \beta_4^Y D_i D_3 + \varepsilon_{it}^Y$$

où $D_i = 1$ si i est traité, 0 sinon, et $D_t = 1$ si $t = 3$ et 0 sinon, et l'erreur ε_{it} n'est pas corrélée aux variables explicatives du modèle, D_i, D_t et $D_i D_t$. On estime le modèle par les MCO. On note les estimateurs $\hat{\beta}_1^Y$, $\hat{\beta}_2^Y$, $\hat{\beta}_3^Y$ et $\hat{\beta}_4^Y$. Montrer le résultat utilisé dans la **section 8.1.2** :

$$\begin{aligned}
\hat{\beta}_1^Y &= \bar{Y}_{02}, \\
\hat{\beta}_2^Y &= \bar{Y}_{12} - \bar{Y}_{02}, \\
\hat{\beta}_3^Y &= \bar{Y}_{03} - \bar{Y}_{02}, \\
\hat{\beta}_4^Y &= \bar{Y}_{13} - \bar{Y}_{03} - (\bar{Y}_{12} - \bar{Y}_{02}).
\end{aligned}$$

(Conseil : résoudre le système des équations normales)

8.4.3) (i) On considère le même modèle que dans l'**exercice 8.4.2**, mais avec la variable de confusion supplémentaire X_{it} :

$$Y_{i,t} = \beta_1^{YX} + \beta_2^{YX} D_i + \beta_3^{YX} D_3 + \beta_4^{YX} D_i D_3 + \beta_5^{YX} X_{it} + \varepsilon_{it}^{YX}$$

où D_i et D_3 sont définis comme précédemment et l'erreur ε_{it}^{YX} est définie comme ε_{it}^Y. Montrer que

$$\begin{aligned}
\hat{\beta}_1^{YX} &= \hat{\beta}_1^Y - \bar{X}_{02}\hat{\beta}_5^{YX}, \\
\hat{\beta}_2^{YX} &= \hat{\beta}_2^Y - (\bar{X}_{12} - \bar{X}_{02})\hat{\beta}_5^{YX}, \\
\hat{\beta}_3^{YX} &= \hat{\beta}_3^Y - (\bar{X}_{03} - \bar{X}_{02})\hat{\beta}_5^{YX}, \\
\hat{\beta}_4^{YX} &= \hat{\beta}_4^Y - [\bar{X}_{13} - \bar{X}_{03} - (\bar{X}_{12} - \bar{X}_{02})]\hat{\beta}_5^{YX}, \text{ et}
\end{aligned}$$

$$\hat{\beta}_5^{YX} = \frac{\sum[\sum_{i:D_i=0}(Y_{it}-\bar{Y}_{0t})(X_{it}-\bar{X}_{0t})+\sum_{i:D_i=1}(Y_{it}-\bar{Y}_{1t})(X_{it}-\bar{X}_{1t})]}{\sum[\sum_{i:D_i=0}(X_{it}-\bar{X}_{0t})^2+\sum_{i:D_i=1}(X_{it}-\bar{X}_{1t})^2]},$$

avec $\hat{\beta}_1^Y$, $\hat{\beta}_2^Y$, $\hat{\beta}_3^Y$ et $\hat{\beta}_4^Y$ les solutions de l'**exercice 8.4.2**.

(ii) Que devient $\hat{\beta}_4^{YX}$ si X varie identiquement dans le temps dans les deux groupes de traitements ?

Corrections des exercices du chapitre 8

8.4.1) Sous CIA, on passe de (8.1) à la différence de différences (8.2) dans laquelle on peut omettre D dans chaque terme :

$$\begin{aligned} &E(Y(1)|D_3=1,X)-E(Y(0)|D_3=0,X)-[E(Y(0)|D_3=1,X)-E(Y(0)|D_3=0,X)] \\ &=E(Y(1)|D_3=1,X)-E(Y(0)|D_3=1,X) \\ &=E(Y(1)-Y(0)|D_3=1,X). \end{aligned}$$

■

8.4.2) On peut commencer par écrire la matrice $\boldsymbol{X}$ des variables explicatives du modèle pour les $2N$ observations :

$$\boldsymbol{X} \coloneqq \begin{matrix} \left.\begin{matrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{matrix}\right) & 2N_0 \\ \left.\begin{matrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{matrix}\right) & 2N_1 \end{matrix}$$

Le système des équations normales $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}^Y = \boldsymbol{X}'\boldsymbol{Y}$ prend la forme explicite suivante :

$$\begin{bmatrix} 2N & 2N_1 & N & N_1 \\ 2N_1 & 2N_1 & N_1 & N_1 \\ N & N_1 & N & N_1 \\ N_1 & N_1 & N_1 & N_1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1^Y \\ \hat{\beta}_2^Y \\ \hat{\beta}_3^Y \\ \hat{\beta}_4^Y \end{bmatrix} = \begin{bmatrix} 2N\bar{Y}_{..} \\ 2N_1\bar{Y}_{1.} \\ N\bar{Y}_{.3} \\ N_1\bar{Y}_{13} \end{bmatrix}$$

Effectuons le produit matriciel :

$$\begin{aligned} 2N\hat{\beta}_1^Y + 2N_1\hat{\beta}_2^Y + N\hat{\beta}_3^Y + N_1\hat{\beta}_4^Y &= 2N\bar{Y}_{..} && (1) \\ 2N_1\hat{\beta}_1^Y + 2N_1\hat{\beta}_2^Y + N_1\hat{\beta}_3^Y + N_1\hat{\beta}_4^Y &= 2N_1\bar{Y}_{1.} && (2) \\ N\hat{\beta}_1^Y + N_1\hat{\beta}_2^Y + N\hat{\beta}_3^Y + N_1\hat{\beta}_4^Y &= N\bar{Y}_{.3} && (3) \\ N_1\hat{\beta}_1^Y + N_1\hat{\beta}_2^Y + N_1\hat{\beta}_3^Y + N_1\hat{\beta}_4^Y &= N_1\bar{Y}_{13} && (4) \end{aligned}$$

Si on prend (2) moins (4) et (1) moins (3) on obtient après simplification les équations (2)' et (1)' suivantes :

$$\begin{aligned} \hat{\beta}_1^Y + \hat{\beta}_2^Y &= 2\bar{Y}_{1.} - \bar{Y}_{13} && (2') \\ N\hat{\beta}_1^Y + N_1\hat{\beta}_2^Y &= 2N\bar{Y}_{..} - N\bar{Y}_{.3} && (1') \end{aligned}$$

Puis on calcule $N \times (2') - (1')$, ce qui permet d'isoler $\hat{\beta}_2^Y = \frac{N}{N_0}(2\bar{Y}_{1.} - \bar{Y}_{13} - 2\bar{Y}_{..} + \bar{Y}_{.3})$.

Or, $2\bar{Y}_{1.} - \bar{Y}_{13} = 2\frac{\sum\sum_{i:D_i=1} Y_{it}}{2N_1} - \frac{\sum_{i:D_i=1} Y_{i3}}{N_1} = \frac{\sum_{i:D_i=1} Y_{i2}}{N_1}$, et $-2\bar{Y}_{..} + \bar{Y}_{.3} = -2\frac{\sum\sum Y_{it}}{2N} + \frac{\sum Y_{i3}}{N} = -\frac{\sum Y_{i2}}{N} = -\frac{\sum_{i:D_i=1} Y_{i2}}{N} - \frac{\sum_{i:D_i=0} Y_{i2}}{N}$. D'où,

$$2\bar{Y}_{1.} - \bar{Y}_{13} - 2\bar{Y}_{..} + \bar{Y}_{.3} = \frac{\sum_{i:D_i=1} Y_{i2}}{N_1} - \frac{\sum_{i:D_i=1} Y_{i2}}{N} - \frac{\sum_{i:D_i=0} Y_{i2}}{N} = \frac{N_0}{N}\frac{\sum_{i:D_i=1} Y_{i2}}{N_1} - \frac{\sum_{i:D_i=0} Y_{i2}}{N}.$$

Par conséquent, $\boxed{\hat{\beta}_2^Y = \frac{N}{N_0}\left(\frac{N_0}{N}\frac{\sum_{i:D_i=1} Y_{i2}}{N_1} - \frac{\sum_{i:D_i=0} Y_{i2}}{N}\right) = \frac{\sum_{i:D_i=1} Y_{i2}}{N_1} - \frac{\sum_{i:D_i=0} Y_{i2}}{N_0} = \bar{Y}_{12} - \bar{Y}_{02}}$, que l'on peut mettre dans (2'), ce qui permet de déduire $\hat{\beta}_1^Y = 2\bar{Y}_{1.} - \bar{Y}_{13} - \bar{Y}_{12} + \bar{Y}_{02}$. Or, $2\bar{Y}_{1.} = 2\frac{\sum\sum_{i:D_i=1} Y_{it}}{2N_1} = \frac{\sum_{i:D_i=1} Y_{i2}}{N_1} + \frac{\sum_{i:D_i=1} Y_{i3}}{N_1} = \bar{Y}_{12} + \bar{Y}_{13}$. Donc, $\boxed{\hat{\beta}_1^Y = \bar{Y}_{02}}$.

On peut trouver $\hat{\beta}_3^Y$ en calculant (1) moins (2), puis en substituant $\bar{Y}_{02}$ à $\hat{\beta}_1^Y$, ce qui donne l'égalité $2N_0\bar{Y}_{02} + N_0\hat{\beta}_3^Y = 2N\bar{Y}_{..} - 2N_1\bar{Y}_{1.}$ Or, $2N\bar{Y}_{..} = 2N_0\bar{Y}_{0.} + 2N_1\bar{Y}_{1.}$, d'où $2N_0\bar{Y}_{02} + N_0\hat{\beta}_3^Y = 2N_0\bar{Y}_{0.}$ On peut multiplier à gauche et à droite par $1/N_0$ et isoler $\hat{\beta}_3^Y$. Ainsi, $\hat{\beta}_3^Y = 2\bar{Y}_{0.} - 2\bar{Y}_{02}$. Or, $2\bar{Y}_{0.} = \bar{Y}_{02} + \bar{Y}_{03}$. Par conséquent, $\boxed{\hat{\beta}_3^Y = \bar{Y}_{03} - \bar{Y}_{02}}$.

Enfin, on met $\hat{\beta}_1^Y$, $\hat{\beta}_2^Y$ et $\hat{\beta}_3^Y$ dans l'équation (4) multipliée par $1/N_1$, ce qui donne :

$$\bar{Y}_{02} + (\bar{Y}_{12} - \bar{Y}_{02}) + (\bar{Y}_{03} - \bar{Y}_{02}) + \hat{\beta}_4^Y = \bar{Y}_{13},$$

d'où $\boxed{\hat{\beta}_4^Y = \bar{Y}_{13} - \bar{Y}_{12} - (\bar{Y}_{03} - \bar{Y}_{02})}$.

C'est l'estimateur avant-après du groupe test moins celui du groupe témoin. En réorganisant les terme, on peut noter qu'il s'agit de l'ECMT après moins l'ECMT avant.

8.4.3)
(i) Comme dans l'exercice précédent, on a la matrice $\boldsymbol{X}$, cette fois-ci augmentée des observations de la variable explicative X :

$$\left(X_{1,2}, X_{1,3}, \dots, X_{2N_0,2}, X_{2N_0,3}, X_{2N_0+1,2}, X_{2N_0+1,3}, \dots, X_{2N,2}, X_{2N,3}\right)' \coloneqq \boldsymbol{Z}'.$$

Le système des équations normales comporte une colonne et une ligne de plus :

$$\begin{aligned}
2N\hat{\beta}_1^{YX} + 2N_1\hat{\beta}_2^{YX} + N\hat{\beta}_3^{YX} + N_1\hat{\beta}_4^{YX} + 2N\bar{X}_{..}\hat{\beta}_5^{YX} &= 2N\bar{Y}_{..} && (1)\\
2N_1\hat{\beta}_1^{YX} + 2N_1\hat{\beta}_2^{YX} + N_1\hat{\beta}_3^{YX} + N_1\hat{\beta}_4^{YX} + 2N_1\bar{X}_{1.}\hat{\beta}_5^{YX} &= 2N_1\bar{Y}_{1.} && (2)\\
N\hat{\beta}_1^{YX} + N_1\hat{\beta}_2^{YX} + N\hat{\beta}_3^{YX} + N_1\hat{\beta}_4^{YX} + N\bar{X}_{.3}\hat{\beta}_5^{YX} &= N\bar{Y}_{.3} && (3)\\
N_1\hat{\beta}_1^{YX} + N_1\hat{\beta}_2^{YX} + N_1\hat{\beta}_3^{YX} + N_1\hat{\beta}_4^{YX} + N_1\bar{X}_{13}\hat{\beta}_5^{YX} &= N_1\bar{Y}_{13} && (4)\\
2N\bar{X}_{..}\hat{\beta}_1^{YX} + 2N_1\bar{X}_{1.}\hat{\beta}_2^{YX} + N\bar{X}_{.3}\hat{\beta}_3^{YX} + N_1\bar{X}_{13}\hat{\beta}_4^{YX} + \sum\sum X_{it}^2\hat{\beta}_5^{YX} &= \sum\sum Y_{it}^2 && (5)
\end{aligned}$$

Ce système de 5 équations-5 inconnues inclut le système précédent, $\boldsymbol{X'X\beta}^Y = \boldsymbol{X'Y}$.

> Nous pourrions être tentés d'inverser $\boldsymbol{X'X}$ et calculer :
>
> $$\begin{bmatrix}\boldsymbol{\beta}_4^{YX}\\ \beta_5^{YX}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X'X} & \boldsymbol{X'Z}\\ \boldsymbol{Z'X} & \boldsymbol{Z'Z}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{X'Y}\\ \boldsymbol{Z'Y}\end{bmatrix}$$
>
> La matrice inverse est plus compliquée qu'il n'y paraît ; voir Magnus et Neudecker (1988). En effet, $\begin{bmatrix}\boldsymbol{X'X} & \boldsymbol{X'Z}\\ \boldsymbol{Z'X} & \boldsymbol{Z'Z}\end{bmatrix}^{-1} \coloneqq \begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix}$, avec $\boldsymbol{A}$ par exemple qui est égal à $(\boldsymbol{X'X})^{-1} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z}[\boldsymbol{Z'Z} - \boldsymbol{Z'X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'Z}]\boldsymbol{Z'X}(\boldsymbol{X'X})^{-1}$. Nous n'empruntons pas cette voie.

Comme dans l'**exercice 8.4.2**, faisons (2) moins (4) et (1) moins (3). Après simplification, nous obtenons $\hat{\beta}_1^{YX} = \bar{Y}_{02} - \bar{X}_{02}\hat{\beta}_5^{YX}$ et $\hat{\beta}_2^{YX} = \bar{Y}_{12} - \bar{Y}_{02} - (\bar{X}_{12} - \bar{X}_{02})\hat{\beta}_5^{YX}$. On reconnaît $\hat{\beta}_1^Y$ et $\hat{\beta}_2^Y$ à droite du signe '='. Les solutions pour $\hat{\beta}_3^{YX}$ et $\hat{\beta}_4^{YX}$ dépendent de $\hat{\beta}_5^{YX}$, en ont la même structure avec $\hat{\beta}_3^Y$ et $\hat{\beta}_4^Y$ à droite du signe '= : $\hat{\beta}_3^{YX} = \bar{Y}_{03} - \bar{Y}_{02} - (\bar{X}_{03} - \bar{X}_{02})\hat{\beta}_5^{YX}$, $\hat{\beta}_4^{YX} = \bar{Y}_{13} - \bar{Y}_{03} - (\bar{Y}_{12} - \bar{Y}_{02}) - [\bar{X}_{13} - \bar{X}_{03} - (\bar{X}_{12} - \bar{X}_{02})]\hat{\beta}_5^{YX}$. Enfin, $\hat{\beta}_5^{YX}$ s'obtient en remplaçant $\hat{\beta}_1^{YX}$, $\hat{\beta}_2^{YX}$, $\hat{\beta}_3^{YX}$ et $\hat{\beta}_4^{YX}$ par leurs valeurs dans (5). Après simplification :

$$\hat{\beta}_5^{YX} = \frac{\sum[\sum_{i:D_i=0}(Y_{it} - \bar{Y}_{0t})(X_{it} - \bar{X}_{0t}) + \sum_{i:D_i=1}(Y_{it} - \bar{Y}_{1t})(X_{it} - \bar{X}_{1t})]}{\sum[\sum_{i:D_i=0}(X_{it} - \bar{X}_{0t})^2 + \sum_{i:D_i=1}(X_{it} - \bar{X}_{1t})^2]}.$$

■

(ii) Si X varie identiquement, alors $\bar{X}_{13} - \bar{X}_{12} = \bar{X}_{03} - \bar{X}_{02}$, et par conséquent :

$$\begin{aligned}
\hat{\beta}_4^{YX} &= \hat{\beta}_4^Y - [\bar{X}_{13} - \bar{X}_{03} - (\bar{X}_{12} - \bar{X}_{02})]\hat{\beta}_5^{YX}\\
&= \hat{\beta}_4^Y - [\bar{X}_{13} - \bar{X}_{12} - (\bar{X}_{03} - \bar{X}_{02})]\hat{\beta}_5^{YX}
\end{aligned}$$

$= \hat{\beta}_4^Y - 0\hat{\beta}_5^{YX}$
$= \hat{\beta}_4^Y .$

9. Sélection sur facteurs non-observables et variables instrumentales

Dans ce chapitre, la sélection des individus dans les groupes de traitement dépend aussi de variables omises (VO) ; Rosenbaum (2010) parle de *hidden bias* (par opposition au BS manifeste/*overt bias* vu jusqu'à présent). On parle aussi de sélection sur (facteurs) non-observables (SsnO).

Lee (2016, 15) donne la définition suivante du problème que SsnO entraîne :

$$E(Y(d)|D,X) \neq E(Y(d)|X).$$

En revanche,

$$E(Y(d)|D,X,\epsilon) = E(Y(d)|X,\epsilon),$$

où ϵ est la VO (de confusion) non observée.

Les VO non-observées peuvent entraîner une hétérogénéité de l'effet causal entre individus. Pour le montrer, il suffit de déplacer les termes d'erreur dans le modèle de régression avec résultats potentiels (PO regression model). En effet, on avait montré (cf. supra **sous-section 7.1.3**) :

$$Y = E\big(Y(0)\big) + D\left(E\big(Y(1)\big) - E\big(Y(0)\big)\right) + D\epsilon_1 + (1-D)\epsilon_0.$$

Il y a visiblement une corrélation entre le terme d'erreur composite $D\epsilon_1 + (1-D)\epsilon_0$ et la variable « causale », D. Nous avions vu dans la **chapitre 7** que la condition d'ignorabilité suffisait : $E(\epsilon_d|D) = 0 \Leftrightarrow E(Y(d)|D) = E\big(Y(d)\big)$.

C'est juste une question de présentation de l'équation ci-dessus : on peut faire apparaître un effet causal hétérogène :

$$Y = E\big(Y(0)\big) + D\big(E\big(Y(1)\big) - E\big(Y(0)\big) + \epsilon_1 - \epsilon_0\big) + \epsilon_0.$$

Pour Heckman (1996), il s'agit d'un modèle à coefficient aléatoire ; il propose des solutions d'estimation. La seule variable explicative endogène étant dichotomique, on parle aussi de *dummy endogenous variable model* (Wooldridge, 2003).

Si les VO varient individuellement mais sont fixes dans le temps (ex., les ϵ_d incluent un effet fixe plus un bruit blanc, le problème peut être atténué avec un protocole de type différence de différences (**chapitre 8**). Ex. : la VO est le plus haut diplôme (fixe dans le temps), qui n'est pas observé par l'évaluateur.

Nous n'allons pas prendre cette voie dans ce cours. Heckman (1990) pousse la discussion assez loin, que reprend Wooldridge (2003). La résolution repose sur des hypothèses fines, difficiles à tenir en pratique. C'est sans doute la raison pour laquelle, on ne voit peu d'utilisation de cette approche, à part en économie du travail.

Nous allons voir deux ensembles de situations du problème de SsnO :

(i) Le biais d'estimation est réduit grâce à des variables instrumentales (VI), et différents estimateurs sont disponibles (VI, DMC (2SLS), 3SLS, MM, …).

Qu'il s'agisse des problèmes classiques (VO, MES, EM), ou du problème éligibilité-participation (LATE), les instruments visent à neutraliser une source de variation de ***Y*** qui transite par ***D*** afin de ne pas brouiller l'effet causal de ***D***.

(ii) Le problème de SsnO peut être est plus grave

Les résultats des individus du groupe de contrôle ne sont pas observables. Le cas emblématique est celui où le traitement est une

formation, dans le cadre d'une évaluation de la formation professionnelle sur les salaires. On a un problème d'observations omises (OO), le modèle est tronqué ; l'évaluateur n'a les salaires que des individus ayant trouvé un travail ; c'est le fameux modèle heckit)

On rappelle l'approche classique des VI dans la **section 9.1**, avec les estimateurs DMC et VI. Ce sera utile avant d'introduire l'estimateur LATE que nous verrons dans la **section 9.2**, développé au milieu des années 1990, pour tenir compte de situations par exemple où des individus éligibles ne participent pas. Cette situation fait intervenir une nouvelle classe d'effet causal, appelée *local average treatment effect* (LATE). Imbens et Angrist (1994) ont montré que l'estimateur à VI peut mesurer un effet causal au sens de Rubin, mais pour une sous-population particulière (les compliers/conformistes). Dans la **section 9.3** vous trouverez le problème de SsnO à la Heckman (1979), lui aussi classique, bien que moins utilisé aujourd'hui.

9.1. Estimateur à VI : approche classique

Pour Wooldridge (2003, 603), il y a deux catégories d'estimateurs de l'ECM :
Les estimateurs qui reposent sur CIA.[38]
Les estimateurs à VI, qui sont ceux que nous allons voir dans ce **chapitre**.

Les différentes situations classiques où les variables instrumentales (VI) sont utiles :

1) **Les VI règlent le problème de BVO, déjà discuté dans le chapitre 8.**
2) **La méthode des VI fut d'abord inventée pour estimer de manière consistante les paramètres d'un modèle à équation simultanée (MES). Les raisonnements que l'on fait pour estimer ces modèles peuvent néanmoins être utiles, pour chercher justement de bons instruments ; nous allons donner une illustration.**
3) **Enfin, la méthode des VI fut parallèlement inventée pour régler le problème d'erreur de mesure lorsque, par exemple, une variable explicative est mesurée avec erreur.**

Nous commençons par rappeler ces trois situations en théorie (**sous-section 9.1.1**). Puis nous illustrons leur résolution de manière standard à partir de la relation salaire-éducation (**sous-section 9.1.2**).

9.1.1. Trois situations théoriques

1) **BVO**. Naïvement, nous estimerions le modèle

$$Y = \beta_{11} + D\beta_{12} + \epsilon_1, \qquad (M_1)$$

sachant que X n'est pas observé. Comment fait-on pour estimer β_{12} sans biais ? Avant de répondre à cette question, nous pouvons mesurer l'ampleur du **BVO**. Pour cela, écrivons le modèle qu'il serait préférable d'estimer :

$$Y = \beta_{21} + D\beta_{22} + \beta_{23}X + \epsilon_2. \qquad (M_2)$$

Nous voyons tout de suite, par identification terme à terme de M_1 et M_2, que $\epsilon_1 = \beta_{23}X + \epsilon_2$. Nous allons supposer que $Cov(\epsilon_2, D) = 0$; dans le modèle « vrai » M_2, il n'y a pas de VO corrélée avec D. Les MC pour β_{12} sont néanmoins biaisés et inconsistants :

[38] Approche analogue à celle des proxy du problème de VO ; voir chapitre 4 de Wooldridge (2003).

$$\begin{aligned}\beta_{12}^{MC} &= \frac{Cov(Y,D)}{V(D)} \\ &= \frac{Cov(\beta_{12}+D\beta_{22}+\beta_{23}X+\epsilon_2,D)}{V(D)} \\ &= \beta_{22} + \beta_{23}\frac{Cov(X,D)}{V(D)} + \frac{Cov(\epsilon_2,D)}{V(D)} \\ &= \beta_{22} + \beta_{23}\frac{Cov(X,D)}{V(D)}.\end{aligned}$$

La source du biais est le terme $Cov(X, D)$, une corrélation potentielle entre la VO X et le traitement D (c'est le BVO, un type de biais d'endogénéité). En tenant compte de $V(D)$, on voit que le biais est le coefficient de pente dans la régression de X sur D :

$$X = \beta_{31} + D\beta_{32} + \epsilon_3, \qquad (M_3)$$

Nous pouvons donc écrire :

$$\beta_{12}^{MC} = \beta_{22} + \beta_{23}\beta_{32}^{MC}.$$

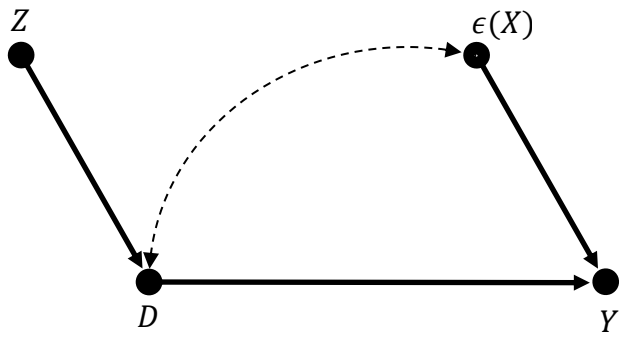
La solution consiste à instrumenter D dans une première étape à l'aide d'au moins un instrument Z tel que $Cov(D, Z) \neq 0$, mais on a la restriction d'exclusion $Cov(\epsilon_1, Z) = 0 \Rightarrow Cov(X, Z) = 0$ et $Cov(\epsilon_2, Z) = 0$. Solution :

1) Spécifier un modèle de régression reliant D à Z : $D = \beta_{41} + Z\beta_{42} + \epsilon_4$
2) Obtenir $D^a \coloneqq Z\beta_{42}^{MC}$
3) Remplacer D par D^a dans M_1 et calculer $\beta_{12}^{VI} \coloneqq Cov(Y, D^a)/V(D^a)$.

On peut montrer que $\beta_{12}^{VI} = \beta_{22}$. Il y a une « porte » $D \leftarrow X \rightarrow Y$ à fermer, ce que l'on peut illustrer avec un graphe de type acyclique dirigé (*acyclic directd graph*). Ce genre de graphique est très développé dans l'approche de la causalité à la Pearl, Glymour et Jewell (2016). Nous empruntons la représentation suivante à Morgan et Winship (2007, 188).

[En profiter pour introduire l'approche graphique en termes de GAD]

Graphique 9.1. Représentation sous forme de graphe acyclique dirigé de la méthode des VI



Variable de collision : Y
Variable de médiation : D
Variable non-observée : ϵ
Back-door path : le chemin détourné de $D \rightarrow Y$ confond la relation causale.
ϵ et D sont affectées de déterminants communs non observés (ce qui ne veut pas nécessairement dire qu'elles sont corrélées)
Instrument : Z

Y est relié à Z à travers D seulement (Z et Y n'ont pas de déterminants communs).

2) Considérons maintenant le problème d'estimation d'un MES ; Angrist et Pischke (2009) et Hamilton (1994). On pense estimer $\boldsymbol{M_1}$, pour étudier l'effet de $\boldsymbol{D}$ sur $\boldsymbol{Y}$. Imaginons qu'il existe un « rapport de force » entre l'offre $\boldsymbol{Y^O}$ et la demande $\boldsymbol{Y}$ d'un bien, ce que le modèle suivant symbolise :

Ex. : Y^O et Y sont l'offre et la demande de travail, et D est une indicatrice (1 si le diplôme est un M2, 0 si c'est une licence).

$$\begin{aligned} Y &= \beta_{11} + D\beta_{12} + \epsilon_1, && (M_1)\\ Y^O &= \gamma_{41} + D\gamma_{42} + \epsilon_4, && (M_4)\\ Y &= Y^O && (M_4') \end{aligned}$$

On suppose que $\boldsymbol{Cov(\epsilon_1, \epsilon_4) = 0}$. On peut simplifier la discussion sans perdre en généralité en supposant que l'on centre $\boldsymbol{Y}$, $\boldsymbol{Y^O}$ et $\boldsymbol{D}$ autour de leur moment d'ordre 1, de sorte que les ordonnées à l'origine ne servent à rien. On suppose également que l'offre est plus pentue que la demande ($\boldsymbol{\beta_{12} - \gamma_{42} > 0}$).

$$\begin{aligned} y &= d\beta_{12} + \epsilon_1, && (m_1)\\ y^O &= d\gamma_{42} + \epsilon_4, && (m_4)\\ y &= y^O && (m_4') \end{aligned}$$

$\boldsymbol{m_4'}$ entraîne $\boldsymbol{d\beta_{12} + \epsilon_1 = d\gamma_{42} + \epsilon_4}$, ce qui, après quelques calculs donne la forme réduite suivante :

$$\begin{cases} d^* = \dfrac{\epsilon_4 - \epsilon_1}{\beta_{12} - \gamma_{42}} \\ y^* = \dfrac{\beta_{12}}{\beta_{12} - \gamma_{42}}\epsilon_4 - \dfrac{\gamma_{42}}{\beta_{12} - \gamma_{42}}\epsilon_1. \end{cases}$$

On voit que $\boldsymbol{d}$ est endogène : $\boldsymbol{Cov(d, \epsilon_1) = -V(\epsilon_1)/(\beta_{12} - \gamma_{42}) \neq 0}$.

De plus, d'après $\boldsymbol{m_1}$, pour différentes valeurs de $\boldsymbol{d}$ on a différentes valeurs de $\boldsymbol{y}$, mais la forme réduite ci-dessus suggère que des facteurs de l'autre équation, $\boldsymbol{m_4}$, remontent vers $\boldsymbol{m_1}$, c.-à-d. on va de $\boldsymbol{y^O}$ vers $\boldsymbol{d}$, donc, de $\boldsymbol{y}$ vers $\boldsymbol{d}$ puisque $\boldsymbol{y = y^O}$ (causalité inverse).

Notons $\boldsymbol{V(\epsilon_4)/(V(\epsilon_4) + V(\epsilon_1))}$ par $\boldsymbol{r}$. Le calcul de $\boldsymbol{\beta_{12}}$ par les MC dans la population (une régression de $\boldsymbol{y}$ sur $\boldsymbol{d}$) donne :

$$\boldsymbol{\beta_{12}^{MC} = \frac{Cov(y,d)}{V(d)} = r\beta_{12} + (1 - r)\gamma_{42} \neq \beta_{12}.}$$

On a une moyenne pondérée de $\boldsymbol{\beta_{12}}$ et $\boldsymbol{\gamma_{42}}$, avec des coefficients de pondération qui dépendent des variances de $\boldsymbol{\epsilon_1}$ et $\boldsymbol{\epsilon_4}$. C'est ce qu'on appelle le biais d'équations simultanées (*simultaneous equation bias*). Si $\boldsymbol{V(\epsilon_4)}$ était grand relativement à $\boldsymbol{V(\epsilon_1)}$, alors $\boldsymbol{r}$ serait proche de 1 ($\boldsymbol{1 - r}$ proche de 0), et la méthode des MC identifierait plutôt la demande ($\boldsymbol{m_1}$).
Dit autrement, si $\boldsymbol{\epsilon_4}$ contenait une variable (instrumentale !!!) qui fasse bouger $\boldsymbol{y^O}$, un curve shifter (Angrist et Krueger, 2001, 70), sans effet direct sur $\boldsymbol{y}$ (seulement *via* $\boldsymbol{d}$), de sorte que $\boldsymbol{V(\epsilon_4) \gg V(\epsilon_1)}$, alors $\boldsymbol{m_4}$ se déplacerait relativement plus que $\boldsymbol{m_1}$. Formellement, si on avait $\boldsymbol{\epsilon_4 = hZ + \epsilon_4'}$, avec $\boldsymbol{Cov(Z, \epsilon_4') = Cov(\epsilon_1, \epsilon_4') = 0}$ par

définition, $\boldsymbol{Cov(Z,\epsilon_1)=0}$ (restriction d'exclusion) et $\boldsymbol{Cov(Z,d)\neq 0}$, alors, on aurait :

$$d=\frac{1}{\beta_{12}-\gamma_{42}}hZ+\frac{\epsilon_4'-\epsilon_1}{\beta_{12}-\gamma_{42}}$$
$$y=\frac{\beta_{12}}{\beta_{12}-\gamma_{42}}hZ+\frac{\beta_{12}\epsilon_4'-\gamma_{42}\epsilon_1}{\beta_{12}-\gamma_{42}}.$$

Soit $\boldsymbol{d^{MC}=\delta^{MC}Z}$ la projection linéaire de $\boldsymbol{d}$ sur $\boldsymbol{Z}$, i.e. $\boldsymbol{Cov(d,Z)/V(Z)\coloneqq\delta^{MC}}$. Or, on a supposé que $\boldsymbol{Z}$ n'était corrélé ni avec $\boldsymbol{\epsilon_4'}$, ni avec $\boldsymbol{\epsilon_1}$, donc $\boldsymbol{Z}$ n'est corrélée ni avec la différence $\boldsymbol{\epsilon_4'-\epsilon_1}$, ni avec une fonction de cette différence $\boldsymbol{\beta_{12}\epsilon_4'-\gamma_{42}\epsilon_1}$. Alors :

$$\boldsymbol{\beta_{12}^{DMC}=\frac{Cov(y,d^{MC})}{V(d^{MC})}=\frac{Cov(d\beta_{12}+\epsilon_1,\delta^{MC}Z)}{V(\delta^{MC}Z)}=\frac{\beta_{12}\delta^{MC}Cov(d,Z)}{(\delta^{MC})^2V(Z)}+\frac{\delta^{MC}Cov(\epsilon_1,Z)}{(\delta^{MC})^2V(Z)}=\beta_{12}.}$$

Z est un instrument. C'est le méthode des DMC. Le nombre d'instruments (ici 1) étant égal au nombre de variables explicatives endogènes, c'est aussi $\boldsymbol{\beta_{12}^{IV}}$

3) Considérons maintenant le problème d'erreur de mesure ; Angrist et Pischke (2009, 114). On estime M1, alors que le bon modèle est le suivant :

$$Y=\beta_{11}+D^*\beta_{12}+\epsilon_1, \qquad (M_1^*)$$
$$D=D^*+\epsilon_5. \qquad (M_5)$$

On combine M1 et M5 :

$$Y=\beta_{11}+D\beta_{12}+\epsilon_1-\epsilon_5\beta_{12}. \qquad (M_1')$$

Ce modèle est le modèle standard avec erreur de mesure (Wooldridge, 2009, 319-322, 525-), en supposant que $E(\epsilon_5|D^*)=0$ (l'erreur de mesure n'est pas corrélée à la variable sans erreur), de sorte que $V(D)=V(D^*)+V(\epsilon_5)\coloneqq\sigma_{D^*}^2+\sigma_{\epsilon_5}^2$. De plus, $E(\epsilon_1|D^*,D)=E(\epsilon_1|D^*)$ (car D dépend de D^* ; on peut aussi dire que conditionnellement à D^*, ϵ_1 ne dépend plus que de ϵ_5, donc c'est comme si D était randomisé), et $Cov(D,\epsilon_1)=0$.

On déduit $Cov(D,\epsilon_5)=\sigma_{\epsilon_5}^2$, et $Cov(D,\epsilon_1-\epsilon_5\beta_{12})=-\beta_{12}\sigma_{\epsilon_5}^2$. Donc, une régression de Y sur D (le modèle M_1 qui, en réalité est M_1') est biaisée :

$$\beta_{12}^{MC}=\frac{Cov(\beta_{11}+D\beta_{12}+\epsilon_1-\epsilon_5\beta_{12},D)}{V(D)}$$
$$=\beta_{12}-\frac{\beta_{12}\sigma_{\epsilon_5}^2}{\sigma_{D^*}^2+\sigma_{\epsilon_5}^2}$$
$$=\frac{\sigma_{D^*}^2}{\sigma_{D^*}^2+\sigma_{\epsilon_5}^2}\beta_{12}.$$

Cette quantité est plus petite que β_{12} (on parle de biais d'atténuation).

La solution consiste à trouver un instrument pour D, donc une variable telle que $Cov\ (\epsilon_1,Z)=0$ et $Cov\ (\epsilon_5,Z)=0$, de sorte que la covariance entre la valeur ajustée de D et l'erreur composite soit nulle (il n'y a plus de biais en théorie).

Z est une mesure alternative à D de D^* : $Z=D^*+\epsilon_6$ et $Cov(\epsilon_5,\epsilon_6)=0$. Cette variable est donc utilisée comme VI de D. On régresse D sur Z, ce qui permet d'obtenir $D^a=\gamma^{MC}Z$. Alors, $Cov(D^a,\epsilon_5)=\gamma^{MC}Cov(Z,\epsilon_5)=0$ par définition.

Il faut reconnaître un problème avec cette approche : il n'est pas évident de trouver un tel instrument, dans la mesure où nous ne sommes déjà pas arrivés à observer la « vraie » variable D^*.

si D^* était un niveau d'études d'une étudiante, Z pourrait être le niveau de sa jumelle, une situation plutôt rare. Ou bien,

Z pourrait aussi être le niveau d'étude des parents. Ici, c'est plus l'idée de trouver des variables exogènes, que d'avoir une seconde mesure, qui compte. On pourra consulter l'exemple 15.5, pp. 523-526, et la page 512 de Wooldridge (2009).

9.1.2. Illustrations

L'illustration qui suit concerne la première situation de recours à une variable instrumentale. C'est l'exemple bien connu de la relation salaire-éducation, largement utilisée par Wooldridge (2003, 2009, 2010) dans ses ouvrages. Le graphique illustre le cas d'un BVO négatif car les aptitudes et le niveau d'années d'études sont supposées substituables à salaire donné. Evidemment, il existe aussi une complémentarité entre les deux. Comme le montre Wooldridge (2009) dans son chapitre 15, on peut avoir les deux signes, selon le choix de l'instrument. En ce sens, il n'y a pas d'instrument plus réaliste qu'un autre pour résoudre le problème de BVO. D'où l'idée d'en considérer plusieurs simultanément.

Parfois, les instruments peuvent être très subtiles, comme l'illustrent Angrist et Pischke (2009, 115-129), qui reprennent un travail fameux coécrit par le premier auteur et Alan Krueger (Angrist et Krueger, 1991), sur la relation salaire-niveau d'études aux États-Unis. Il s'agit du trimestre de naissance.

La motivation est que le trimestre est corrélé à l'âge d'arrêt des études, et donc le nombre d'années d'études à l'âge d'arrêt légal, avec des variations sur cet âge légal entre États. *Idem* concernant le trimestre de naissance et l'âge d'entrée à l'école. Pour un âge légal d'entrée à l'école donné, le fait d'être né en début ou en fin d'année est corrélé au nombre d'années d'études.
Et pour une année de naissance donnée, les auteurs trouvent également que le trimestre de naissance est corrélé au résultat, le salaire.

[Programmes Stata + encadré page suivante]

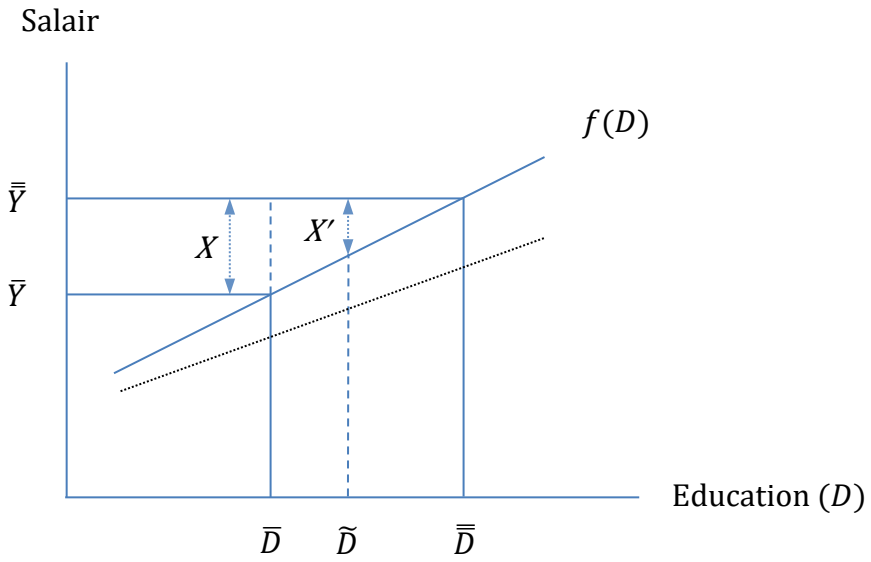
La relation éducation-salaire

Angrist et Pischke (2009)

Typiquement, en matière de relation salaire-éducation, Y est le salaire, D le niveau d'éducation ; il peut s'agir d'un variable dichotomique (l'étudiant a une Licence ou un Master) ou entière (le nombre d'années d'études), comme dans l'article d'Angrist et Krueger (1991), repris dans Angrist et Pischke (2009). Autrement dit, le salaire est la variable de résultat et le niveau d'études le traitement.

La VO X symbolise des caractéristiques intrinsèques des étudiants (*ability*). Quant à l'instrument, Z, on a le choix, même trop de choix, de sorte qu'il y a des instruments plus forts que d'autres. Sans rentrer dans la théorie, il y a des instruments qui remplissent plus des conditions à satisfaire. Pour ce qui est de la relation salaire-éducation, la note à un test de QI, une mention au BAC (ou équivalent) sont des VI possibles, quoiqu'imparfaits.

Le graphique suivant représente la source du problème, le rôle que peut jouer la VO



Pour un même niveau d'éducation $\bar{D}$, un étudiant a des aptitudes, des compétences acquises en dehors du système scolaire, qui font que son salaire, au lieu d'être $\bar{Y}$, sera $\bar{\bar{Y}}$, qui correspond en fait au salaire d'un étudiant qui aurait $\bar{\bar{D}} > \bar{D}$ années d'études. On peut voir qu'avec des aptitudes d'un niveau $X' < X$, il faut un niveau d'éducation $\tilde{D}$ supérieur à $\bar{D}$ mais inférieur à $\bar{\bar{D}}$, afin d'espérer le même salaire, $\bar{\bar{Y}}$.

Dans cet exemple on a une relation inverse entre le traitement (nombre d'années d'études) et la VO (aptitudes) pour un résultat (salaire) donné. On devrait donc s'attendre à un BVO négatif. Par conséquent, $\beta_{12}^{MC} < \beta_{12}$, le coefficient qui nous intéresse, mais calculé à partir du modèle faux, M_1.

Dans la réalité, aptitudes et années d'études sont aussi complémentaires, de sorte que la relation entre le traitement et la VO n'est pas forcément négatif. Le fait de conditionner sur le salaire entraîne une relation théorique négative entre ces deux variables. On pourrait plutôt envisager qu'un étudiant ayant moins d'aptitudes, obtient toujours un salaire inférieur, quel que soit le nombre d'années d'études (la courbe en trait pointillé).

Note : « VO » (variable omise), « VI » (variable instrumentale).

9.2. Estimateur LATE

Il est normal de penser que les individus qui s'enrôlent dans un programme, lors d'une expérimentation (ou qui bénéficient d'une politique publique), sont seulement ceux qui y avaient été affectés (ou qui y étaient « éligibles »). On peut aussi se demander ce qu'auraient fait ces individus s'ils n'avaient pas été éligibles. Le fait de distinguer l'éligibilité au traitement du traitement lui-même, le traitement est souvent synonyme de participation dans la littérature.

Cette distinction, entre éligibilité (*eligibility*, *qualification*) en t et traitement (ou participation) en t, a deux conséquences :

Elle élargit la liste des variables (en fait, des instruments) permettant d'identifier l'effet causal de la politique publique. Nous avons vu ce point dans l'exemple de la lutte contre la pauvreté (**chapitre 7**), dans lequel une famille bénéficie d'une aide ou pas, et la variable d'éligibilité est la surface agricole cultivée par cette famille.

On a vu dans la **section 9.1** qu'une VI, au sens classique, est une variable qui n'est pas corrélée au terme d'erreur de l'équation reliant le résultat à la variable dont on veut mesurer l'effet (restriction d'exclusion).

Ici, l' instrument Z idéal est :

Randomisé ! C'est l'affectation aléatoire dans les groupes de traitement (MATAC) ; Heckman (1996).

Réalisé avant D, afin d'être sûr que $D \nrightarrow Z$.

Formellement, le PFIC s'applique désormais à D ; n'importe quelle unité de la population étudiée est soit éligible, soit ne l'est pas *a priori*. Ainsi :

$$D = \begin{cases} D(1) \; si \; Z = 1 \\ D(0) \; si \; Z = 0 \end{cases} \Leftrightarrow D = D(1)Z + D(0)(1 - Z). \qquad (9.1)$$

$D(z)$, $z \in \{0\,;1\}$ sont les traitements contrefactuels. On a écrit ci-dessus une équation de Rubin pour le traitement, qui est une manière de montrer que l'instrument (l'éligibilité) a un effet causal sur la participation.

Au passage, si les individus respectent leur affectation, alors $D(1) = 1$ et $D(0) = 0$, de sorte que $D = Z$!

Comme Imbens et Angrist (1994) et Imbens et Wooldridge (2009), la présentation qui suit ne considère pas de variable explicatives supplémentaires (on n'a pas besoin de CIA). Nous suivons de près Wooldridge (2003) et l'article qui a introduit l'estimateur LATE, Imbens et Angrist (1994).

Combinons $D = D(1)Z + D(0)(1 - Z)$ avec $Y = Y(1)D + Y(0)(1 - D)$, après avoir réécrit ces équations ainsi : $D = D(0) + (D(1) - D(0))Z$ et $Y = Y(0) + (Y(1) - Y(0))D$. Nous obtenons :

$$\begin{aligned} Y &= Y(0) + (Y(1) - Y(0))D \\ &= Y(0) + \big(Y(1) - Y(0)\big)[D(0) + \big(D(1) - D(0)\big)Z\,] \\ &= Y(0) + \big(Y(1) - Y(0)\big)D(0) + \big(Y(1) - Y(0)\big)\big(D(1) - D(0)\big)Z. \end{aligned}$$

Cette extension au cas de traitements contrefactuels conduit au concept de RP généralisés, selon l'expression de Angrist et Pischke (2009, 151).

On ajoute une supposition clé sur les résultats et traitements contrefactuels :

$$Z \perp (Y(1), Y(0), D(1), D(0)).$$

Attention, **ici** l'argument est D et **là** Z.

Comme pour l'ignorabilité, il faut comprendre ici que l'indépendance porte entre Z et n'importe qu'elle fonction des variables $Y(1), Y(0), D(1), D(0)$. Une interprétation possible de cette supposition :

> L'instrument est déterminé avant la décision de participation, et indépendamment des résultats qu'entrainerait l'exposition ou pas au traitement. Cette supposition veut dire que nous avons entre autres :

S9.1 : $E(Y(d)|z) = E(Y(d))\ \forall\ d \in \{0\,;1\}, \forall\ z \in \{0\,;1\}$,
S9.2 : $E(D(1)|z) = E(D(1))\ \forall\ z \in \{0\,;1\}$, et idem pour $E(D(0)|z)$.
S9.3 : $\Pr(D = 1|Z = 1) \neq \Pr(D = 0|Z = 0)$

> La première supposition signifie que le RP d'une unité d'observation selon les différents traitements ne change pas avec l'éligibilité. On peut l'écrire autrement, $E(Y(d)|Z = 1) = E(Y(d)|Z = 0)$. Nous verrons l'importance de **S9.3** plus loin.

Par conséquent, nous avons :

$$E(Y|Z = 1) - E(Y|Z = 0) = E\big[\big(Y(1) - Y(0)\big)\big(D(1) - D(0)\big)\big].$$

En effet,

$$\begin{aligned} E(Y|Z = 1) &= E(Y(0)|Z = 1) + E\big[\big(Y(1) - Y(0)\big)D(0)|Z = 1\big] \\ &\quad + E\big[\big(Y(1) - Y(0)\big)\big(D(1) - D(0)\big)Z|Z = 1\big] \\ &= E(Y(0)) + E\big[\big(Y(1) - Y(0)\big)D(0)\big] + E\big[\big(Y(1) - Y(0)\big)\big(D(1) - D(0)\big)\big], \end{aligned}$$

et

$$\begin{aligned} E(Y|Z = 0) &= E(Y(0)|Z = 0) + E\big[\big(Y(1) - Y(0)\big)D(0)|Z = 0\big] \\ &\quad + E\big[\big(Y(1) - Y(0)\big)\big(D(1) - D(0)\big)Z|Z = 0\big] \\ &= E(Y(0)) + E\big[\big(Y(1) - Y(0)\big)D(0)\big]. \end{aligned}$$

D'où :

$$E(Y|Z = 1) - E(Y[Z = 0) = E\big[\big(Y(1) - Y(0)\big)\big(D(1) - D(0)\big)\big].$$

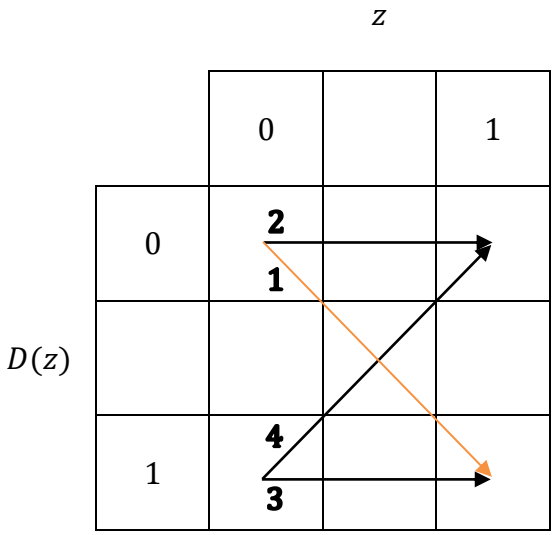
■

Deux points importants sur le produit $\big(Y(1) - Y(0)\big)\big(D(1) - D(0)\big)$

> La présence de $D(1) - D(0)$ suggère de considérer le passage des états « pas éligibles » à « éligible » ($Z = 0$ à $Z = 1$), sans qu'il y ait le passage du temps (il s'agit d'états contrefactuels).
> Le produit peut clairement être négatif ! $D(z) \in \{0\,;1\}$, $\forall\ z \in \{0\,;1\}$, donc la différence $D(1) - D(0)$ ci-dessus peut prendre trois valeurs : -1, 0 et 1.

[**Un graphique + un tableau**]

Graphique. Scénarios éligibilité-participation



Chaque flèche va dans le sens de « pas éligible » ($Z = 0$) à « éligible » ($Z = 1$). Nous avons numéroté les différents cas dans le tableau suivant, qui existe sous d'autres formes plus ou moins formelles dans la littérature académique. Il s'agit de compléter la discussion menée par Imbens et Angrist (1994) ; voir également Wooldridge (2003).

Tableau. Effet causal de l'éligibilité sur la participation

		participation des non-éligibles [a]	participation des éligibles [a]	
	$D(0) - D(1)$	$D(0)$	$D(1)$	Type [b]
1	-1	0	1	Conformiste (*complier*)
2	0	0	0	Prudent (*never-taker*)
3	0	1	1	Opportuniste (*always-taker*)
4	1	1	0	Anti conformiste (*defier*)

Notes : a. 0 « participe », 1 « ne participe pas ».
b. Les adjectifs décrivant les différents types en anglais sont ceux que l'on trouve dans la littérature. Il s'agit de mots génériques. Les traductions en français sont personnelles et pourraient être améliorées.

Devrait-on retenir tous ces cas dans l'effet causal que nous estimons ? On peut donner une illustration sur les différents types.

Il est normal d'exclure **4** car on s'attendrait plutôt à ce qu'un individu qui ne participe pas, s'il était éligible, ne participe pas non-plus s'il ne l'était pas, le cas **2**. Donc, d'une certaine manière, **4** et **2** ne sont pas cohérents.

On constate qu'une fois **4** éliminé, les trois cas qui restent vérifient $D(1) \geq D(0)$, i.e. :

S9.4 : $D(w)$ est monotone (Imbens et Angrist, 1994, 469).

Ensuite, il y a deux cas qui, par construction, ne vont pas jouer de rôle sur $E\big[\big(Y(1)-Y(0)\big)\big(D(1)-D(0)\big)\big]$. En effet, cette espérance peut se réécrire à partir de la loi des espérances itérées de manière à faire apparaître les différentes valeurs de $D(1)-D(0)$. Notons les distributions jointe de $Y(1), Y(0), D(1), D(0)$ par J, la distribution conditionnelle de $Y(1), Y(0)$ sachant $D(1), D(0)$ par C, et la marginale de $D(1), D(0)$ par M et la probabilité que $D(1)-D(0)=\delta$ par $m(\delta)$. Alors l'espérance se réécrit :

$$\begin{aligned} E\big[\big(Y(1)-Y(0)\big)\big(D(1)-D(0)\big)\big] &= -E\big[\big(Y(1)-Y(0)\big)\big(D(0)-D(1)\big)\big] \\ &= -E_D\big\{\big(D(0)-D(1)\big)E_C[Y(1)-Y(0)|D(0)-D(1)]\big\} \\ &= -1E_C[Y(1)-Y(0)|D(0)-D(1)=-1]m(-1)+ \\ &\quad 0E_C[Y(1)-Y(0)|D(0)-D(1)=0]m(0). \end{aligned}$$

Par conséquent, le 0 élimine les cas **2** et **3**. Il ne reste que le cas **1**, le moins inattendu à vrai dire. Par conséquent, on a :

$$E\big[\big(Y(1)-Y(0)\big)\big(D(1)-D(0)\big)\big] = E_C[Y(1)-Y(0)|D(1)-D(0)=1]m(1),$$

et donc :

$$E(Y|Z=1)-E(Y[Z=0)=E_C[Y(1)-Y(0)|D(1)-D(0)=1]m(1).$$

Imbens et Angrist (1994) appellent la quantité $E_C[Y(1)-Y(0)|D(1)-D(0)=1]$ LATE. C'est l'effet de la participation sur la sous-population des conformistes. Nous souhaitons estimer cet effet. De l'égalité ci-dessus, on déduit LATE :

$$\text{LATE} \coloneqq E_C[Y(1)-Y(0)|D(1)-D(0)=1]=\frac{E(Y|Z=1)-E(Y[Z=0)}{m(1)}.$$

Dans la sous-population des conformistes, $E(Y|Z=1)-E(Y|Z=0)$ est observable, mais également $m(1)$, grâce à la supposition **S9.2**. et l'équation de Rubin appliquée au traitement, S9.1. En effet,

$$\begin{aligned} m(1) &= \Pr(D(1)-D(0)=1) \\ &= \mathrm{E}\big(D(1)-D(0)\big) \\ &= E\big(D(1)\big)-E\big(D(0)\big) \\ \textbf{(S9.2)}\quad &= E(D(1)|Z=1)-E(D(0)|Z=0) \\ \textbf{(9.1)}\quad &= E(D|Z=1)-E(D|Z=0) \\ &= \Pr(D=1|Z=1)-\Pr(D=0|Z=0). \end{aligned}$$

D'où, $E(Y|Z=1)-E(Y[Z=0)=E_C[Y(1)-Y(0)|D(1)-D(0)=1][\Pr(D=1|Z=1)-\Pr(D=0|Z=0)]$, et donc :

$$E_C[Y(1)-Y(0)|D(1)-D(0)=1]=\frac{E(Y|Z=1)-E(Y[Z=0)}{\Pr(D=1|Z=1)-\Pr(D=0|Z=0)}.$$

Cette quantité est le ratio des coefficients de pente des régressions de Y sur Z et D sur Z, donc entre $Cov(Y,Z)/V(Z)$ et $Cov(D,Z)/V(Z)$. Nous avons donc :

$$E_C[Y(1) - Y(0)|D(1) - D(0) = 1] = \frac{Cov(Y,Z)}{Cov(D,Z)}.$$

Application à l'évaluation de la microfinance (Khandker, Koolwal et Hussain, 2010)

[**late.do**]

Remarques :

Nous faisons remarquer que nous n'avons pas exclu les cas 2 et 3. Ils sont mathématiquement éliminés dans le développement de $E\big[\big(Y(1) - Y(0)\big)\big(D(1) - D(0)\big)\big]$, mais des individus qui se comporteraient comme l'indique cette hypothèse peuvent très bien figurer dans l'échantillon étudié.

Par exemple, si l'éligibilité est déterminée de manière aléatoire, en tirant au hasard des étudiant dans la population, que l'on rend ainsi éligibles à une bourse. On les met au courant. Puis, on les laisse décider librement de déposer un dossier ou pas. Alors, **S9.4** signifie qu'un étudiant qui déposerait un dossier s'il n'était pas éligible, $D(0) = 1$, déposerait aussi un dossier s'il était éligible, $D(1) = 1$.

Il y a au moins deux problèmes avec le LATE :

La sous-population sur laquelle on estime l'effet causal, Pr $(D(1) - D(0) = 1)$ n'est pas facile à identifier. On sait qu'il s'agit *in fine* de la sous population qui vérifie **S9.3**.

La définition de « local » dépend des VI à disposition.

9.3. Estimateur heckit

9.3.1. BS à la Heckman (1979)

Distinguons immédiatement le BS à la Heckman (1979) de celui vu jusqu'à présent :

> Supposons que nous nous intéressions au salaire d'une population de travailleurs ayant suivi une formation professionnelle, pendant que d'autres travailleurs ne l'ont pas suivie. On ne devrait pas supposer que ces deux groupes sont comparables. En effet, les travailleurs qui ne l'ont pas suivie n'en avaient peut-être pas besoin, ou bien n'ont pas les moyens de suivre ce type de formation, etc. Ces facteurs sont non-seulement corrélés à la participation, mais aussi au résultat.
>
> > Chez les travailleurs qui ne suivent pas de formation, il y en a qui ont acquis suffisamment de compétences durant leur études, ou qui ont suffisamment d'expérience professionnelle, de sorte que la formation n'aurait pas d'effet sur eux s'ils la suivaient
>
> C'est la source du BS que nous avons vu jusqu'ici. Cependant, il n'est possible d'étudier la relation salaire-formation que pour ceux qui ont trouvé du travail. Ces derniers sont peut-être différents de ceux qui n'ont pas trouvé de travail, qu'ils aient eu la formation ou pas. C'est le BS à la Heckman (1979).

Une difficulté lorsque l'on s'intéresse au BS à la Heckman (1979) est que ce dernier change d'approche pour adopter le MCR dans les années 1980. C'est Rubin (2004, p. 345) qui nous le dit :

> « This framework using potential outcomes to define causal effects in general is now relatively well accepted in many fields […], in economics, see the transition to adopt it reflected by comparing Heckman (1979) to Heckman (1989) […] »

Les motivations d'Heckman sont si proches de celles de Rubin, Rosenbaum, …, que nous pouvons nous demander pourquoi il n'utilise pas le MCR (Rubin, 1974) pour parler du BS vu dans le **chapitre 4** [peut-être que Rubin ne parlait pas encore de BS en 1974-].

L'exemple pris par Heckman (1979, 153) renforce notre étonnement :

> « The wages of migrants do not, in general, afford a reliable estimate of what nonmigrants would have earned had they migrated. The earnings of manpower trainees do not estimate the earnings that nontrainees would have earned had they opted to become trainees. […] Comparisons of the wages of migrants with the wages of nonmigrants (or trainee earnings with nontrainee earnings, etc.) result in a biased estimate of the effect of a *random* 'treatment' of migration, manpower training […] »

Heckman (1979, 153-154) suggère que $E(Y|D=1)$ n'est pas un estimateur fiable de $E(Y(1)|D=0)$, et que $E(Y|D=1)-E(Y|D=0)$ est un estimateur biaisé de $E(Y(1)-Y(0))$, sans utiliser ces notations. Probablement du fait que dans le problème de BS à la Heckman, il existe des individus pour lesquels $E(Y|D=d)$, $d=0{,}1$ ne sont pas observés. De plus, il s'intéresse à $E(Y(1)|D=0)$, alors que le pour estimer l'ECMT, le problème est surtout de trouver un estimateur de $E(Y(0)|D=1)$. Enfin, il compare à la fois des RP et des RO tronqués.

> Cet étonnement est encore plus fort sachant qu'Heckman (1979, 154) parle de *confounding*.

> « Such procedures have the same effect on structural estimates as self-selection: fitted regression functions confound the behavioural parameters of interest with parameters of the function determining the probability of entrance into the sample. »

À la différence du problème de biais de variable omise (BVO), on a ici un problème de biais d'individus omis (BIO). C'est ce que l'on appelle un échantillon tronqué. La question est de savoir si cet troncature est une « sélection » non-aléatoire (le MAT est caractérisé par une « *self selection* », pour reprendre les mots de l'auteur, d'individus qui ont décidé de ne pas travailler, pour reprendre l'exemple d'introduction, ou pas). On peut « introduire » dans la régression d'intérêt ces individus omis via l'inclusion d'une variable supplémentaire : la probabilité que ces individus soient observés.

En fait, on va introduire l'inverse du ratio de Mills

Du coup, c'est comme-ci Heckman (1979) modélisait ce problème de BS comme une mauvaise spécification économétrique de l'équation d'intérêt, le problème étant renvoyé dans l'erreur de cette équation. Ce biais de spécification est atténué quand $N \to \infty$. En plus de corriger le BS, on peut tester sa présence.

L'auteur est certainement influencé par le test d'exogénéité d'Hausman (1978).

Cette section s'appuie sur Morgan et Winship (2007, 184-186)

La commande sur Stata est `heckman`

[Documentation Stata, 2013]

9.3.2. Problème de troncature

Heckman (1979) se focalise sur une variable dépendante continue Y_1 (la variable de l'équation d'intérêt), une variable de sélection Y_2, des vecteurs de variables explicatives dans chacune, X_1 et X_2, ainsi que deux termes d'erreurs : U_1 et U_2. Chez Heckman, le problème est le suivant :

$$Y_{1i} = X'_{1i}\beta_1 + U_{1i}$$
$$Y_{2i} = X'_{2i}\beta_2 + U_{2i}$$

On se situe dans le cadre d'un échantillon aléatoire de N individus $i \in \{1\ ;\ \dots\ ;\ N\} \coloneqq I$. X_{1i} et β_1 sont de dimension $K_1 \boxtimes 1$ et X_{2i} et β_2 sont de dimension $K_2 \boxtimes 1$, $\boldsymbol{X}_1 \coloneqq (X_{11} \cdots X_{1N})'$ et $\boldsymbol{X}_2 \coloneqq (X_{21} \cdots X_{2N})'$ sont de plein rang, $E(U_{1i}U_{2j}) \coloneqq 1_j(i)\sigma_{12}$. Pour rester cohérent avec ce que nous avons vu jusqu'à présent, nous allons supposer $K_1 = 2$, $K_2 = 3$, $X'_1 \coloneqq (1, D)$, $X'_2 \coloneqq (1, D, Z)$, $U_1 \coloneqq \epsilon_1$, $U_2 \coloneqq \epsilon_2$ et travailler dans la population. (on retire les indices « i »)

$$Y_1 = \beta_{11} + D\beta_{12} + \epsilon_1 \qquad (9.1)$$
$$Y_2 = \beta_{21} + D\beta_{22} + Z\beta_{23} + \epsilon_2 \qquad (9.2)$$

On est dans la situation de troncature suivante : on observe Y_1 si $Y_2 \geq 0$, autrement dit, nous pouvons définir la variable aléatoire $S = 1(Y_2 \geq 0)$ et réécrire 9.1 et 9.2 :

$$SY_1 = S\beta_{11} + SD\beta_{12} + S\epsilon_1. \qquad (9.1')$$
$$S = 1(\beta_{21} + D\beta_{22} + Z\beta_{23} + \epsilon_2 \geq 0) \qquad (9.2')$$

si $S = 1$, (9.1') est équivalente à (9.1), sinon, (9.1') est équivalente à $0 = 0 + 0 + 0$. Contrairement à Heckman, Wooldridge suppose que bien que Y_1 ne soit pas observé pour certaines valeurs de Y_2, en revanche, X_1 (ici D) est toujours observé.

Comme précédemment, il faut faire quelques hypothèses :

S9.5 $E(\epsilon_1|D,Z) = 0$,
S9.6 $E(\epsilon_2|D,Z) = 0$,
S9.7 $\epsilon_2 \sim N(0,\sigma_{22})$,
S9.8 $c(\epsilon_1,\epsilon_1|D,Z) = j(\epsilon_1,\epsilon_1)$.

Calculons pour $S = 1$

$$E(Y_1|D,Z,\epsilon_2) = E(\beta_{11} + D\beta_{12} + \epsilon_1|D,Z,\epsilon_2) = \beta_{11} + D\beta_{12} + E(\epsilon_1|D,Z,\epsilon_2).$$

Or S9.6 et S9.8 impliquent $E(\epsilon_1|D,Z,\epsilon_2) = E(\epsilon_1|\epsilon_2)$.

En effet
$$\begin{aligned}
\int \epsilon_1 c(\epsilon_1|D,z,\epsilon_2) &= \int \epsilon_1 \frac{j(\epsilon_1,\epsilon_2,D,Z)}{j(\epsilon_2,D,Z)} \\
&= \int \epsilon_1 \frac{c(\epsilon_1,\epsilon_2|D,Z)j(D,Z)}{j(\epsilon_2,D,Z)} \\
\text{(S9.8)} \quad &= \int \epsilon_1 \frac{j(\epsilon_1,\epsilon_2)j(D,Z)}{j(\epsilon_2,D,Z)} \\
\text{(S9.7)} \quad &= \int \epsilon_1 \frac{j(\epsilon_1,\epsilon_2)j(D,Z)}{m(\epsilon_2)j(D,Z)} \\
&= \int \epsilon_1 \frac{j(\epsilon_1,\epsilon_2)}{m(\epsilon_2)} \\
&= \int \epsilon_1 c(\epsilon_1|\epsilon_2) \\
&= E(\epsilon_1|\epsilon_2).
\end{aligned}$$

Par conséquent,

$$E(Y_1|D,Z,\epsilon_2) = \beta_{11} + D\beta_{12} + E(\epsilon_1|\epsilon_2).$$

L'estimateur de Heckman implique le calcul de $E(\epsilon_1|\epsilon_2)$ sachant que (ϵ_1,ϵ_2) suit une loi normale bivariée. Alors,

$$E(Y_1|D,Z,\epsilon_2) = \beta_{11} + D\beta_{12} + \rho\epsilon_2/\sigma_{22}^{1/2}.$$

Dans la mesure où l'on n'observe pas ϵ_2, on passe à l'espérance suivante (Wooldridge, 2009, 609) où l'on conditionnera sur $S = 1$ (que l'on observe) plutôt que ϵ_2 :

$$E(Y_1|D = d, Z = z, S = 1) = \beta_{11} + d\beta_{12} + \frac{\rho}{\sigma_{22}^{1/2}} E(\epsilon_2|\epsilon_2 \geq -(\beta_{21} + d\beta_{22} + z\beta_{23})).$$

Notons que $E(\epsilon_2|\cdot)$ apparaît, du fait que l'on ne conditionne plus sur ϵ_2. D'autre part, nous avons sauté une étape. En effet, conditionnellement à S, $E(\epsilon_2/\sigma_{22}^{1/2}|D = d, Z = z, S = 1)$ ne dépend ni de D, ni de Z, de sorte que l'on a $E(\epsilon_2/\sigma_{22}^{1/2}|S = 1)$. Or, $S = 1$ quand $\epsilon_2 \geq -(\beta_{21} + d\beta_{22} + z\beta_{23})$, d'où $E(\epsilon_2|\epsilon_2 \geq -(\beta_{21} + d\beta_{22} + z\beta_{23}))$.

Notons $(\beta_{21} + d\beta_{22} + z\beta_{23})/\sigma_{22}^{1/2}$ par $x\beta_2$, alors

$$\frac{\rho}{\sigma_{22}^{1/2}} E(\epsilon_2/\sigma_{22}^{1/2}|\epsilon_2/\sigma_{22}^{1/2} \geq -x\beta_2) = \frac{\rho}{\sigma_{22}^{1/2}} E(N(0\,;\,1)|N(0\,;\,1) \geq -x\beta_2)$$

$$= \frac{\rho}{\sigma_{22}^{1/2}} \frac{\phi(x\beta_2)}{\Phi(x\beta_2)}.$$

Pour conclure,

$$E(Y_1|D = d, Z = z, S = 1) = \beta_{11} + d\beta_{12} + \frac{\rho}{\sigma_{22}^{1/2}} \frac{\phi(x\beta_2)}{\Phi(x\beta_2)}$$

Notons le dernier terme de la régression ci-dessus par $\rho\lambda(x\beta_2)/\sigma_{22}^{1/2}$. La quantité $\lambda(x\beta_2)$ est l'inverse du ratio de Mills évalué au point $x\beta_2$. On peut déduire de ce résultat, que le paramètre qui nous intéresse, β_{12}, ne pourra être estimé sans biais que si l'on tient compte de λ. Une estimation par les moindres carrés n'élimine pas le biais. En effet,

$$\beta_{12}^{MC} = \beta_{12} + \frac{\rho}{\sigma_{22}^{1/2}} \left(\frac{\phi(x_1\beta_2)}{\Phi(x_1\beta_2)} - \frac{\phi(x_0\beta_2)}{\Phi(x_0\beta_2)} \right),$$

où $x_1\beta_2 \coloneqq -(\beta_{21} + \beta_{22} + z\beta_{23})/\sigma_{22}^{1/2}$ et $x_0\beta_2 \coloneqq -(\beta_{21} + z\beta_{23})/\sigma_{22}^{1/2}$.

Il est généralement admis que le biais est négatif dans la mesure où $\lambda(x\beta_2)$ est une fonction monotone croissante de $x\beta_2$. Or, $x_1\beta_2 - x_0\beta_2 = -\beta_{22}$. Donc, $x\beta_2$ diminue. À condition que $\beta_{22} > 0$, $\lambda(x\beta_2)$ diminue aussi. Par conséquent, $\lambda(x_1\beta_2) < \lambda(x_0\beta_2)$, et donc $\beta_{12}^{MC} < \beta_{12}$.

Nous voyons au passage l'importance (i) de supposer que D figure aussi dans l'équation de sélection, et (ii) qu'il existe au moins une variable de l'équation de sélection qui ne figure pas dans l'équation d'intérêt 9.1.

(i) Nous n'aurions pas sinon $\boldsymbol{\beta_{22}}$ dans 9.2', alors $\boldsymbol{x_1\beta_2 = x_0\beta_2}$ et $\boldsymbol{\lambda(x_1\beta_2) = \lambda(x_0\beta_2)}$, laissant croire que l'estimation par les MC est sans biais. Non, car $\frac{\boldsymbol{\phi(x\beta_2)}}{\boldsymbol{\Phi(x\beta_2)}}$ ne dépendrait pas de $\boldsymbol{D}$. […]

(ii) Si $\boldsymbol{D}$ est la seule variable qui figurait dans les équations 9.1 et 9.2, il y aurait un problème de multicolinéarité évident entre cette variable et $\boldsymbol{\lambda(x\beta_2)}$ dont les valeurs ne varieraient qu'en fonction de $\boldsymbol{D}$. De ce point de vue, $\boldsymbol{Z}$ joue le rôle d'instrument, c'est une variable qui détermine la sélection et qui a un effet sur $\boldsymbol{Y_1}$ à travers $\boldsymbol{\lambda}$, pas directement.

Une possibilité est d'ajouter $\lambda(x\beta_2)$ dans la régression 9.1. Car, on peut considérer que ne pas le faire c'est comme si on avait un biais de VO. Si ϵ_1 et ϵ_1 ne sont pas très corrélés, alors ρ est proche de 0 (en théorie), et le biais est atténué.

D'après S9.7, la variable S dans 9.2' suit un probit :

$$\begin{aligned}
&\Pr(S = 1|D = 1, Z = z) = \Pr(\beta_{21} + \beta_{22} + z\beta_{23} + \epsilon_2 \geq 0|D, Z) \\
&= \Pr(\epsilon_2 \geq -(\beta_{21} + \beta_{22} + z\beta_{23})|D, Z) \\
&= \Pr\left(\epsilon_2/\sigma_{22}^{1/2} \geq -x_1\beta_2\middle|D, Z\right) \\
&= 1 - \Pr\left(\epsilon_2/\sigma_{22}^{1/2} \leq -x_1\beta_2\middle|D, Z\right) \\
&= 1 - \Phi(-x_1\beta_2) \\
&= 1 - (1 - \Phi(x_1\beta_2)) \\
&= \Phi(x_1\beta_2).
\end{aligned}$$

La procédure est la suivante (Wooldridge, 2009, 610) :

1) Estimer le probit sur tout l'échantillon, afin d'obtenir $\boldsymbol{\beta_2^{probit}}$, l'estimation probit de $\boldsymbol{\beta_2}$. La variable expliquée est la variable $\boldsymbol{S}$ qui prend la valeur 1 si $\boldsymbol{Y_1}$ est observé et 0 sinon. Les variables explicatives sont $\boldsymbol{D}$, $\boldsymbol{Z}$ (ne pas oublier la constante).
2) Calculer l'inverse du ratio de Mills, $\boldsymbol{\lambda(x\beta_2^{probit})}$ pour chaque individu pour qui $\boldsymbol{S = 1}$.
3) Régresser $\boldsymbol{Y_1}$ sur $\boldsymbol{D}$ et $\boldsymbol{\lambda(x\beta_2^{probit})}$ afin d'obtenir $\boldsymbol{\beta_{12}^{heckit}}$ l'estimation selon cette méthode du coefficient qui nous intéresse depuis le départ. Le coefficient devant la variable $\boldsymbol{\lambda(x\beta_2^{probit})}$ peut être testé. Il s'agit d'un test du BS à la Heckman pour l'hypothèse nulle $\boldsymbol{H_0 : \rho = 0}$ (en effet, on a $\boldsymbol{\rho/\sigma_{22}^{1/2}}$ devant $\boldsymbol{\lambda(x\beta_2^{probit})}$, donc, si le coefficient devant $\boldsymbol{\lambda(x\beta_2^{probit})}$ n'est pas significativement différent de 0, c'est nécessairement que $\boldsymbol{\rho/\sigma_{22}^{1/2}}$ donc $\boldsymbol{\rho}$ ne l'est pas, en supposant bien sûr que $\boldsymbol{\sigma_{22}^{1/2}}$ ne soit pas énorme). Si l'on ne rejette pas $\boldsymbol{H_0}$, il n'y a pas de BS ni besoin de correction. Dans le cas contraire, on a corrigé le biais, mais on a un problème de précision, les erreurs standards sont mal estimées.

[Encadré (page suivante)]

Exemple de problème de biais de sélection à la Heckman (1979)

Les travaux originaux de James Heckman sur le BS au début des années 1970 ne sont pas très éloignés de ceux que l'on a vu jusqu'à présent. Une petite recherche de paternité divise cet auteur et Rubin, et plus généralement des économètres et des statisticiens, à propos de l'approche contrefactuelle en inférence causale.

En revanche, s'il y a un travail dont il ne fait aucun doute qu'il a été développé par Heckman, c'est le modèle avec troncage que nous avons reproduit dans cette section.

Par exemple, si la première équation décrit les salaires de femmes (l'exemple classique de la littérature), selon la décision d'une femme de travailler ou pas, nous observerons ou pas son salaire. Cette décision dépend d'une variable observée, X, le niveau d'études salaire proposé, par exemple. La variable D indique si cette femme a suivi une formation ou pas.

Cette situation est très probable pour les femmes à qui le salaire proposé est bas. Une autre manière de dire cela est que l'échantillon des salaires est « biased upward » ; le salaire moyen que l'on observerait si toutes les femmes travaillaient (la population), serait plus bas que celui de l'échantillon.

> En termes économiques, le salaire proposé est inférieur au salaire de réservation. Et, de ce point de vue, il y a des femmes ne désirant pas travailler dont le salaire de réservation est élevé au point que le salaire auquel elles seraient prêtes à travailler est supérieur au salaire observé de celles qui travaillent.
>
> Le salaire de réservation est la VO. Ce n'est pas la peine de chercher à l'observer. En revanche, nous pouvons chercher un instrument pour la variable de traitement, qui satisfasse les conditions usuelles : être corrélé avec le traitement, ne pas être corrélé (directement) avec le résultat et les VO qui affectent ce résultat.

Dans cet exemple, il peut s'agir du nombre d'enfants de la femme (Stata 13, `heckman`, p. 782).

9.4. Exercices … à développer

9.4.1)

9.4.2)

Bibliographie

6. ABADIE, A. (2021): "Using synthetic controls: feasibility, data requirements, and methodological aspects," *Journal of Economic Literature*, 59, 391-425.

7. 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*, 105, 493-505.

8. — (2015): "Comparative politics and the synthetic control method," *American Journal of Political Science*, 59, 495–510.

9. ABADIE, A., DRUKKER, D., HERR, J. L., ET IMBENS, G. W. (2004): "Implementing matching estimators for average treatment effects in Stata," *The Stata Journal*, 4, 290-311.

10. ABADIE, A., ET GARDEAZABAL, J. (2003): "The economic costs of conflict: a case study of the Basque Country," *American Economic Review*, 93, 112-132.

11. ABADIE, A., ET IMBENS, G. W. (2011): "Bias-corrected matching estimators for average treatment effects," *Journal of the American Statistical Association*, 29, 1-11.

12. ANGRIST, J. D., ET KRUEGER, A. B. (1991): "Does compulsory schooling attendance affect schooling and earnings?," *The Quarterly Journal of Economics*, 106, 976-1014.

13. — (2001): "Instrumental variables and the search for identification: from supply and demand to Natural Experiments," *Journal of Economic Perspectives*, 15, 69-85.

14. ANGRIST, J. D., ET PISCHKE, J.-S. (2009): *Mostly Harmless Econometrics - An Empiricist's Companion*. Princeton, New Jersey: Princeton University Press.

15. — (2014): *Mastering 'Metrics: The Path from Cause to Effect*. : Princeton University Press.

16. — (2015): *Mastering 'Metrics: The Path from Cause to Effect*. : Princeton University Press.

17. ATHEY, S., ET IMBENS, G. W. (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685-725.

18. BAI, J. (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77, 1229-1279.

19. BALTAGI, B. (2005): *Econometric Analysis of Panel Data*. : John Wiley & Sones.

20. — (2021): *Econometric Analysis of Panel Data*. : Springer.

21. BALTAGI, B. H., ET GRIFFIN, J. M. (1983): "Gasoline demand in the OECD: an application of pooling and testing procedures," *European Economic Review*, 22, 117-137.

22. BARHOUMI, K., DARNÉ, O., ET FERRARA, L. (2012): "Une revue de la littérature des modèles à facteur dynamiques," *Economie et Prévision*, 199, 51-77.

23. BAUDRY, M. (2022): "Prévisions des effectifs dans l'enseignement supérieur – Rentrées 2022 et 2023," 2 pp.

24. BECKER, S. O., ET ICHINO, A. (2002): "Estimation of average treatment effects based on propensity scores," *The Stata Journal*, 2, 358-377.

25. BENKIMOUN, P. (2016): "L'effet cancérogène du café ou du maté n'est pas prouvé," Paris, 1.

26. BERTRAND, M., DUFLO, E., ET MULLAINATHAN, S. (2004): "How much should we trust differences-in-differences estimates?," *The Quarterly Journal of Economics*, 119, 249-275.

27. BERTRAND, M., ET MULLAINATHAN, S. (2004): "Are emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, 94, 991-1013.

28. BIA, M., ET MATTEI, A. (2008): "A Stata package for the estimation of the dose-response function through adjustment for the generalized propensity score," *The Stata Journal*, 8, 354-373.

29. — (2012): "Assessing the effect of the amount of financial aids to Piedmont firms using the generalized propensity score," *Statistical Methods & Applications*, 21, 485-516.

30. BILLMEIER, A., ET NANNICINI, T. (2013): "Assessing economic liberalization episodes: a synthetic control approach," *The Review of Economics and Statistics*, 95, 983-1001.

31. BONO, P.-H., DEBU, S., DESPLATZ, R., HAYET, M., LACOUETTE-FOUGÈRE, C., ET TRANNOY, A. (2018): "Vingt ans d'évaluations d'impact en France et à l'étranger. Analyse quantitative de la production scientifique," 56 pp.

32. BOTOSARU, I., ET FERMAN, B. (2019): "On the role of covariates in the synthetic control method," *Econometrics Journal*, 22, 117-130.

33. BOX, G., CONNOR, L., COUSINS, W., DAVIES, O., HIMSWORTH, F., ET SILLITTO, G. (1978): *The Design and Analysis of Industrial Experiments*. New York: Longman group ltd.

34. BOZIO, A., COTTET, S., ET PY, L. (2019): "Evaluation d'impact de la réforme 2008 du crédit d'impôt recherche," : IPP, 87 pp.

35. BOZIO, A., IRAC, D., ET PY, L. (2014): "Impact of research tax credit on R&D and innovation: evidence from the 2008 French reform," : Banque de France.

36. BUNEL, S., ET SICSIC, M. (2024): "Les incitations fiscales à la recherche et développement et à l'innovation : état des lieux, effets et alternatives," 29.

37. CACOULLOS, T. (1965): "A relation between t and F-distributions," *Journal of the American Statistical Association*, 60, 528-531.

38. CAE (2013): "Evaluation des politiques publiques," *Les notes du conseil d'analyse économique*, 1, 1-12.

39. CAMPOS, N. F., CORICELLI, F., ET MORETTI, L. (2019): "Institutional integration and economic growth in Europe," *Journal of Monetary Economics*, 103, 88-104.

40. CARD, D., ET KRUEGER, A. B. (1994): "Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772-784.

41. — (2000): "Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania: reply," *American Economic Review*, 90, 1397-1420.

42. CERULLI, G. (2015): *Econometric Evaluation of Socio-Economic Programs: Theory and Applications*. Berlin Heidelberg: Springer.

43. CONSEIL-DÉPARTEMENTAL (2016): "Evaluation de l'accompagnement des allocataires du RSA par les PVD," Département de la Seine-Saint-Denis: Conseil Départemental.

44. DEHEJIA, R., ET WAHBA, S. (2002): "Propensity score-matching methods for nonexperimental causal studies," *The Review of Economics and Statistics*, 84, 151-161.

45. DUNNING (2012): *Natural Experiments in the Social Sciences*. : Cambridge University Press.

46. EFRON, B., ET HASTIE, T. (2016): *Computer Age Statistical Inference - Algorithms, Evidence, and Data Science*. UK: Cambridge University Press.

47. ENGLE, R. F., HENDRY, D. F., ET RICHARD, J.-F. (1983): "Exogeneity," *Econometrica*, 51, 277-304.

48. EUROPEAN COMMISSION (2016): "The 2016 EU Industrial R&D Investment Scoreboard."

49. FLORENS, J.-P., HECKMAN, J. J., MEGHIR, C., ET VYTLACIL, E. (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76, 1191-1206.

50. GASPAROTTI, A., ET KULLAS, M. (2019): "20 years of the euro: winners and losers," 20.

51. GELMAN, A., ET HILL, J. (2006): *Data Analysis Using Regression and Multilevel/Hierarchical Models*. : Cambridge University Press.

52. — (2007): *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

53. GLIMCHER, P. W., ET FEHR, E. (2014): "Neuroeconomics: decision making and the brain," : Academic Press.

54. GOBILLON, L., ET MAGNAC, T. (2016): "Regional policy evaluation: interactive fixed effects and synthetic controls [+ supplementary + R replication data]," *The Review of Economics and Statistics*, 98, 535-551.

55. GRANGER, C. W. J. (1986): "Statistics and causal inference: comment," *Journal of the American Statistical Association*, 81, 967-968.

56. GRANGER, C. W. J., ET NEWBOLD, P. (1974): "Spurious regressions in econometrics," *Journal of Econometrics*, 2, 111-120.

57. GREENE, W. H. (2008): *Econometric Analysis*. : Pearson Prentice Hall.

58. GRIMA, P. (2013): *La certitude absolue et autres illusions*. Paris, France: RBA.

59. HAGEN, T., ET MOHL, P. (2008): "Which is the right dose of EU cohesion policy for economic growth?," : ZEW - Center for European Economic Research.

60. HAMILTON, J. (1994): *Time Series Analysis*. : Princeton University Press.

61. HARRISON, G. W., ET LIST, J. A. (2004): "Field experiments," *Journal of Economic Literature*, 42, 1009-1055.

62. HAUSMAN, J. A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251-1270.

63. HECKMAN, J. J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153-161.

64. — (1989): "Causal inference and nonrandom samples," *Journal of Educational Statistics*, 14, 159-168.
65. — (1990): "Varieties of selection bias," *American Economic Review*, 80, 313-318.
66. — (1996): "Randomization as an instrumental variable," *The Review of Economics and Statistics*, 78, 336-341.
67. HECKMAN, J. J. (2005): "The scientific model of causality," *Sociological Methodology*, 35, 1-97.
68. HECKMAN, J. J. (2020): "Randomization and social policy evaluation revisited," in *Randomized controlled trials in the field of development: a critical perspective*, ed. by F. Bédécarrats, I. Guérin, and F. Roubaud. Oxford: Oxford University Press, 46 pp.
69. HECKMAN, J. J., LOCHNER, L., ET TABER, C. (1998): "General-equilibrium treatment effects: a study of tuition policy," *American Economic Review: Papers and Proceedings*, 88, 381-386.
70. HIRANO, K., ET IMBENS, G. W. (2004): "The propensity score with continuous treatments," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, ed. by A. Gelman, and X.-L. Meng. Chichester, UK: John Wiley & Sons Ltd, 73-84.
71. HOLLAND, P. W. (1986): "Statistics and causal inference," *Journal of the American Statistical Association*, 81, 945-960.
72. IMBENS, G. W. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, 706-710.
73. — (2004): "Nonparametric estimation of average treatment effects under exogeneity: a review," *The Review of Economics and Statistics*, 86, 4-29.
74. — (2015): "Matching methods in practice - Three examples," *The Journal of Human Resources*, 50, 373-419.
75. IMBENS, G. W., ET ANGRIST, J. D. (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467-475.
76. IMBENS, G. W., ET KOLESAR, M. (2016): "Robust standard errors in small samples: some practical advice," *The Review of Economics and Statistics*, 98, 701-712.
77. IMBENS, G. W., ET RUBIN, D. B. (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge, USA: Cambridge University Press.
78. IMBENS, G. W., ET WOOLDRIDGE, J. M. (2009): "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47, 5-86.
79. KAHNEMAN, D. (2011): *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.
80. KHANDKER, S. R. (2005): "Microfinance and poverty: evidence using panel data from Bangladesh," *The World Bank Economic Review*, 19, 263-286.
81. KHANDKER, S. R., KOOLWAL, G. B., ET HUSSAIN, A. S. (2010): "Handbook on Impact Evaluation: Quantitative Methods and Practices," Washington, D.C.: The World Bank, 262.
82. KIEL, K. A., ET MCCLAIN, K. T. (1995): "The effect of an incinerator siting on housing appreciation rates," *Journal of Urban Economics*, 37, 311-323.
83. KLEIN, E. (2018): *Matière à contredire*. : Editions de l'Observatoire.
84. — (2022): "Comment construire une théorie de la gravitation en faisant chuter des corps ?," Sorbonne Université, Pierre et Marie Curie.
85. KRUEGER, A. (1999): "Experimental estimates of education production functions," *The Quarterly Journal of Economics*, 114, 497-532.
86. KRUEGER, A. B. (1999): "Experimental estimates of education production functions," *The Quarterly Journal of Economics*, 114, 497-532.
87. LALONDE, R. J. (1986): "Evaluating the econometric evaluations of training programs with experimental data," *American Economic Review*, 76, 604-620.
88. LECHNER, M. (2011): "The relation of different concepts of causality used in time series and microeconometrics," *Econometric Reviews*, 30, 109-127.
89. LEE, M.-J. (2016): *Matching, Regression Discontinuity, Difference in Differences and Beyond*. : Oxford University Press.
90. LEUVEN, E., ET SIANESI, B. (2003): "PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing," : ideas.

91. LIST, J. A. (2009): "An introduction to field experiments in economics," *Journal of Economic Behavior & Organization*, 70, 439-442.

92. LORD, F. M. (1967): "A paradox in the interpretation of group comparisons," *Psychological Bulletin*, 65, 304-305.

93. MAGNUS, J. R., ET NEUDECKER, H. (1988): *Matrix differential calculus with applications in statistics and econometrics*. Chippenham, Wiltshire: John Wiley & Sons.

94. MARINO, M., LHUILLERY, S., PARROTTA, P., ET SALA, D. (2016): "Additionality or crowding-out? An overall evaluation of public R&D subsidy on private R&D expenditure," *Research Policy*, 45, 1715-1730.

95. MENESR (2010): "Crédit d'impôt recherche : chiffres 2008 et évolutions récentes."

96. MERCER, W. B., ET HALL, A. D. (1911): "The experimental error of field trials," *The Journal of Agricultural Science*, 4, 107-132.

97. MEYER, B. D. (1995): "Natural and quasi-experiments in economics," *Journal of Business & Economic Statistics*, 13, 151-160.

98. MOCZALL, A. (2014): "Effets d'aubaine et de substitution d'un dispositif allemand de subvention salariale pour demandeurs d'emploi difficielement employables," *Travail et emploi*, 139, 39-59.

99. MOREL-À-L'HUISSIER, P., ET PETIT, V. (2018): "Évaluation des dispositifs d'évaluation des politiques publiques - Rapport d'information du comité d'évaluation et de contrôle des politiques publiques."

100. MORGAN, S. L., ET WINSHIP, C. (2007): *Counterfactuals and Causal Inference - Methods and Principles for Social Research*. USA, New York: Cambridge University Press.

101. NELSON, D. (2004): *The Penguin Dictionary of Statistics*. : Penguin Books.

102. NEWBOLD, P., CARLSON, W. L., ET BETTY, T. (2007): *Statistics for Business and Economics*. New Jersey: Pearson Prentice Hall.

103. NICHOLS, A. (2007): "Causal inference with observational data," *The Stata Journal*, 7, 507-541.

104. — (2008): "Erratum and discussion of propensity-score reweighting," *The Stata Journal*, 8, 532-539.

105. PEARL, J., GLYMOUR, M., ET JEWELL, N. P. (2016): *Causal Inference in Statistics: a Primer*. : Wiley.

106. PELIKS, G. (2014): "In data veritas, et le Big data dans tout ça ?," *Tangente Sup*, 77-78, 52-56.

107. QUANDT, R. E. (1972): "A new approach to estimating switching regressions," *Journal of the American Statistical Association*, 67, 306-310.

108. ROSENBAUM, P. R. (2010): *Observational Studies*. USA, New York: Springer-Verlag.

109. ROSENBAUM, P. R., ET RUBIN, D. B. (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41-55.

110. RUBIN, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688-701.

111. — (1977): "Assignment to treatment group on the basis of a covariate," *Journal of Educational Statistics*, 2, 1-26.

112. — (2004): "Teaching statistical inference for causal effects in experiments and observational studies," *Journal of Educational and Behavioral Statistics*, 29, 343-367.

113. — (2005): "Basic conepts of statistical inference for causal effects in experiments and observational studies," Harvard University.

114. SALSBURG, D. (2002): *The Lady Tasting Tea*. : Holt.

115. SFE (2006): "Charte de l'évaluation des politiques publiques et des programmes publics," Paris: Société Française de l'Evaluation.

116. SHADISH, W. R., COOK, T. D., ET CAMPBELL, D. T. (2002): *Experimental and quasi-experimental designs for generalized causal inference*. : Cengage Learning.

117. TVERSKY, A., ET KAHNEMAN, D. (1981): "The framing of decisions and the psychology of choice," *Science*, 211, 453-458.

118. VANDENBUSSCHE, H., ET VIEGELAHN, C. (2015): "Trade protection and input switching/Input reallocation within firms," 1-60.

119. WASMER, E. (2010): *Principes de Microéconomie: Méthodes Empiriques et Théories Modernes*.

120. WING, C., SIMON, K., ET BELLO-GOMEZ, R. A. (2018): "Designing difference-in-difference studies: best practices for public health policy research," *Annual Review of Public Health*, 39, 453-469.

121. WOOLDRIDGE, J. M. (2003): *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press.
122. — (2005): "Violating ignorability of treatment by controlling for too many factors," *Econometric Theory*, 21, 1026-1028.
123. — (2009): *Introductory Econometrics: A Modern Approach*. Cambridge, Massachusetts: The MIT Press.
124. — (2009): *Introductory Econometrics: A Modern Approach (Limited Dependent Variable)*. Cambridge, Massachusetts: The MIT Press.
125. — (2009): *Introductory Econometrics: A Modern Approach: Ch. 17*. Cambridge, Massachusetts: The MIT Press.
126. — (2010): *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press.
127. YULE, G. U. (1925): "Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series."
128. ZUBIZARRETA, J. R., SMALL, D. S., ET ROSENBAUM, P. R. (2014): "Isolation in the construction of natural experiments," *Annals of Applied Statistics*, 8, 2096-2121.