Université de Montréal

Méthode d'inférence par bootstrap pour l'estimateur sisVIVE en randomisation mendélienne

par

Tatiana Dessy

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en statistique

28 novembre 2018

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé

Méthode d'inférence par bootstrap pour l'estimateur sisVIVE en randomisation mendélienne

présenté par

Tatiana Dessy

a été évalué par un jury composé des personnes suivantes :

Pierre Duchesne
(président-rapporteur)

Marie-Pierre Sylvestre
(directeur de recherche)

Christian Léger
(codirecteur)

Martin Bilodeau
(membre du jury)

Mémoire accepté le
27 novembre 2018

SOMMAIRE

Dans le contexte des études observationnelles, l'estimation de l'effet d'une exposition sur une issue doit gérer l'effet de variables de confusion non mesurées qui affectent à la fois l'exposition et l'issue, sans quoi l'estimation de l'effet causal sera biaisée. En réponse à ce problème de biais de confusion, la discipline de l'économétrie a développé la méthode des variables instrumentales. Cette dernière permet d'inférer un effet causal lorsqu'il y a de la confusion en utilisant des variables ayant la propriété d'être « aléatoires » dans le modèle d'estimation, en plus d'être fortement associées à l'exposition. Ainsi peuvent-elles être utilisées comme des instruments pour randomiser l'exposition. Le contexte d'estimation devient alors comparable à celui d'un essai randomisé, reconnu comme le *gold standard* pour inférer la causalité. La solution se traduit en génétique par la technique de la randomisation mendélienne. Cette technique tire avantage de l'allocation aléatoire des allèles à la naissance et utilise des polymorphismes nucléotidiques (SNPs) comme instruments pour obtenir un phénotype exposition « quasi randomisé ». Par contre, les SNPs doivent satisfaire certaines suppositions de validité qui ne sont pas toutes vérifiables dans un jeu de données. Avec des instruments multiples, l'estimation de l'effet se fait fréquemment par la méthode des moindres carrés en deux étapes. Or, cette méthode suppose que tous les SNPs sont valides, alors qu'il est probable qu'en réalité certains SNPs soient invalides : par exemple, dû au phénomène de la pléiotropie. Pour tenir compte de la possibilité que certains SNPs soient invalides, Kang et al. (2014) ont proposé l'estimateur sisVIVE, qui gère l'effet de l'invalidité des SNPs jusqu'à un seuil de 50% de SNPs invalides et offre une estimation ponctuelle de l'effet causal. Nous contribuons à la littérature de sisVIVE en explorant une méthode bootstrap pour construire des intervalles de confiance. Les résultats de l'étude obtenus par simulations ainsi qu'une application à une base de données de la Biobanque de l'Institut de Montréal seront présentés.

Mots-clés : inférence causale, génétique, biais de confusion, pléiotropie, randomisation mendélienne, variable instrumentale, sisVIVE, bootstrap, intervalle de confiance, Biobanque.

SUMMARY

In the observational data framework, estimation must account for unmeasured confounders, which affect both exposure and outcome. The reason for this is that if the effect of confounders is not considered, the causal effect estimate will suffer from confounding bias. The econometrics literature addresses this problem by using the instrumental variables method, which enables causal inference when confounding is present by using variables which have the property of being "random" in the posited model and strongly associated with the exposure. Hence, they can serve as instruments to randomize exposure, thus mimicking the setting of randomized control trials—the gold standard to establish causality. With genetic data, the issue can be approached similarly using Mendelian randomization. Mendelian randomization studies apply the instrumental variables approach to genetics by exploiting the natural randomness of allele allocation and using single-nucleotide polymorphisms (SNPs) as instrumental variables to obtain a "quasi-randomized" exposure phenotype. However, the choice of SNPs is not arbitrary ; instead, it relies on a set of assumptions for validity, some of which cannot be verified in most empirical settings. When using multiple SNPs as instruments, one might be tempted to use the widely known two-stage least squares estimator, given its simplicity. Yet, the use of this estimator comes at the high price of assuming that all SNPs verify the set of assumptions, whereas in reality, SNPs can be invalid, for instance, due to pleiotropy. To include the possibility that some SNPs may be invalid, Kang et al. (2014) proposed the sisVIVE estimator, which mitigates the effect of invalid SNPs up to a threshold of 50% of invalid SNPs, thereby providing a point estimate of the causal effect. We add to their contribution by exploring a bootstrap method to construct confidence intervals for sisVIVE. Results obtained from simulations and an application to a real dataset from the Montreal Heart Institute's Biobank are presented.

Keywords : causal inference, genetics, confounding bias, pleiotropy, Mendelian randomization, instrumental variable, sisVIVE, bootstrap, confidence interval, Biobank.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Plus qu'un travail de formation académique, ce mémoire est le fruit d'une étape de croissance personnelle fort appréciée. Plusieurs personnes y ont participé et j'aimerais, par ce qui suit, offrir mes remerciements plus particuliers à certaines d'entre elles.

À ma directrice et à mon codirecteur, la professeure Marie-Pierre Sylvestre et le professeur Christian Léger, pour m'avoir offert l'unique opportunité de faire mes premiers pas dans l'univers de la recherche scientifique. Mes sincères remerciements à Marie-Pierre, pour m'avoir confié un projet de recherche stimulant, en plus d'être à la fois socialement et scientifiquement pertinent ; pour m'avoir soutenue et guidée face aux nombreux défis auxquels j'ai été confrontée durant toute la réalisation de ce projet ; et pour m'avoir ainsi donné la chance de croire un peu plus en moi, de viser toujours un peu plus haut. C'est une leçon énorme pour laquelle je lui serai toujours reconnaissante. Ma profonde gratitude à Christian, pour m'avoir communiqué les exigences de rigueur, de cohérence et de pertinence qui m'ont permis de développer ma capacité à rédiger des documents scientifiques ; et pour son soutien constant et sa disponibilité tout au long de l'exécution de ce projet. Tous les deux ont fait preuve d'une grande patience à mon égard et leur dévouement à ce projet m'a permis de le conduire jusqu'à la fin.

À la professeure Marie-Pierre Dubé, qui m'a généreusement accueillie au sein de son équipe de recherche et sans qui la réalisation de l'application aux données réelles dans ce projet n'aurait pas eu lieu. Dès notre première rencontre, son attention et ses conseils m'ont été des plus précieux.

Je remercie également toute l'équipe de StatGen et toutes les personnes dont j'ai fait la connaissance au Centre de Pharmacogénomique Beaulieu-Saucier, pour leur aide inoubliable.

Aux amitiés développées tout au long de mes études. Si je peux enfin leur exprimer mes sentiments, c'est qu'elles m'ont épaulée et inspirée plus que je ne pouvais m'y attendre. Simplement, merci.

Mes remerciements les plus tendres à toute ma famille, celle qui est restée près de moi lors des moments plus incertains, qui m'a insufflé toutes les qualités personnelles essentielles à la complétion de ce projet et la continuation vers bien d'autres.

ex nihilo nihil fit.

INTRODUCTION

À l'issue d'une étude scientifique menée en 1950, Richard Doll et Austin Bradford Hill, deux épidémiologistes, ont abouti à la conclusion qu'en absence de toute autre explication possible, la consommation du tabac cause le cancer du poumon [**11**]. En réponse à leur conclusion, le statisticien Ronald A. Fisher a suggéré que le lien entre le tabagisme et le cancer pulmonaire n'était qu'une association, plutôt qu'un lien de causalité. Il a ensuite attribué l'association observée entre la consommation du tabac et le cancer du poumon à une cause commune aux deux, la génétique, dont les effets sur le choix de fumer et le développement d'un cancer au poumon étaient non mesurés [**15**]. Bien que nous sachions aujourd'hui que le tabagisme est une cause du cancer du poumon, la contribution de Fisher demeure pertinente, parce qu'elle illustre la question de savoir à quel moment la corrélation implique la causalité.

L'inférence causale appliquée à la médecine a évolué suite à de nombreuses discussions autour du débat association-causalité. Celles de Fisher, Doll et Bradford Hill en sont un exemple parmi d'autres. Pour les études basées sur des données observationnelles, où il n'y a aucune randomisation, l'inférence causale doit se faire à partir d'une méthodologie qui traite, entre autres, du problème de variables non mesurées qui causent à la fois l'effet étudié, c'est-à-dire, l'issue, et sa cause, l'exposition. De là, l'inférence causale en médecine a adopté la pratique d'introduire des facteurs génétiques dans l'explication des causes de maladies par la technique de la randomisation mendélienne.

En tant que méthodologie d'inférence causale basée sur l'utilisation de données observationnelles, la randomisation mendélienne permet d'estimer l'effet d'une exposition sur une issue en présence de causes communes non mesurées, en exploitant la technique de l'instrumentation [**29**]. Cette approche à l'inférence causale tire profit du fait qu'une mutation génétique présente pour un certain pourcentage d'une population donnée, soit le résultat d'un polymorphisme nucléotidique (ci-après, SNP, de l'anglais *single nucleotide polymorphism*), est corrélée avec une maladie (exposition). En effet, un SNP peut mener à l'expression de divers traits génétiques qui pourraient, à leur tour, être responsables de différentes maladies. De plus, la probabilité d'exprimer une maladie est quantifiée par

le nombre d'allèles (0, 1 ou 2) impliqués dans le SNP. Le processus génétique mendélien implique que les allèles dans une population donnée sont distribués indépendamment des nombreux facteurs qui peuvent avoir un impact à la fois sur une exposition et sur son issue [**33**], parce que l'allocation des allèles au moment de la formation des gamètes est un processus aléatoire. C'est ce que nous entendons par « randomisation mendélienne ». Cette technique permet ainsi de cadrer une étude observationnelle en un devis « quasi randomisé ».

Cependant, un problème qui persiste dans cette méthodologie est le fait que les suppositions sont difficilement vérifiables dans un vrai jeu de données. La validité d'un SNP en randomisation mendélienne implique que (A1) le SNP est associé avec l'exposition ; (A2) le SNP est indépendant de toutes causes communes (non mesurées) à l'exposition et à l'issue ; et (A3) le SNP est indépendant de l'effet si nous tenons compte de l'exposition et des causes communes non mesurées [**29**]. Ainsi, la possibilité qu'un SNP utilisé comme instrument puisse avoir un effet sur un autre trait physique ou maladie, appelée la « pléiotropie », qui affecte aussi l'issue étudiée invalide le SNP du fait de la violation de (A3). En d'autres termes, la pléiotropie rend possible le fait qu'un SNP puisse avoir un effet sur l'issue qui n'est pas médié par l'exposition. Par conséquent, la randomisation mendélienne n'est pas valide.

Par exemple, nous savons que les SNPs du gène FTO (de l'anglais, *fat mass and obesity-associated protein*) sont fortement associés à l'obésité, qui est une cause de l'hypertension [**14**]. Or, dépendamment de l'effet étudié, ces SNPs peuvent être invalides comme instruments pour l'obésité s'ils affectent l'hypertension autrement que par leur effets sur l'obésité. Ainsi, FTO est un gène souvent utilisé en randomisation mendélienne dans l'étude de l'obésité, mais dont les connaissances génétiques ne sont pas assez étendues pour déterminer la pléiotropie de ses SNPs. Outre les SNPs de FTO, il existe d'autres SNPs qui sont corrélés avec l'obésité et tous ces SNPs peuvent être utilisés dans une même analyse en randomisation mendélienne pour avoir une meilleure prévision de l'obésité. Il faut alors avoir recours à des méthodes d'estimation dîtes à « instruments multiples » pour combiner les effets des SNPs.

Un des estimateurs à instruments multiples les plus conventionnellement utilisés est l'estimateur des moindres carrés en deux étapes ou *two-stage least squares* en anglais (ci-après, 2SLS), d'abord proposé en 1953 par l'économiste Henri Theil [**37**]. L'estimateur 2SLS est simple à mettre en œuvre, puisque la méthode fait essentiellement une série de deux estimations par les moindres carrés ordinaires. Une limite importante de cet estimateur est qu'il s'appuie sur l'hypothèse que les suppositions (A1)-(A3) de la validité des instruments sont satisfaites pour tous les SNPs, sans quoi l'estimation de l'effet sera biaisée. D'où l'intérêt singularisé pour les estimateurs qui sont plus robustes à l'utilisation

d'instruments possiblement invalides, tels que des SNPs pléiotropiques en randomisation mendélienne.

Parmi les estimateurs à instruments multiples capables d'estimer l'effet causal d'une exposition sur une issue en présence de SNPs possiblement pléiotropiques se trouve l'estimateur sisVIVE (*some invalid some valid instrumental variables estimator*), développé par Kang et al. (2014). Une étude par des simulations démontre qu'il est possible d'identifier et d'estimer un effet causal avec sisVIVE lorsque plus de la moitié des SNPs sont des instruments valides, sans nécessairement savoir lesquels [**23**]. Toutefois, aucune façon de calculer la variance ou les intervalles de confiance de l'estimateur sisVIVE n'y a été proposée.

Dans ce mémoire, nous explorons une méthodologie bootstrap simple pour calculer des intervalles de confiance pour l'effet causal estimé par sisVIVE et ce dans le contexte de la randomisation mendélienne. Les méthodologies de rééchantillonnage telles que le bootstrap (p.e. [**13**]) sont généralement utilisées du fait qu'elles sont facilement mises en œuvre sans avoir à connaître la distribution asymptotique de l'estimateur et, en particulier, d'avoir une forme analytique pour sa variance.

À l'aide d'une analyse par simulations, nous verrons qu'il est possible de calculer des intervalles de confiance bootstrap de niveau 95% pour l'estimation de sisVIVE. Nous calculons trois types d'intervalles : l'intervalle percentile, l'intervalle de base et l'intervalle basé sur la loi normale. Nous les étudions en présence de SNPs invalides en raison d'un effet du SNP sur l'issue qui n'est pas médié par l'exposition, soit une violation de (A3), et de différentes tailles d'échantillon. Pour illustrer notre méthodologie, nous faisons une application à un jeu de données réelles où nous analysons l'effet de l'obésité sur la pression artérielle avec les données de la Biobanque de l'Institut de Cardiologie de Montréal.

Ce mémoire est structuré comme suit. Le chapitre 1 présente certains concepts de base en génétique et une explication mathématique de l'effet de l'absence de la randomisation dans l'estimation d'un effet causal à partir d'études observationnelles. Le chapitre 2 présente la méthodologie de la randomisation mendélienne et les suppositions qu'un SNP doit satisfaire pour être considéré comme un instrument valide. Le chapitre 3 présente deux estimateurs adaptés à la randomisation mendélienne : la méthode conventionnelle 2SLS et la méthode sisVIVE. Le chapitre 4 présente la méthodologie du bootstrap et notre algorithme bootstrap simple pour calculer des intervalles de confiance de niveau 95% pour sisVIVE. Le chapitre 5 présente les simulations. Le chapitre 6 présente l'application au jeu de données réelles. Le chapitre 7 revient sur l'objectif du mémoire et en fait la conclusion.

Chapitre 1

INFÉRENCE CAUSALE ET CONCEPTS GÉNÉTIQUES

Dans les études observationnelles étudiant le lien entre deux phénotypes, les méthodes statistiques actuelles ont tendance à modéliser un effet d'association statistique : par exemple, l'association entre l'obésité et la pression artérielle. Pourtant, si une étude a pour fin d'établir une intervention pour prévenir une maladie ou promouvoir un comportement bénéfique pour la santé dans une population, ce qui est souhaitable est plutôt de déterminer si l'exposition cause l'issue, et donc de pouvoir identifier un effet causal entre elles.

Afin d'identifier un effet causal, les chercheurs et chercheuses ont recours, dans la mesure du possible, aux essais randomisés, parce qu'ils font usage de la randomisation et que celle-ci, lorsque parfaite, élimine le biais de confusion, c'est-à-dire, le biais qui résulte de l'effet de variables qui ne sont pas contrôlées mais qui agissent à la fois sur l'exposition et l'issue, influençant ainsi le vrai effet causal entre ces dernières. Or, les essais randomisés s'avèrent parfois impossibles à mettre en œuvre pour des raisons pragmatiques, éthiques et/ou économiques, ce qui explique pourquoi les études observationnelles sont utilisées. Par contre, les études observationnelles sont plus susceptibles de mener à un effet causal qui souffre d'un biais de confusion, parce qu'il peut arriver qu'on soit confronté à des situations où l'exposition et l'issue sont expliquées par une même variable qui est inconnue ou mal mesurée dans l'étude.

Dans ce premier chapitre, nous faisons d'abord la distinction entre la modélisation d'un effet causal et celle d'association statistique. Puis, nous présentons la problématique du biais de confusion dans l'estimation d'un effet causal. Enfin, nous définissons des concepts génétiques de base qui seront nécessaires à la compréhension de la randomisation mendélienne, qui est la solution au biais de confusion que ce mémoire a pour but d'étudier.

1.1. Association et causalité

Dans cette section, nous abordons le concept de causalité. D'abord, nous soulignons la distinction entre le concept d'association et le concept de causalité en statistique. Ensuite, nous décrivons deux devis d'étude : les essais randomisés, qui se prêtent plus naturellement à l'inférence causale, et les études observationnelles, qui nécessitent plus d'attention pour pouvoir inférer l'effet causal. Puis, nous présentons un outil de modélisation en inférence causale, soit le graphe orienté acyclique. À titre d'exemple, nous présentons l'étude de l'effet causal de l'obésité (mesurée par l'indice de masse corporelle) sur la pression artérielle.

1.1.1. Distinction entre association et causalité

La causalité peut se comprendre comme étant la réponse de l'issue suite à un changement dans l'exposition, et sa mesure est invariante aux changements qui peuvent s'opérer dans d'autres variables qui auraient une influence sur l'exposition et l'issue.

Lorsque nous voulons étudier la possibilité qu'une variable d'issue réponde aux changements dans une variable d'exposition, nous avons habituellement recours aux techniques d'association statistique, par exemple, la régression linéaire, qui se basent sur la distribution conjointe des variables. Dans un environnement contrôlé, nous pouvons clairement identifier et mesurer les variables qui entrent en jeu dans la structure de cause à effet pour une exposition et une issue donnée. Dans ce cas, nous savons quelles variables inclure ou exclure du modèle d'estimation. L'effet d'association estimé est alors bel et bien l'effet causal, parce que nous avons une connaissance parfaite de la structure de cause à effet qui a mené à l'exposition et l'issue, et le cadre d'interprétation de l'effet d'association, soit « toutes choses étant égales par ailleurs », coïncide avec celle-ci. Autrement, dans un environnement non contrôlé, nous ne connaissons habituellement pas la véritable structure de cause à effet, donc certaines variables peuvent agir de sorte à ce que l'association observée entre l'exposition et l'issue soit conditionnelle à l'influence de ces variables. Lorsque ces variables ne sont pas incluses dans le modèle d'estimation, l'effet estimé variera selon les changements dans ces variables, ce qui n'est pas un effet causal.

Ainsi, un effet causal ne peut pas être défini par une distribution conjointe seule, mais nécessite des suppositions qui identifient les relations qui demeurent invariantes lorsque les conditions expérimentales changent.

Pour illustrer la différence entre le concept d'association et le concept de causalité, considérons une régression linéaire multiple, où l'issue Y est régressée sur l'exposition D afin d'estimer l'effet d'intérêt. En principe, nous ajustons pour les variables que nous croyons être en mesure de changer l'effet d'intérêt en les incluant comme régresseurs C :

$$Y = D\beta + C\gamma + \epsilon \tag{1.1.1}$$

où β est l'effet de D sur Y et γ est l'effet d'une covariable C sur Y. Le modèle de régression (1.1.1) nous permet de savoir si l'effet de l'exposition est statistiquement significatif avec ou sans ajustement pour des covariables, mais ne nous révèle rien au sujet de la structure de cause à effet entre les variables. En particulier, nous pouvons inférer que β est causal pour une valeur fixe γ et en supposant qu'aucune autre variable n'a un effet sur l'issue.

1.1.2. Essais randomisés et études observationnelles

Nous procédons maintenant à la description des devis d'étude qui seront utiles à la compréhension du concept de confusion et du biais qui s'en suit.

1.1.2.1. *Essai randomisé*

Nous avons mentionné précédemment que l'effet causal d'une exposition sur une issue peut être estimé par un modèle d'association statistique si les conditions expérimentales sont parfaitement connues. Un devis qui satisfait cette condition est l'essai randomisé. Dans un essai randomisé, les participants de l'étude sont aléatoirement assignés à l'un de deux ou plusieurs groupes d'exposition. En d'autres termes, l'exposition est « randomisée » et ainsi toute autre variable qui ne fait pas partie des variables de l'étude est en moyenne distribuée de façon équilibrée dans les groupes d'exposition. Par conséquent, l'association observée entre l'exposition et l'issue est indépendante de ces variables.

Une fois l'exposition assignée, les données expérimentales recueillies sont analysées à l'aide d'un modèle d'association statistique approprié pour parvenir à l'inférence de l'effet causal de l'exposition sur l'issue.

1.1.2.2. *Étude observationnelle*

Le devis observationnel est, comme son nom l'indique, basé sur des données observées où l'exposition n'est pas contrôlée. Les participants de l'étude forment deux ou plusieurs groupes d'exposition, mais celle-ci n'est pas randomisée. Il est possible que certaines variables non mesurées dans l'étude influencent à la fois l'exposition et l'issue. Par conséquent, l'association observée entre l'exposition et l'issue dépendra de ces variables et ne correspondra donc pas à l'effet causal. Ceci est la difficultée inhérente d'estimer un effet causal dans une étude observationnelle : si ces variables sont inconnues ou mal mesurées et donc impossibles à inclure dans le modèle d'estimation, elles vont biaiser l'estimation de l'effet causal à l'insu des chercheurs et/ou des chercheuses.

1.1.3. Graphe orienté acyclique

Nous pouvons constater qu'il est important de bien conceptualiser les relations influençant l'association entre l'exposition et l'issue pour s'assurer de contrôler le mieux que possible l'influence d'autres variables sur l'effet de l'exposition sur l'issue et inférer la causalité. Un outil permettant de représenter la structure de cause à effet entre les variables, qui représente le vrai modèle, est le graphe orienté acyclique (en anglais, *directed acyclic graph* ou DAG). Le DAG illustre toutes les suppositions faites à propos des liens de causalité entre les variables d'exposition et d'issue. Ainsi, il contient non seulement l'exposition et l'issue, mais également les variables qui peuvent les influencer, l'inclusion desquelles est faite à la connaissance des chercheurs et/ou des chercheuses.

L'interprétation d'un DAG se fait comme suit. D'abord, une flèche entre deux variables implique qu'il existe un lien de cause à effet entre elles, la pointe de la flèche indiquant la variable d'effet. Puis, à l'inverse, une absence de flèche entre deux variables implique qu'il n'y a pas de lien causal. Notons que ceci n'empêche pas qu'il puisse exister une association entre deux variables et donc que l'absence d'une flèche n'implique pas l'indépendance, mais plutôt l'indépendance conditionnelle.

1.1.4. Exemple

L'exemple suivant est tiré d'une étude observationnelle sur la santé des femmes [**12**]. Nous voulons étudier l'effet entre l'obésité (l'exposition, mesurée par l'indice de masse corporelle, ci-après IMC) et la pression artérielle (ci-après PA, soit l'issue). Nous disposons également d'une mesure du statut socio-économique (SSE). Les informations contenues dans la littérature nous ont permis de construire le DAG suivant :

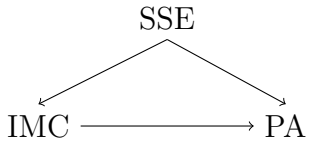


FIGURE 1.1. DAG représentant la structure de cause à effet pour l'étude de l'effet de l'obésité (mesurée par l'IMC) sur la pression artérielle (PA).

Nous interprétons la figure 1.1 de la façon suivante. D'abord, la pression artérielle de la femme est causée par son IMC et son statut socio-économique. Ensuite, l'IMC de la femme est causé par son statut socio-économique. Ainsi, nous pouvons constater que le SSE est un facteur qui cause à la fois l'IMC et la PA. Par conséquent, l'effet d'association observé entre les données de l'IMC et celles de la PA incorpore l'effet du SSE et donc est

différent de l'effet causal de l'IMC sur la PA. Notons que ce DAG illustre volontairement un cas simple, mais qu'en réalité, il comporterait davantage de variables.

1.2. Biais de confusion

Dans cette section, nous formalisons les concepts de cause commune et de biais de confusion. Ce dernier sera une motivation principale pour la technique de la randomisation mendélienne.

1.2.1. Cause commune, confusion et randomisation

Nous entendons par cause commune tout facteur qui influence deux variables à la fois. Par exemple, à la figure 1.1, le statut socio-économique de l'individu est une cause commune à son IMC et sa pression artérielle. Les causes communes sont d'une grande importance dans la modélisation de l'effet causal d'une exposition sur une issue. En effet, s'il existe une cause commune à ces deux variables, alors l'association observée entre elles réflète l'effet causal, mais aussi l'effet de la cause commune. Par exemple, si la cause commune a un effet positif sur l'exposition et l'issue, alors l'association observée sera plus grande que l'effet causal [1], reflétant l'effet de la cause commune. La différence entre l'effet d'association observé et le vrai effet causal est dû à la confusion de l'effet par la cause commune. Cette différence est appelée biais de confusion.

Afin de souligner l'importance du biais de confusion, considérons les deux DAGs suivants qui illustrent les situations où l'effet causal est confondu par l'influence d'une cause commune à l'exposition et l'issue. Dans les exemples qui suivent, nous supposons que toutes ces variables sont parfaitement mesurées et que les effets sont tous positifs.

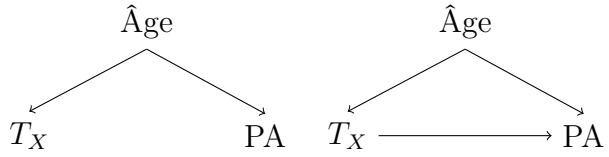


Figure 1.2. Structures de base illustrant des situations où il existe un biais de confusion potentiel dans l'estimation de l'effet causal du traitement T_X sur la pression artérielle.

Supposons que le traitement T_X est un médicament ayant pour effet secondaire d'augmenter la pression artérielle. Selon les DAGs à la figure 1.2, l'âge de l'individu est la seule cause commune au traitement T_X et à la pression artérielle de l'individu. Plus les individus sont âgés, plus ils sont susceptibles de recevoir le traitement. De plus, la pression artérielle

1. supposant que l'effet causal est positif

a tendance à augmenter naturellement avec l'âge. Dans la situation de gauche, tout effet observé entre T_X et PA est entièrement dû à l'influence de l'âge sur chacune des variables : l'association mesurée entre T_X et PA en ignorant l'âge peut alors être qualifiée d'artificielle. Dans la situation de droite, soit le cas où il existe réellement un effet du traitement sur la pression artérielle, celui-ci est confondu par l'effet de l'âge. Ceci implique que l'effet observé est plus grand que le vrai effet causal, parce que l'âge a un effet positif sur le traitement T_X et la pression artérielle.

Nous allons présenter le problème de la confusion à l'aide des deux devis expliqués précédemment, soit l'essai randomisé et l'étude observationnelle. Rappelons qu'une différence cruciale entre ces deux devis est que l'exposition est randomisée dans le premier cas, mais ne l'est pas dans le deuxième. Ceci implique que l'exposition dans une étude observationnelle ne sera pas assignée lors du devis, donc l'effet de l'exposition sur l'issue dépendra de plusieurs causes communes hors du contrôle de l'investigateur ou de l'investigatrice. Or, dans le cas randomisé, nous avons indépendance de l'effet. Notre explication de la confusion est inspirée de [**8**].

Supposons que le modèle de régression employé pour estimer l'effet du traitement T_X sur la pression artérielle soit :

$$PA = \alpha + \beta_1 Age + \beta_2 T_X + \epsilon. \tag{1.2.1}$$

Alors l'espérance conditionnelle de PA étant donné le traitement T_X et l'âge est

$$E[PA|T_X,\, Age] = \alpha + \beta_1 Age + \beta_2 T_X. \tag{1.2.2}$$

Nous allons démontrer l'effet de la randomisation du traitement T_X dans l'estimation de l'effet causal β_2. Pour simplifier le problème, nous regroupons les individus en deux groupes d'âge, soit inférieur à 55 ans (Age = 0) et supérieur ou égal à 55 ans (Age = 1), et deux groupes d'exposition, soit témoin ($T_X = 0$) et traitement ($T_X = 1$). Notons la probabilité de l'âge avancé par $p_{age} = P(Age = 1)$ et la probabilité de recevoir le traitement T_X par $p_{T_X} = P(T_X = 1)$. Le tableau 1.1 illustre les valeurs possibles de l'espérance conditionnelle (1.2.2).

Âge	T_X	$E[PA\|T_X,\, Age]$	$P(Age = a,\, T_X = x)$
0	0	α	p_{00}
0	1	$\alpha + \beta_2$	p_{01}
1	0	$\alpha + \beta_1$	p_{10}
1	1	$\alpha + \beta_1 + \beta_2$	p_{11}

TABLEAU 1.1. Valeurs possibles de l'espérance conditionnelle de PA étant donné le traitement T_X et l'âge.

Définissons l'effet d'association entre T_X et PA par la différence des espérances conditionnelles,

$$\beta_2 = E[PA|T_X = 1] - E[PA|T_X = 0]. \tag{1.2.3}$$

À partir du tableau 1.1, nous pouvons calculer ces espérances comme suit :

$$\begin{aligned} E[PA|T_X = 0] &= \frac{\sum_a E[PA|T_X = 0, Age = a]P(T_X = 0,\ Age = a)}{P(T_X = 0)} \\ &= \frac{\alpha p_{00} + (\alpha + \beta_1)p_{10}}{p_{00} + p_{10}} \\ &= \alpha + \beta_1\Big(\frac{p_{10}}{p_{00} + p_{10}}\Big) \end{aligned} \tag{1.2.4}$$

pour le groupe témoin, et

$$\begin{aligned} E[PA|T_X = 1] &= \frac{\sum_a E[PA|T_X = 1,\ Age = a]P(T_X = 1,\ Age = a)}{P(T_X = 1)} \\ &= \frac{(\alpha + \beta_2)p_{01} + (\alpha + \beta_1 + \beta_2)p_{11}}{p_{01} + p_{11}} \\ &= \alpha + \beta_2 + \beta_1\Big(\frac{p_{11}}{p_{01} + p_{11}}\Big) \end{aligned} \tag{1.2.5}$$

pour le groupe traitement. Ainsi, l'effet d'association (1.2.3) est

$$\begin{aligned} \beta_2 &= E[PA|T_X = 1] - E[PA|T_X = 0] \\ &= \Big[\alpha + \beta_2 + \beta_1\Big(\frac{p_{11}}{p_{01} + p_{11}}\Big)\Big] - \Big[\alpha + \beta_1\Big(\frac{p_{10}}{p_{00} + p_{10}}\Big)\Big] \\ &= \beta_2 + \beta_1\Big(\frac{p_{11}}{p_{01} + p_{11}} - \frac{p_{10}}{p_{00} + p_{10}}\Big). \end{aligned} \tag{1.2.6}$$

Si le traitement T_X est randomisé, alors nous avons que T_X est indépendant de l'âge. Ceci implique que les probabilités au tableau 1.1 peuvent s'écrire,

$$\begin{aligned} p_{00} &= (1 - p_{age})(1 - p_{T_X}) \\ p_{01} &= (1 - p_{age})p_{T_X} \\ p_{10} &= p_{age}(1 - p_{T_X}) \\ p_{11} &= p_{age}p_{T_X} \end{aligned} \tag{1.2.7}$$

donc les fractions à l'équation (1.2.6) sont égales à

$$\begin{aligned} \frac{p_{11}}{p_{01} + p_{11}} &= \frac{p_{age}p_{T_X}}{p_{T_X}} = p_{age} \\ \frac{p_{10}}{p_{00} + p_{10}} &= \frac{p_{age}(1 - p_{T_X})}{(1 - p_{T_X})} = p_{age} \end{aligned} \tag{1.2.8}$$

et par conséquent l'effet d'association entre T_X et PA en (1.2.6) reflète uniquement l'effet de T_X. Dans le cas où T_X n'est pas randomisé, par exemple si les données sont issues d'une étude observationnelle plutôt que d'un essai randomisé, nous ne pouvons pas arriver à l'effet non-confondu par l'âge puisque les fractions en (1.2.8) ne sont pas égales dans ce cas. Donc, nous encourons un biais de confusion dans l'estimation de l'effet causal du traitement T_X sur la pression artérielle, et l'effet estimé de β_2 est sur- ou sous-estimé par $\left(\frac{p_{11}}{p_{01}+p_{11}} - \frac{p_{10}}{p_{00}+p_{10}}\right)$.

Nous allons illustrer par un exemple quantitatif l'aspect problématique dans l'estimation de l'effet du traitement T_X sur la pression artérielle, soit la dépendance (potentielle) de T_X sur l'âge. Cet exemple est inspiré de [**8**].

Nous pouvons faire face à l'un des deux cas illustrés ci-dessous.

Âge	T_X	Probabilité observée
0	0	$p_{00} = 0{,}8$
0	1	$p_{01} = 0{,}8$
1	0	$p_{10} = 0{,}2$
1	1	$p_{11} = 0{,}2$

TABLEAU 1.2. Données issues d'une étude où le traitement T_X est randomisé : l'âge est équilibré entre les groupes de T_X.

Âge	T_X	Probabilité observée
0	0	$p_{00} = 0{,}2$
0	1	$p_{01} = 0{,}4$
1	0	$p_{10} = 0{,}8$
1	1	$p_{11} = 0{,}6$

TABLEAU 1.3. Données issues d'une étude où le traitement T_X n'est pas randomisé : l'âge est déséquilibré entre les groupes de T_X.

Nous pouvons imaginer que les données du tableau 1.2 proviennent d'un essai randomisé, donc l'âge est équilibré dans les groupes du traitement T_X, et que les données du tableau 1.3 proviennent d'une étude observationnelle, donc l'âge n'est pas équilibré dans les groupes de traitement, par exemple, parce que les individus du groupe témoin sont plus jeunes et donc moins à risque de recevoir le traitement et d'avoir une pression artérielle élevée. Puisque nous avons l'indépendance entre le traitement T_X et l'âge au tableau 1.2, les effets conditionnels sont

$$\begin{aligned} E[PA|T_X = 0] &= \alpha + (0{,}2)\beta_1 \\ E[PA|T_X = 1] &= \alpha + \beta_2 + (0{,}2)\beta_1 \end{aligned} \tag{1.2.9}$$

et l'effet de T_X sur PA défini en (1.2.3) est égal à

$$E[PA|T_X = 1] - E[PA|T_X = 0] = \beta_2. \tag{1.2.10}$$

Dans le cas illustré au tableau 1.3, où le traitement T_X dépend de l'âge, les effets conditionnels sont

$$\begin{aligned} E[PA|T_X = 0] &= \alpha + (0{,}8)\beta_1 \\ E[PA|T_X = 1] &= \alpha + \beta_2 + (0{,}6)\beta_1 \end{aligned} \tag{1.2.11}$$

et l'effet de T_X sur PA défini en (1.2.3) est égal à

$$E[PA|T_X = 1] - E[PA|T_X = 0] = (0{,}2)\beta_1 + \beta_2. \tag{1.2.12}$$

Ainsi, au tableau 1.2 nous avons les rapports égaux $(p_1 \ : \ p_3) = (p_2 \ : \ p_4)$, ce qui nous permet d'estimer l'effet (1.2.3) sans confusion, alors qu'au tableau 1.3 nous avons des rapports différents, ce qui rend l'estimation biaisée de $0{,}2\beta_1$. Notons que l'inclusion de l'âge dans le modèle ferait disparaître ce biais [2].

1.2.2. Biais de confusion non mesuré

Dans le cadre des études observationnelles, le problème du biais de confusion dû à l'existence de causes communes à une exposition et une issue dont l'association nous intéresse est fréquent. Notons que le biais introduit par une variable de confusion parfaitement mesurée ne pose pas de problème pour l'estimation de l'effet, parce qu'il suffit alors d'ajuster pour cette variable dans le modèle statistique en l'incluant, par exemple, comme régresseur. Ainsi, en supposant que la vraie structure de cause à effet entre les variables est celle décrite par le DAG, l'inférence faite à partir de l'estimation est causale dans le contexte où il n'existe qu'une seule cause commune à l'exposition et l'issue, en autant que l'exposition précède l'issue dans le temps et que la modélisation du DAG en termes de modèles statistiques soit la bonne. Par exemple, si la relation entre l'exposition et l'issue est linéaire, alors un modèle qui utilise une relation quadratique mènerait à une estimation biaisée du vrai effet causal.

Le problème survient lorsque la variable de confusion est non mesurée. Si elle est non mesurée, par exemple, parce qu'elle est inconnue, alors l'ajustement ne peut pas se faire dans le modèle statistique. Par conséquent, il faut avoir recours à une autre technique pour éliminer la confusion de l'effet dans le modèle.

2. pourvu que l'erreur de mesure de la variable d'âge soit minime

1.3. Concepts génétiques

En randomisation mendélienne, nous exploitons des phénomènes génétiques pour simuler une étude où l'exposition est randomisée. Ainsi, afin de faciliter la compréhension de la technique de la randomisation mendélienne, nous présentons dans cette section une explication simplifiée des notions principales en génétique. Cette section ne prétend en aucun cas couvrir la théorie de base en génétique de façon exhaustive. Si le lecteur ou la lectrice a déjà une bonne connaissance des notions de base, il ou elle peut terminer sa lecture par la conclusion de ce chapitre.

1.3.1. Gène

Précisons d'abord qu'une séquence d'ADN est formée à partir d'une chaîne de quatre bases nucléotidiques : A, C, G et T [**1**]. Un gène est une séquence de bases nucléotidiques appartenant à une région spécifique du génome. Plusieurs gènes sont responsables de la production de protéines à la base des fonctions du corps humain. Par exemple, le gène *BRCA1* est responsable des mécanismes de réparation de l'ADN. De façon générale, la fonction d'un gène dans le corps humain est de coder pour une protéine. Une protéine est transcrite en termes d'un phénotype, c'est-à-dire le trait physique observable qui est associé au trait génétique codé par un ou plusieurs gènes. Un fragment de gène (appelé SNP, défini dans la section suivante), lorsque muté, fait que le gène peut modifier la production des protéines associées résultant en une sur- ou sous-expression du phénotype associé.

1.3.2. Polymorphisme nucléotidique

Un polymorphisme nucléotidique (en anglais *single nucleotide polymorphism*, ci-après SNP) est la variation d'une seule paire de bases nucléotidiques à une position précise du génome, entre les êtres humains (voir la figure 1.3). Certains SNPs sont situés sur des gènes, alors que d'autres sont situés entre des gènes ou dans des régions non codantes, c'est-à-dire qu'ils ne sont pas associés avec la production de protéines. La mutation d'un SNP, que ce soit par la modification d'une base nucléotidique ou par sa disparition, par exemple, peut résulter en une modification du phénotype qui lui est associé.

Ainsi, certains SNPs sont à l'origine de certaines différences dans notre prédisposition à développer une maladie et à exprimer un certain phénotype. Plus encore, les SNPs sont à l'origine de la sévérité avec laquelle une maladie se manifeste et la réponse physiologique à un traitement. Puisque différents SNPs peuvent avoir différents effets sur les phénotypes et les maladies qui peuvent atteindre un individu, il est possible d'étudier quels SNPs sont

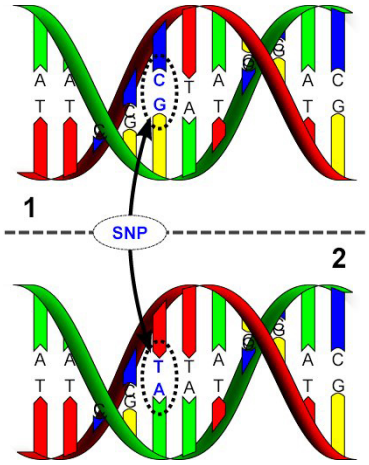


FIGURE 1.3. Illustration de la modification d'une paire de bases nucléotidiques [**19**].

associés avec quelles maladies, dans le but d'évaluer la prédisposition d'un individu à développer une certaine condition : par exemple, l'obésité. À cette fin, les études d'association pangénomiques (en anglais, *genome wide association studies*, ci-après GWAS) ont pour objectif de mesurer le degré d'association entre des millions de SNPs et un phénotype, individuellement.

Il est important de noter que les SNPs ne sont pas transmis de façon complètement aléatoire des parents vers l'enfant. Plutôt, les SNPs sont généralement transmis en groupes de SNPs avoisinants, appelés haplotypes. Ceci implique que les SNPs d'un même haplotype seront habituellement corrélés entre eux, un phénomène appelé déséquilibre de liaison. Cela importe considérablement pour les études d'association. Supposons qu'un SNP donné cause un changement dans l'expression d'un phénotype (nous disons que ce SNP est causal). Ce SNP appartient à un haplotype, c'est-à-dire qu'il est habituellement transmis avec une série de SNPs avoisinants qui bien que corrélés avec le SNP causal, ne sont pas causaux eux-mêmes. Ils sont plutôt considérés comme des marqueurs du SNP causal. Or, les GWAS ne sont pas en mesure de déterminer si les SNPs détectés ou non sont causaux. Par exemple, certains SNPs situés sur le gène FTO sont des excellents marqueurs pour le phénotype de l'obésité, sans pour autant être une cause de celle-ci [**24**]. Des études poussées, au-delà des GWAS, sont nécessaires pour établir la causalité d'un SNP (p.e. le *fine-mapping* [**20**]).

1.3.3. Allèle de risque

Au niveau de la population, un SNP peut habituellement se décomposer en trois groupes distincts composés de deux allèles conférés par chacun des parents [3]. Ainsi, si chacun des parents donne l'allèle A, alors l'enfant sera un homozygote commun. Si chacun des parents donne l'allèle a, alors l'enfant sera un homozygote rare. Les termes rare et commun font référence à la fréquence avec laquelle les allèles se retrouvent dans la population. Cette fréquence n'est pas aléatoire et sera expliquée dans la prochaine section.

Pour l'instant, clarifions la notion d'allèle de risque. Nous définissons l'allèle de risque comme étant la mutation (de type a ou A) qui confère un risque accru d'expression d'un phénotype. Sous ce modèle, si a est l'allèle de risque, alors les homozygotes rares en comptent 2, les hétérozygotes 1 et les homozygotes communs 0. Il est ainsi possible de représenter les trois groupes formés par un SNP en termes de nombre d'allèles de risque. L'unité de mesure d'une variable génétique est le nombre d'allèles de risque sur le SNP. Une telle variable prend donc des valeurs de 0, 1 ou 2.

1.3.4. Équilibre de Hardy-Weinberg

Nous avons établi précédemment que sur un SNP, nous pouvons habituellement observer l'une des trois combinaisons suivantes :

— AA : homozygote commun (0 allèles de risque),

— Aa : hétérozygote (1 allèle de risque),

— aa : homozygote rare (2 allèles de risque).

En effet, il existe une variation génétique à l'intérieur d'une population. De plus, lorsque l'accouplement est fait de façon aléatoire dans une grande population et que des facteurs tels que la mutation sont absents [4], la loi de Hardy-Weinberg (HWE) stipule que les fréquences génotypiques et alléliques demeureront en équilibre d'une génération à l'autre [**1**]. En d'autres termes, si nous dénotons par p la fréquence d'allèles de type A dans une population et par q la fréquence d'allèles de type a, alors dans cette population, la fréquence des allèles est dîte en équilibre de Hardy-Weinberg si

$$p^2 + 2pq + q^2 = 1. \tag{1.3.1}$$

Les variables génétiques qui sont des SNPs suivent la loi (1.3.1). Ainsi, la simulation des SNPs dans le chapitre 5 sera faite selon cette loi.

3. Dans la vaste majorité des cas, la région du chromosome sur laquelle un SNP est situé implique deux allèles. Nous nous restreignons donc à ces cas pour la suite de ce mémoire.

4. La loi de Hardy-Weinberg repose sur un total de cinq suppositions. Pour une liste plus exhaustive, voir p.e. [**4**].

Bien que les allèles de risque ne soient pas conférés de façon complètement aléatoire dans une population, nous trouvons généralement qu'ils ne sont pas associés à des variables de confusion potentielles dans une population donnée [**25**]. Par conséquent, nous pouvons faire l'hypothèse qu'au niveau de la population, la valeur que prend un SNP se décompose en trois groupes indépendants rendant ainsi tout phénotype résultant de ce SNP « randomisé » dans cette population. En d'autres termes, les individus peuvent être considérés comme avoir été assignés au hasard soit au groupe homozygotes communs, au groupe hétérozygotes, ou bien au groupe homozygotes rares pour le phénotype en question. Ainsi, nous pouvons faire une analogie au processus de randomisation employé dans les essais randomisés, et bénificions du fait que les distributions des variables de confusion dans les groupes du phénotype exposition sont équilibrées, nous permettant ainsi d'éliminer le problème du biais de confusion et d'inférer la causalité.

Chapitre 2

RANDOMISATION MENDÉLIENNE

Dans ce chapitre, nous définissons la méthode de la randomisation mendélienne. En premier lieu, nous introduisons la technique sur laquelle la randomisation mendélienne repose, soit l'instrumentation. Ensuite, nous présentons la notion d'un instrument génétique valide et illustrons les suppositions sur la validité à l'aide de l'exemple de l'effet de l'obésité sur la pression artérielle. Enfin, nous expliquons les motivations de combiner plusieurs instruments génétiques en randomisation mendélienne et en exposons les défis.

2.1. Définition de l'instrumentation et de la randomisation mendélienne

D'abord développée dans la discipline de l'économétrie [**43**], l'instrumentation est une technique qui permet d'estimer un effet causal entre une exposition D et une issue Y lorsque cet effet est influencé par une (ou plusieurs) variable de confusion non mesurée U_1 [**36**] tel qu'illustré à la figure 2.1.

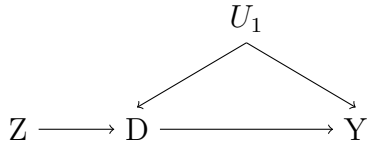


Figure 2.1. DAG représentant la situation idéale pour utiliser l'instrumentation, où D représente le phénotype exposition, Y, le phénotype issue, Z, l'instrument génétique et U_1 la variable de confusion non mesurée.

Cette technique utilise une troisième variable, l'instrument Z, qui est corrélée avec D mais indépendante des variables de confusion, pour estimer l'effet causal de D sur Y, évitant ainsi tout biais de confusion dû à U_1. Notons qu'à la figure 2.1 nous utilisons la flèche $Z \rightarrow D$ pour représenter un lien entre Z et D qui n'est pas nécessairement causal,

mais ceci est un abus de notation qui apparaît uniquement dans la représentation du lien entre l'instrument et l'exposition.

La randomisation mendélienne est une application de la technique de l'instrumentation dans laquelle les instruments sont des marqueurs génétiques [**6**]. Bien que plusieurs marqueurs génétiques puissent servir d'instruments en randomisation mendélienne, par simplicité, nous nous restreindrons au cas le plus fréquemment utilisé, soit l'utilisation d'un ou plusieurs SNPs comme instrument(s). Dans le chapitre 3, nous verrons deux méthodes d'estimation par l'instrumentation dans le cadre de leur application en randomisation mendélienne.

2.2. Validité d'un instrument en randomisation mendélienne

Dans cette section, nous définissons plus précisément ce qu'est un instrument dans le contexte de la randomisation mendélienne. Pour ce faire, nous introduisons le concept de la validité d'un instrument, soit l'ensemble des suppositions que doivent satisfaire un SNP candidat à l'instrumentation avant d'être utilisé pour estimer l'effet causal. Enfin, nous illustrerons comment ces suppositions peuvent ne pas être respectées dans la pratique à l'aide de l'exemple de l'effet de l'obésité sur la pression artérielle.

2.2.1. Définition d'un instrument génétique valide

Dans un problème donné, un SNP peut être utilisé comme instrument s'il satisfait la définition suivante.

Définition 2.2.1. *Un SNP Z est un instrument valide si :*

- *(A1) il est corrélé avec le phénotype exposition D ;*
- *(A2) il n'existe pas de cause commune non mesurée entre Z et le phénotype issue Y ;*
- *(A3) tout chemin dirigé (séquence de flèches) de Z allant vers Y passe uniquement par D* [**18**].

La supposition (A1) implique que l'instrument Z prédit D. Par exemple, puisque (A1) requiert que Z prédise D, et non pas que Z cause D, n'importe quel SNP qui a une association forte avec D peut être un instrument valide au niveau de (A1). Soulignons que plus l'association est forte entre Z et D, plus la portion de D qui est prédite par Z est importante, et meilleure est la force de l'instrument Z. Notons que (A1) peut être vérifiée dans les données à l'aide de modèles d'association statistique appropriés : une régression logistique si D est binaire et une régression linéaire simple si D est continue.

La supposition (A2) implique qu'il n'existe pas de variable non ou mal mesurée précédant l'assignation de Z dans le temps qui influence à la fois Z et Y. Remarquons d'abord

qu'une variable pouvant influencer à la fois Z et Y doit être préalable à Z et Y chacun. Dans la vaste majorité des cas, la seule variable qui peut être préalable à Z est l'origine ethnique puisque la composition génétique d'un individu est déterminée avant sa naissance. L'origine ethnique est une cause de Z parce qu'elle cause l'apparition plus ou moins fréquente de mutations génétiques dans une population donnée. De plus, puisque l'origine a une influence sur les gènes d'individus issus d'une population donnée, elle influence également l'apparition de certains traits physiques, donc de phénotypes, à l'intérieur d'une population donnée. Ainsi, l'origine ethnique est une variable qui pourrait être une cause commune à l'instrument et au phénotype issue. Cet exemple est illustré à la figure 2.2.

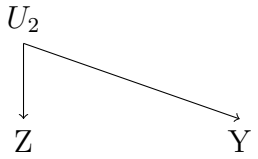


FIGURE 2.2. DAG partiel représentant une violation de (A2).

La flèche de gauche à la figure 2.2 implique que U_2 précède l'assignation des allèles de risque à l'individu lors de sa conception, tandis que la flèche de droite implique que U_2 influence le phénotype issue indépendamment de la fréquence des allèles de risque. C'est l'existence des deux flèches dans un même DAG qui implique une violation de (A2). Dans le cas où U_2 représente l'origine ethnique, il y aurait un problème si la fréquence des allèles de risque ainsi que la prévalence de l'issue différaient selon les origines ethniques et que l'étude contenait des individus de plusieurs ethnies différentes. Toutefois, en pratique, il est habituellement possible de déterminer et d'ajuster pour l'origine ethnique dans le modèle de randomisation mendélienne, minimisant ainsi la possibilité d'une violation de (A2) [**29**].

La supposition (A3) implique que Z influence Y à travers D uniquement. Supposons que Z influence Y par un chemin autre que $Z \to D \to Y$ tel qu'illustré dans les figures 2.3 et 2.4.

À la figure 2.3, l'existence de la flèche allant de Z vers Y directement implique que le SNP utilisé comme instrument influence non seulement l'expression du phénotype exposition, mais aussi celle du phénotype issue. De façon similaire, à la figure 2.4, si U_3 est un autre phénotype inconnu, l'existence du chemin $Z \to U_3 \to Y$ implique que le SNP utilisé comme instrument influence l'expression d'un autre phénotype qui a également une influence sur le phénotype issue. Notons que, au contraire de (A2) qui ne pose pas de problèmes importants qui ne peuvent pas être résolus en pratique, la préoccupation avec (A3) n'est pas que théorique, car les gènes peuvent agir sur plusieurs phénotypes et la plupart de ces mécanismes d'action sont inconnus ou mal compris [**34**].

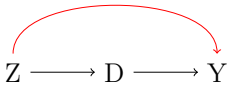


FIGURE 2.3. Exemple d'une violation de la supposition (A3).

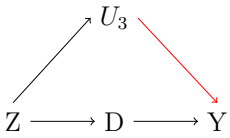


FIGURE 2.4. Autre violation de la supposition (A3), U_3 est un phénotype inconnu.

Ainsi, puisque (A1) peut être vérifiée dans les données et qu'une violation de (A2) est, en pratique, possible à minimiser en ajustant pour l'origine ethnique, la préoccupation majeure pour la sélection d'instruments en randomisation mendélienne est la supposition (A3) : s'il existe des variables telles que U_3 qui sont non ou mal mesurées, alors nous ferons face à un problème dans l'estimation de l'effet causal de D sur Y.

2.2.2. Illustration de la validité en randomisation mendélienne

Reprenons l'exemple de l'effet de l'obésité (mesurée par l'IMC) sur la pression artérielle afin d'illustrer le concept de la validité d'un instrument (voir la figure 2.5). Pour cet exemple, nous supposons comme variable de confusion pour l'effet causal de l'IMC sur la PA le statut socio-économique de l'individu (SSE) qui n'est pas mesuré pour la population étudiée. Un exemple d'un tel méchanisme est le suivant. D'un côté, nous supposons que le SSE influence l'obésité si le fait d'avoir un statut socio-économique plus bas implique que l'individu fait le choix d'aliments moins dispendieux et donc moins nutritifs : par exemple, des aliments ayant une majorité d'aliments transformés. De l'autre côté, nous supposons que le SSE influence la PA si le fait d'avoir un statut socio-économique plus bas implique que l'individu a un travail plus demandant et donc doit gérer un niveau de stress plus élevé, ce qui pourrait mener à de l'hypertension.

Comme nous l'avons vu au chapitre 1, l'effet de l'IMC sur la PA estimé par un modèle de régression linéaire simple qui n'ajuste pas pour la variable SSE serait biaisé parce que le SSE est une variable de confusion. Afin d'éliminer le biais de confusion qui s'en suit, nous utilisons un SNP du gène FTO, nommé rs16953002, comme instrument génétique. Nous savons, de par les études d'association pangénomiques, que rs16953002 est fortement associé à l'IMC, ce qui suggère que ce sera un instrument fort. Nous pouvons vérifier que cette association existe aussi dans nos données. Ainsi, rs16953002 satisfait (A1). Le DAG de la figure 2.5 illustre comment les suppositions (A2) et (A3) peuvent être contredites

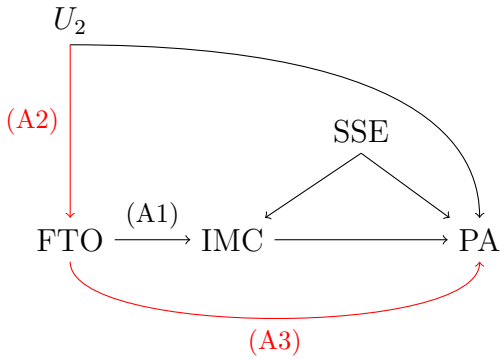


FIGURE 2.5. DAG représentant les suppositions de la validité d'instruments en randomisation mendélienne pour le mécanisme IMC cause PA. Une flèche rouge indique une violation d'une supposition, tandis qu'une flèche noire indique que la supposition est satisfaite.

dans l'application de la randomisation mendélienne pour étudier l'effet de l'IMC sur la PA, confondu par le SSE.

Prenons d'abord la supposition (A2). Dans la figure 2.5, une flèche lie la variable de confusion U_2 et l'instrument FTO. Il y a donc une cause commune non mesurée à FTO et PA. Cela pourrait se produire par l'influence des origines ethniques présentée dans la section 2.3.1. Par exemple, supposons que la population étudiée comprenne un sous-ensemble d'individus d'Afrique de l'Est, et que de par son évolution, il soit associé à une fréquence plus élevée de la mutation rs16953002 de FTO (flèche $U_2 \to FTO$) et une prévalence d'hypertension plus grande [1] (flèche $U_2 \to PA$). Alors nous ferions face à une violation de (A2).

Prenons maintenant la supposition (A3). Le paradigme actuel en génétique soutient que les SNPs de FTO influencent de multiples traits physiques uniquement par leur association avec l'obésité. Or, FTO est un gène dont la fonction sur le corps humain est probablement plus vaste que son influence sur l'obésité [**21**]. Par conséquent, nous supposons qu'il pourrait influencer le phénotype PA sans passer par l'IMC. En particulier, rs16953002 serait associé au développement de mélanomes (M), un trait physique lié à la prévalence du cancer de la peau, et ce indépendamment de l'IMC [**21**]. De plus, les traitements pour mélanomes (T_M) ont souvent l'effet d'augmenter la pression artérielle [**28**]. Donc, cela pourrait créer un autre chemin allant de FTO à PA, soit $FTO \to M \to T_M \to PA$, ce qui impliquerait une violation de (A3). Dans ce cas, il est possible que rs16953002 soit un instrument invalide. L'utiliser en randomisation mendélienne pourrait mener à des résultats biaisés dans l'estimation de l'effet, entre autres si les changements observés dans le niveau de pression artérielle résultent de facteurs autres que l'IMC qui ne sont pas mesurés [**18**].

1. Ceci est un exemple hypothétique inspiré de [**35**].

Pour résumer cet exemple, notons que (A1) est satisfait puisque nos données montrent une association forte entre l'instrument rs16953002 et l'IMC. Ensuite, pour satisfaire (A2) il suffit d'inclure une variable représentant l'origine ethnique de façon appropriée dans le modèle d'estimation. De plus, il n'y a pas d'inquiétude de biaiser les résultats via (A3) si la population étudiée ne comporte pas de patients qui sont sous traitement pour des mélanomes. Par conséquent, si nous respectons les mises en garde énumérées précédemment, l'instrument rs16953002 est valide pour estimer l'effet causal de l'IMC sur la PA par un modèle de randomisation mendélienne. Enfin, cet exemple souligne l'importance de considérer qu'un SNP peut être invalide pour une population, mais être valide pour une autre. Donc la validité d'un SNP comme instrument en randomisation mendélienne est toujours conditionnelle à la population étudiée.

2.3. Combinaison d'instruments en randomisation mendélienne

En randomisation mendélienne, il est parfois nécessaire d'utiliser plusieurs SNPs comme instruments pour améliorer la prédiction de l'exposition D (A1). Toutefois, plus il y d'instruments, plus il y a de suppositions de type (A2) et (A3) à satisfaire pour que tous ces instruments soient valides. Ainsi, utiliser plusieurs SNPs implique un plus grand risque de multiplier les sources d'invalidité. Dans cette section, nous abordons le défi que pose l'utilisation de multiples SNPs en randomisation mendélienne. Notons que pour la suite, nous nous restreindrons au cas de SNPs peu corrélés entre eux.

2.3.1. Compromis entre la force et la validité des instruments

Le DAG suivant représente l'instrumentation multiple en randomisation mendélienne.

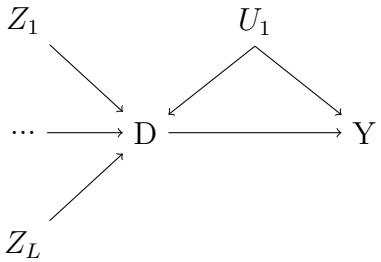


Figure 2.6. DAG illustrant l'instrumentation multiple, où D représente le phénotype exposition, Y, le phénotype issue, Z_1 à Z_L , les instruments génétiques et U_1 la variable de confusion non mesurée.

Augmenter le nombre de SNPs a deux conséquences majeures sur l'application de la randomisation mendélienne pour estimer l'effet causal d'un phénotype exposition sur un phénotype issue. D'une part, la force globale des instruments augmente, parce que chacun

des L SNPs à la figure 2.6 prédit une partie de l'exposition, créant ainsi un signal plus fort pour l'association avec D. Ceci fait référence à la supposition (A1) dans la définition 2.2.1. D'autre part, utiliser plus de SNPs augmente le risque d'invalidité des instruments, en particulier par le biais de (A3). En effet, chacun des SNPs à la figure 2.6 peut provenir d'un gène différent, et chaque gène peut avoir une fonction sur le corps humain qui n'est pas entièrement connue. Alors, le mécanisme $Z \to D \to Y$ se complexifie davantage par la présence de phénotypes inconnus qui créent possiblement des chemins de Z à Y qui ne passent pas par D, soit des violations de (A3). Nous pouvons donc faire face à une plus grande source de biais dans l'estimation de l'effet causal [**25**].

Dans ce chapitre, nous avons présenté les suppositions qui doivent être faites pour qu'un SNP soit valide. Nous avons expliqué les différentes sources d'invalidité à l'aide de DAGs illustrant des situations où l'instrument est simple, c'est-à-dire qu'il comprend un seul SNP. Nous avons aussi mis le lecteur en garde contre le compromis à faire lors de la combinaison de plusieurs SNPs. D'un côté, la force du signal pour le phénotype exposition est meilleure si les SNPs sélectionnés ont une association forte avec ce phénotype. D'un autre côté, utiliser plusieurs SNPs implique que les chances de concevoir une violation des suppositions (A2) et (A3) se multiplient, d'où la nécessité de faire un compromis entre la force (A1) et la validité de (A2)-(A3). Les aspects techniques de l'estimation par l'instrumentation peuvent désormais être élaborés dans le chapitre suivant.

Chapitre 3

ESTIMATEURS À INSTRUMENTS MULTIPLES

Dans ce chapitre, nous expliquons l'estimation de l'effet causal d'une exposition sur une issue en randomisation mendélienne. Rappelons que, dans ce contexte, le fait qu'il y a des causes communes non mesurées dans le vrai modèle causal impliquera un biais de confusion dans le modèle d'estimation. Comme nous allons le démontrer dans les sections qui suivent, l'estimateur des moindres carrés ordinaires n'est pas robuste à la confusion. Par conséquent, pour corriger ce problème, les études de randomisation mendélienne utilisent des méthodes d'estimation par les variables instrumentales pour gérer le biais de confusion.

Nous présentons deux estimateurs utilisés en randomisation mendélienne. Dans un premier temps, nous présentons l'estimateur le plus conventionnellement utilisé en randomisation mendélienne, appelé la méthode des moindres carrés en deux étapes (2SLS). Dans un deuxième temps, nous expliquons l'estimateur sisVIVE qui a été récemment proposé dans la littérature [**23**]. Ce dernier vise à réduire la sévérité des suppositions de l'estimateur 2SLS pour gérer les problèmes d'association entre l'instrument et l'issue qui ne passent pas par l'exposition. Ces problèmes sont très présents dans la pratique, en raison de la pléiotropie d'un SNP par rapport à l'issue étudiée, ce qui le rend alors invalide en tant qu'instrument.

3.1. Contexte de modélisation

Avant de procéder à l'explication des méthodes d'estimation pour instruments multiples, commençons par établir la notation utilisée dans ce chapitre. Notons par n la taille de l'échantillon, soit le nombre d'individus observés. Parmi les variables observées, notons par $\mathbf{Y}$ la variable d'issue et par $\mathbf{D}$ la variable d'exposition, qui auront chacune une distribution continue dans le contexte de ce mémoire. Ainsi, $\mathbf{Y}$ et $\mathbf{D}$ sont des vecteurs $n \times 1$ basés sur les observations des phénotypes issue et exposition, respectivement. Notons par $\mathbf{U}$ le vecteur $n \times 1$ basé sur les observations non mesurées d'une cause commune à l'exposition $\mathbf{D}$ et l'issue $\mathbf{Y}$. Ainsi, $\mathbf{U}$ est une variable de confusion pour l'effet causal de $\mathbf{D}$ sur $\mathbf{Y}$.

Nous allons nous restreindre au cas où nous avons une exposition, une issue et une cause commune non mesurée, mais ce qui sera présenté se généralise bien à plusieurs variables de confusion.

Ensuite, notons par $\mathbf{Z}$ la matrice $n \times L$ formée d'un ensemble de $L \geq 1$ instruments génétiques codés par les valeurs que prennent les SNPs, soit 0, 1 ou 2. Ainsi, les colonnes de $\mathbf{Z}$ correspondent à des SNPs et les rangées de $\mathbf{Z}$ correspondent au nombre d'allèles de risque des SNPs propres à un individu.

Enfin, notons par β^D le paramètre de l'effet causal de l'exposition $\mathbf{D}$ sur l'issue $\mathbf{Y}$ pour lequel nous voulons obtenir une estimation.

Pour la suite de ce mémoire, nous supposons que la corrélation entre les instruments est faible parce qu'ils sont choisis comme n'étant pas en déséquilibre de liaison, tel que mentionné à la section 1.3.2.

3.2. Modèle causal en randomisation mendélienne

Rappelons que la problématique est d'estimer l'effet causal de l'exposition $\mathbf{D}$ sur l'issue $\mathbf{Y}$ par la technique de la randomisation mendélienne pour contourner le problème du biais de confusion.

Avant de procéder à l'explication des estimateurs pour cet effet, nous allons d'abord définir le vrai modèle causal, c'est-à-dire, le processus de génération des données individuelles. Pour ce faire, nous présentons les DAGs et les modèles statistiques qui y sont associés pour deux cas : en premier, celui où les SNPs satisfont la définition (2.2.1) de la validité d'un instrument en randomisation mendélienne, en deuxième, celui où au moins un SNP est invalide, en particulier parce qu'il ne satisfait pas la supposition (A3) de cette définition.

3.2.1. Cas 1 : instruments valides

Dans un premier temps, supposons que les instruments sont valides selon (A1)-(A3). Supposons que le vrai processus de génération des données individuelles est représenté par le DAG à la figure 3.1, où $\mathbf{U}$ est une cause commune à l'exposition et l'issue et $\mathbf{Z}$ représente les SNPs. Les relations statistiques qui découlent du DAG à la figure 3.1 sont :

— $\mathbf{Z}$ est indépendant de $\mathbf{U}$,
— $\mathbf{D}$ est dépendant de $\mathbf{Z}$ et $\mathbf{D}$ est dépendant de $\mathbf{U}$,
— $\mathbf{Y}$ est dépendant de $\mathbf{D}$, $\mathbf{Y}$ est dépendant de $\mathbf{U}$ et $\mathbf{Y}$ est indépendant de $\mathbf{Z}$ conditionnellement à $\mathbf{D}$ et $\mathbf{U}$.

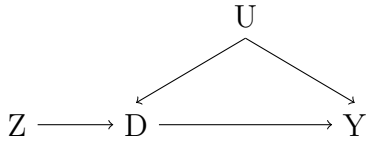


FIGURE 3.1. DAG pour le cas 1, représentant le vrai processus de génération des données individuelles pour l'effet d'une exposition **D** sur une issue **Y** lorsque les SNPs sont des instruments valides pour l'association entre **D** et **Y**.

Ensuite, à partir de ces relations statistiques, nous pouvons traduire le DAG à la figure 3.1 en termes des modèles de régression linéaire multiple suivants :

$$\begin{aligned} \mathbf{D} &= \pi^D \mathbf{1} + \mathbf{U}\delta^D + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\boldsymbol{D}} \\ \mathbf{Y} &= \pi^Y \mathbf{1} + \mathbf{U}\delta^Y + \mathbf{D}\beta^D + \boldsymbol{\epsilon}^{\boldsymbol{Y}}, \end{aligned} \tag{3.2.1}$$

où $\boldsymbol{\gamma}$ est le vecteur $L \times 1$ basé sur l'effet de chaque SNP sur l'exposition **D**. Ainsi, dans le cas illustré par la figure 3.1 où les conditions d'indépendance décrites plus haut sont satisfaites, les SNPs sont des instruments valides selon (A1)-(A3) pourvu que $\gamma_j \neq 0, \quad \forall j = 1, \ldots, L$. Nous faisons également l'hypothèse que $\mathbb{E}(\boldsymbol{\epsilon}^{\boldsymbol{D}}) = \mathbb{E}(\boldsymbol{\epsilon}^{\boldsymbol{Y}}) = 0$ qui correspond à la supposition qu'il n'y a pas de variable de confusion non mesurée pour l'effet causal de **D** sur **Y**.

3.2.2. Cas 2 : instruments invalides

Dans un deuxième temps, supposons plutôt que le vrai processus de génération des données individuelles est représenté par le DAG à la figure 3.2 dans lequel la supposition (A3) n'est pas respectée. Les relations statistiques qui découlent de ce DAG diffèrent de

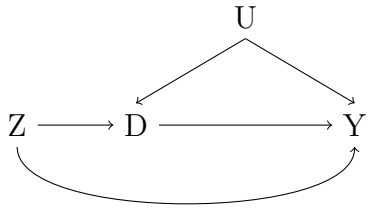


FIGURE 3.2. DAG pour le cas 2, représentant le vrai processus de génération des données individuelles pour l'effet d'une exposition D sur une issue Y lorsqu'au moins un SNP est un instrument invalide selon (A3).

celles pour le DAG représentant les SNPs qui sont des instruments valides. En effet, en raison de la violation de (A3), nous avons que Y et Z sont dépendants même si nous conditionnons sur D et U.

À partir de ces relations statistiques, nous traduisons le DAG à la figure 3.2 en termes des modèles de régression linéaire multiple suivants :

$$\begin{aligned} \mathbf{D} &= \pi^D \mathbf{1} + \mathbf{U}\delta^D + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\boldsymbol{D}} \\ \mathbf{Y} &= \pi^Y \mathbf{1} + \mathbf{U}\delta^Y + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{D}\beta^D + \boldsymbol{\epsilon}^{\boldsymbol{Y}}, \end{aligned} \tag{3.2.2}$$

où $\boldsymbol{\alpha}$ est le vecteur $L \times 1$ basé sur l'effet de chaque SNP sur l'issue $\mathbf{Y}$. Ainsi, dans le cas illustré par la figure 3.2 où les conditions d'indépendance décrites plus haut sont satisfaites, $\mathbf{Z}$ comprend un SNP qui est un instrument invalide selon la supposition (A3) si $\exists j \in 1, \ldots, L$ pour lequel $\alpha_j \neq 0$. Enfin, nous supposons à nouveau qu'il n'y a pas de variable de confusion non mesurée pour l'association entre $\mathbf{D}$ et $\mathbf{Y}$ en faisant l'hypothèse que $\mathbb{E}(\boldsymbol{\epsilon}^{\boldsymbol{D}}) = \mathbb{E}(\boldsymbol{\epsilon}^{\boldsymbol{Y}}) = 0$.

Dans la pratique, l'effet $\boldsymbol{\alpha}$ est attribuable au phénomène de pléiotropie, qui admet qu'un SNP puisse avoir un effet sur plus d'un phénotype [**34**]. Dans le cas qui nous intéresse, cet effet n'est pas seulement sur le phénotype $\mathbf{D}$, mais aussi sur le phénotype $\mathbf{Y}$, par un mécanisme autre que celui qui passe par $\mathbf{D}$.

3.3. Impact de la confusion non mesurée sur le modèle d'estimation de l'effet causal

Nous allons maintenant reformuler le modèle causal du cas des instruments valides en (3.2.1) pour refléter le fait que la variable de confusion $\mathbf{U}$ est non mesurée. Par la suite, nous pourrons expliciter l'impact de l'omission de $\mathbf{U}$ dans les modèles sur (1) les termes d'erreur et (2) l'effet causal de l'exposition $\mathbf{D}$ sur l'issue $\mathbf{Y}$ estimé par la méthode des moindres carrés ordinaires.

Réécrivons les modèles d'estimation de façon à montrer comment les erreurs ne sont plus indépendantes. D'abord, pour l'exposition, le modèle (3.2.1) devient :

$$\begin{aligned} \mathbf{D} &= \pi^D \mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\boldsymbol{D}} + \mathbf{U}\delta^D \\ &= (\pi^D + \mathbb{E}(\mathbf{U}\delta^D))\mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\boldsymbol{D}} + [\mathbf{U} - \mathbb{E}(\mathbf{U})]\delta^D \\ &= c^D \mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}^{\mathbf{D}}, \end{aligned} \tag{3.3.1}$$

où $c^D = \pi^D + \mathbb{E}(\mathbf{U}\delta^D)$ et $\mathbf{e}^{\mathbf{D}} = \boldsymbol{\epsilon}^{\boldsymbol{D}} + [\mathbf{U} - \mathbb{E}(\mathbf{U})]\delta^D$ puisque l'effet de la variable de confusion non mesurée $\mathbf{U}$ sera absorbé dans le terme d'erreur. De plus, comme la contribution de $\mathbb{E}(\mathbf{U})$ entre dans la définition de la constante c^D, alors nous pouvons faire sans perte de généralité l'hypothèse que $\mathbb{E}(\mathbf{U}) = 0$. Ensuite, nous faisons le développement similaire

pour l'issue :

$$\begin{aligned}\mathbf{Y} &= \pi^Y \mathbf{1} + \mathbf{D}\beta^D + \boldsymbol{\epsilon}^{\boldsymbol{Y}} + \mathbf{U}\delta^Y \\ &= (\pi^Y + \mathbb{E}(\mathbf{U}\delta^Y))\mathbf{1} + \mathbf{D}\beta^D + \boldsymbol{\epsilon}^{\boldsymbol{Y}} + [\mathbf{U} - \mathbb{E}(\mathbf{U})]\delta^Y \\ &= c^Y \mathbf{1} + \mathbf{D}\beta^D + \mathbf{e}^{\mathbf{Y}},\end{aligned} \tag{3.3.2}$$

où $c^Y = \pi^Y + \mathbb{E}(\mathbf{U}\delta^Y)$ et $\mathbf{e}^{\mathbf{Y}} = \boldsymbol{\epsilon}^{\boldsymbol{Y}} + [\mathbf{U} - \mathbb{E}(\mathbf{U})]\delta^Y$. À nouveau, nous pouvons faire sans perte de généralité l'hypothèse que $\mathbb{E}(\mathbf{U}) = 0$. Ainsi, dorénavant nous faisons l'hypothèse que $\mathbb{E}(\mathbf{U}) = 0$.

Ainsi, nous remarquons qu'étant donné que $\mathbf{U}$ est non mesurée, $\mathbf{e}^{\mathbf{D}}$ et $\mathbf{e}^{\mathbf{Y}}$ sont tous les deux des fonctions de $\mathbf{U}$, alors ils ne peuvent pas être indépendants. Le même développement est fait pour le cas des instruments invalides à l'annexe B.1.1 et mène aux mêmes conclusions que pour le cas des instruments valides.

Démontrons maintenant que l'estimateur des moindres carrés ordinaires ne parvient pas à gérer l'effet engendré par la variable de confusion non mesurée dans le modèle postulé ci-haut. Rappelons que nous avons les relations suivantes entre les termes du modèle :

$$\mathbf{Z} \perp\!\!\!\perp \mathbf{U} \perp\!\!\!\perp \boldsymbol{\epsilon}^{\boldsymbol{D}} \perp\!\!\!\perp \boldsymbol{\epsilon}^{\boldsymbol{Y}}. \tag{3.3.3}$$

Soit la matrice $\mathbf{X} = \Big(\mathbf{1} \,|\, \mathbf{D} \Big)$. Nous utilisons la notation de [**41**] pour définir la matrice résultante du centrage des données dans $\mathbf{X}$,

$$\mathcal{D} = \begin{bmatrix} D_1 - \bar{D} \\ D_2 - \bar{D} \\ \vdots \\ D_n - \bar{D} \end{bmatrix} = (\mathbf{I} - \mathbf{P_1})\mathbf{D},$$

où $\mathcal{D}$ est la matrice avec les observations D_i centrées par la moyenne $\bar{D}$ et $\mathbf{P_1}$ est la matrice de projection sur l'espace engendré par le vecteur $\mathbf{1}$. Le modèle des moindres carrés ordinaires est :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}^{\mathbf{Y}}, \tag{3.3.4}$$

où $\boldsymbol{\beta} = (c^Y, \beta^D)^T$. Incidemment, ce modèle correspond à celui en (3.3.2). Si nous l'estimons par la méthode des moindres carrés ordinaires (OLS) nous obtenons, pour β^D :

$$\hat{\beta}^D_{OLS} = (\mathcal{D}^T\mathcal{D})^{-1}\mathcal{D}^T\mathbf{Y}. \tag{3.3.5}$$

D'où,

$$\begin{aligned}\hat{\beta}^D_{OLS} &= [\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{D}]^{-1}\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{Y} \\ &= [\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{D}]^{-1}\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})(\pi^Y\mathbf{1}+\mathbf{D}\beta^D+\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\mathbf{U}\delta^Y) \\ &= \beta^D + [\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{D}]^{-1}\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\boldsymbol{\epsilon}^{\boldsymbol{Y}} \\ &+ [\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{D}]^{-1}\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{U}\delta^Y\end{aligned} \tag{3.3.6}$$

Maintenant, en notant que $(\mathbf{I}-\mathbf{P_1})\mathbf{1}=\mathbf{0}$ et en définissant $\mathcal{Z}=(\mathbf{I}-\mathbf{P_1})\mathbf{Z}$ (voir l'équation (3.4.5) plus loin),

$$\begin{aligned}\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{U}\delta^Y &= (\pi^D\mathbf{1}^T+\boldsymbol{\gamma}^T\mathbf{Z}^T+\boldsymbol{\epsilon}^{\boldsymbol{D}^T}+\delta^D\mathbf{U}^T)(\mathbf{I}-\mathbf{P_1})\mathbf{U}\delta^Y \\ &= \boldsymbol{\gamma}^T\mathcal{Z}^T\mathbf{U}\delta^Y+\boldsymbol{\epsilon}^{\boldsymbol{D}^T}(\mathbf{I}-\mathbf{P_1})\mathbf{U}\delta^Y+\delta^D\delta^Y\mathbf{U}^T(\mathbf{I}-\mathbf{P_1})\mathbf{U}.\end{aligned} \tag{3.3.7}$$

Ainsi le dernier terme de (3.3.6) devient,

$$\begin{aligned}n(\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{D})^{-1}n^{-1}[\boldsymbol{\gamma}^T\mathcal{Z}^T+\boldsymbol{\epsilon}^{\boldsymbol{D}^T}(\mathbf{I}-\mathbf{P_1})\mathbf{U}\delta^Y+\delta^D\delta^Y\mathbf{U}^T(\mathbf{I}-\mathbf{P_1})\mathbf{U}] \\ = O_p(1)[O_p(n^{-1})+O_p(n^{-1/2})+O_p(1)] \\ = O_p(1)\end{aligned} \tag{3.3.8}$$

puisque $\mathcal{Z}=O_p(1)$ et $n^{-1}\boldsymbol{\epsilon}^{\boldsymbol{D}^T}(\mathbf{I}-\mathbf{P_1})\mathbf{U}\delta^Y$ étant une moyenne de variables aléatoires centrées, ce terme est $O_p(n^{-1/2})$ alors que $\mathbf{U}^T(\mathbf{I}-\mathbf{P_1})\mathbf{U}$ est une forme quadratique qui est $O_p(n)$. Ainsi,

$$\hat{\beta}^D_{OLS}=\beta^D+(\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{D})^{-1}\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\boldsymbol{\epsilon}^{\boldsymbol{Y}}+O_p(1) \tag{3.3.9}$$

et bien que le second terme de (3.3.6) en moyenne sera

$$\mathbb{E}[(\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})\mathbf{D})^{-1}\mathbf{D}^T(\mathbf{I}-\mathbf{P_1})]\mathbb{E}[\boldsymbol{\epsilon}^{\boldsymbol{Y}}]=0 \tag{3.3.10}$$

puisque $\mathbf{D}$ est indépendant de l'erreur de $\mathbf{Y}$ et que $\mathbb{E}[\boldsymbol{\epsilon}^{\boldsymbol{Y}}]=0$ par supposition, le troisième terme ne s'amenuise pas en augmentant n. Donc, le biais engendré par la variable de confusion non mesurée est non négligeable.

Ce résultat est l'une des motivations statistiques principales pour l'utilisation des estimateurs qui font usage des variables instrumentales. Dans la section suivante, nous en présentons un exemple fréquemment utilisé, soit l'estimateur des moindres carrés en deux étapes.

3.4. Estimateur des moindres carrés en deux étapes (2SLS)

Dans cette section, nous présentons un estimateur conventionnellement utilisé dans les méthodes d'estimation par les variables instrumentales, soit l'estimateur des moindres

carrés en deux étapes (en anglais, *two-stage least squares*, d'où 2SLS). Nous démontrons son utilité à gérer le biais de confusion engendré par l'effet d'une variable de confusion non mesurée. Nous démontrons également ses limites dans l'estimation de l'effet causal lorsque un ou plusieurs instruments sont des SNPs pléiotropiques impliquant l'issue $\mathbf{Y}$.

3.4.1. Modèle de l'estimateur 2SLS

D'abord, rappelons le modèle linéaire additif à estimer, qui correspond à celui en (3.3.1) et (3.3.2) où la variable de confusion $\mathbf{U}$ n'est pas observée :

$$\begin{aligned} \mathbf{D} &= \pi^D \mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\boldsymbol{D}} + \mathbf{U}\delta^D \\ &= c^D \mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}^{\mathbf{D}}, \\ \mathbf{Y} &= \pi^Y \mathbf{1} + \mathbf{D}\beta^D + \boldsymbol{\epsilon}^{\boldsymbol{Y}} + \mathbf{U}\delta^Y \\ &= c^Y \mathbf{1} + \mathbf{D}\beta^D + \mathbf{e}^{\mathbf{Y}}. \end{aligned} \tag{3.4.1}$$

Exprimons la dépendance entre les termes d'erreur $\mathbf{e}^{\mathbf{D}}$ et $\mathbf{e}^{\mathbf{Y}}$ sous la forme de $\sigma_{D,Y}$. Le modèle postulé pour l'estimation par la méthode 2SLS est le suivant :

$$\mathbf{D} = \mu^D \mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\xi}^{\boldsymbol{D}} \tag{3.4.2}$$

pour l'exposition et pour l'issue,

$$\mathbf{Y} = \mu^Y \mathbf{1} + \mathbf{D}\beta^D + \boldsymbol{\xi}^{\boldsymbol{Y}}. \tag{3.4.3}$$

La structure de covariance est,

$$(\xi^D, \xi^Y)^T \sim N\Big(\left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} \sigma_D^2 & \sigma_{D,Y} \\ \sigma_{D,Y} & \sigma_Y^2 \end{smallmatrix}\right]\Big), \tag{3.4.4}$$

où $\sigma_{D,Y} \neq 0$. La première équation illustre la première étape (ou *first stage*, en anglais) où les instruments satisfont les suppositions (A1)-(A3). La deuxième équation illustre la deuxième étape (ou *second stage*, en anglais) de la méthode 2SLS.

3.4.2. Estimation avec 2SLS

Définissons la matrice $\mathbf{W} = \Big(\, \mathbf{1} \mid \mathbf{Z} \,\Big)$. Nous utilisons à nouveau la notation de [**41**] pour définir les matrices résultantes du centrage des données dans $\mathbf{W}$, soit :

$$\mathcal{Z} = \begin{bmatrix} Z_{11} - \bar{Z}_1 & Z_{12} - \bar{Z}_2 & \cdots & Z_{1L} - \bar{Z}_L \\ Z_{21} - \bar{Z}_1 & Z_{22} - \bar{Z}_2 & \cdots & Z_{2L} - \bar{Z}_L \\ \vdots & \vdots & \cdots & \vdots \\ Z_{n1} - \bar{Z}_1 & Z_{n2} - \bar{Z}_2 & \cdots & Z_{nL} - \bar{Z}_L \end{bmatrix} = (\mathbf{I} - \mathbf{P_1})\mathbf{Z}, \tag{3.4.5}$$

où la matrice $\mathcal{Z}$ est la matrice des observations centrées Z_{ij}. Nous faisons les suppositions suivantes entre les termes d'erreur et l'exposition et les instruments :

1. $\mathbb{E}[\mathcal{Z}^T \boldsymbol{\xi}^{\boldsymbol{Y}}] = 0$,
2. $\mathbb{E}[\mathcal{Z}^T \mathcal{D}] \neq 0$.

Notons par $\mathcal{P}_{\mathcal{Z}}$ la matrice de projection orthogonale dans l'espace engendré par les colonnes de $\mathcal{Z}$,

$$\mathcal{P}_{\mathcal{Z}} = \mathcal{Z}(\mathcal{Z}^T\mathcal{Z})^{-1}\mathcal{Z}^T. \tag{3.4.6}$$

Alors, les valeurs ajustées obtenues par l'estimation de la première étape du modèle 2SLS peuvent s'écrire :

$$\hat{\mathbf{D}} = \mathcal{P}_{\mathcal{Z}}\mathcal{D} + \bar{D}\mathbf{1} = \mathcal{P}_{\mathcal{Z}}\mathbf{D} + \bar{D}\mathbf{1}, \tag{3.4.7}$$

puisque $\mathcal{P}_{\mathcal{Z}}(\mathbf{I}-\mathbf{P_1}) = \mathcal{P}_{\mathcal{Z}}$. Pour la seconde étape, nous remplaçons les valeurs observées $\mathbf{D}$ par les valeurs ajustées $\hat{\mathbf{D}}$ dans l'estimateur des moindres carrés. En utilisant la première ligne de l'équation (3.3.6), mais en remplaçant $\mathbf{D}$ par $\hat{\mathbf{D}}$ nous obtenons :

$$\begin{aligned}
\hat{\beta}^D &= (\hat{\mathbf{D}}^T(\mathbf{I}-\mathbf{P_1})\hat{\mathbf{D}})^{-1}\hat{\mathbf{D}}^T(\mathbf{I}-\mathbf{P_1})\mathbf{Y} \\
&= [(\mathbf{D}^T\mathcal{P}_{\mathcal{Z}} + \bar{D}\mathbf{1}^T)(\mathbf{I}-\mathbf{P_1})(\mathcal{P}_{\mathcal{Z}}\mathbf{D} + \bar{D}\mathbf{1})]^{-1}(\mathbf{D}^T\mathcal{P}_{\mathcal{Z}} + \bar{D}\mathbf{1}^T)(\mathbf{I}-\mathbf{P_1})\mathbf{Y} \\
&= (\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}\mathbf{D})^{-1}\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}(\mathbf{I}-\mathbf{P_1})\mathbf{Y},
\end{aligned} \tag{3.4.8}$$

puisque

$$\mathcal{P}_{\mathcal{Z}}(\mathbf{1}-\mathbf{P_1}) = (\mathbf{1}-\mathbf{P_1})\mathbf{Z}[\mathbf{Z}^T(\mathbf{1}-\mathbf{P_1})\mathbf{Z}]^{-1}\mathbf{Z}^T(\mathbf{1}-\mathbf{P_1})(\mathbf{1}-\mathbf{P_1}) = \mathcal{P}_{\mathcal{Z}}. \tag{3.4.9}$$

L'estimateur 2SLS développé en premier par Henri Theil en 1953 est asymptotiquement sans biais pour les covariables non mesurées dans le modèle (3.4.2) et (3.4.3) et est convergent (p.e. [**9**]).

3.4.3. Inférence pour l'estimateur 2SLS

Soit $\mathbb{E}[\boldsymbol{\xi}^{\boldsymbol{Y}}(\boldsymbol{\xi}^{\boldsymbol{Y}})^T] = \sigma_Y^2\mathbf{I}$. L'estimateur de la variance asymptotique de $\hat{\beta}^D$ est,

$$\hat{V}(\hat{\beta}^D) = \hat{\sigma}^2_{\mathbf{Y}}(\hat{\mathcal{D}}^T\hat{\mathcal{D}})^{-1}. \tag{3.4.10}$$

où $\hat{\sigma}^2_{\mathbf{Y}}$ est l'estimation de la variance de la régression de $\mathbf{Y}$ sur $\hat{\mathbf{D}}$ et non pas celle de $\mathbf{Y}$ sur $\mathbf{D}$.

Nous devons mentionner qu'il serait incorrect d'estimer la variance de l'étape 2 de la méthode 2SLS sans tenir compte de la variabilité générée à l'étape 1, c'est-à-dire de l'estimer avec $\mathcal{D}$ au lieu de $\hat{\mathcal{D}}$. Pour ce faire, il existe des fonctions en **R** qui mettent en œuvre la méthode 2SLS en prenant soin de faire la correction nécessaire : par exemple, la fonction `ivreg` dans le package `{AER}`.

3.4.4. Robustesse de l'estimateur 2SLS au biais de confusion

Nous allons expliquer comment l'estimateur 2SLS est robuste au biais engendré par l'effet de la variable de confusion non mesurée dans le vrai modèle causal si tous les instruments sont valides selon la définition (2.2.1).

Supposons à nouveau sans perte de généralité que $\mathbb{E}(\mathbf{U}) = 0$. Nous avons vu à l'équation (3.4.8) que,

$$\hat{\beta}^D = (\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}\mathbf{D})^{-1}\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}(\mathbf{I}-\mathbf{P_1})\mathbf{Y}.$$

Considérons la dernière partie de cette équation. À partir du modèle (3.4.1),

$$\begin{aligned}\mathcal{P}_{\mathcal{Z}}(\mathbf{I}-\mathbf{P_1})\mathbf{Y} &= \mathcal{P}_{\mathcal{Z}}(\mathbf{I}-\mathbf{P_1})(\pi^Y\mathbf{1}+\mathbf{D}\beta^D+\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\delta^Y\mathbf{U})\\ &= \mathcal{P}_{\mathcal{Z}}(\mathbf{I}-\mathbf{P_1})(\mathbf{D}\beta^D+\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\delta^Y\mathbf{U})\\ &= \beta^D\mathcal{P}_{\mathcal{Z}}\mathbf{D}+\mathcal{P}_{\mathcal{Z}}(\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\delta^Y\mathbf{U}),\end{aligned} \tag{3.4.11}$$

en utilisant (3.4.9).

Ainsi,

$$\begin{aligned}\hat{\beta}^D &= (\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}\mathbf{D})^{-1}\mathbf{D}^T[\beta^D\mathcal{P}_{\mathcal{Z}}\mathbf{D}+\mathcal{P}_{\mathcal{Z}}(\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\delta^Y\mathbf{U})]\\ &= \beta^D+(\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}\mathbf{D})^{-1}\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}(\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\delta^Y\mathbf{U})\\ &= \beta^D+(\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}\mathbf{D})^{-1}(\pi^D\mathbf{1}^T+\boldsymbol{\gamma}^T\mathbf{Z}^T+\boldsymbol{\epsilon}^{\boldsymbol{D}^T}+\delta^D\mathbf{U}^T)\mathcal{P}_{\mathcal{Z}}(\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\delta^Y\mathbf{U})\\ &= \beta^D+(\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}\mathbf{D})^{-1}(\boldsymbol{\gamma}^T\mathcal{Z}^T+\boldsymbol{\epsilon}^{\boldsymbol{D}^T}+\delta^D\mathbf{U}^T)\mathcal{P}_{\mathcal{Z}}(\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\delta^Y\mathbf{U}),\end{aligned} \tag{3.4.12}$$

puisque $\mathcal{P}_{\mathcal{Z}}\mathbf{1} = (\mathbf{1}-\mathbf{P_1})\mathbf{Z}[\mathbf{Z}^T(\mathbf{I}-\mathbf{P_1})\mathbf{Z}]^{-1}\mathbf{Z}^T(\mathbf{1}-\mathbf{P_1})\mathbf{1} = \mathbf{0}$, $\mathcal{P}_{\mathcal{Z}}$ est idempotente et

$$\begin{aligned}\mathbf{Z}^T\mathcal{P}_{\mathcal{Z}} &= \mathbf{Z}^T(\mathbf{1}-\mathbf{P_1})\mathbf{Z}[\mathbf{Z}^T(\mathbf{I}-\mathbf{P_1})\mathbf{Z}]^{-1}\mathbf{Z}^T(\mathbf{1}-\mathbf{P_1})\\ &= \mathbf{Z}^T(\mathbf{1}-\mathbf{P_1})\\ &= \mathcal{Z}^T.\end{aligned} \tag{3.4.13}$$

Nous avons donc,

$$\begin{aligned}\hat{\beta}^D &= \beta^D+(\mathbf{D}^T\mathcal{P}_{\mathcal{Z}}\mathbf{D})^{-1}\Big[\boldsymbol{\gamma}^T\mathcal{Z}^T\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\boldsymbol{\epsilon}^{\boldsymbol{D}^T}\mathcal{P}_{\mathcal{Z}}\boldsymbol{\epsilon}^{\boldsymbol{Y}}+\delta^D\mathbf{U}^T\mathcal{P}_{\mathcal{Z}}\boldsymbol{\epsilon}^{\boldsymbol{Y}}\\ &\quad+\delta^Y\boldsymbol{\gamma}^T\mathcal{Z}^T\mathbf{U}+\delta^Y\boldsymbol{\epsilon}^{\boldsymbol{D}^T}\mathcal{P}_{\mathcal{Z}}\mathbf{U}+\delta^D\delta^Y\mathbf{U}^T\mathcal{P}_{\mathcal{Z}}\mathbf{U}\Big]\\ &= \beta^D+O_p(n^{-1})[T_1+T_2+T_3+T_4+T_5+T_6]\end{aligned} \tag{3.4.14}$$

Maintenant, rappelons que nous avons les relations d'indépendance suivantes :

$$\mathbf{Z} \perp\!\!\!\perp \mathbf{U} \perp\!\!\!\perp \boldsymbol{\epsilon}^{\boldsymbol{D}} \perp\!\!\!\perp \boldsymbol{\epsilon}^{\boldsymbol{Y}}$$

et que $\mathbb{E}(\mathbf{U}) = 0$, par supposition. Ceci implique qu'en moyenne les termes $T1 - T5$ s'annulent. En effet,

$$\begin{aligned}
\mathbb{E}[T_1] &= \mathbb{E}(\boldsymbol{\gamma}^T \mathcal{Z}^T \boldsymbol{\epsilon}^{\boldsymbol{Y}}) = 0, \\
\mathbb{E}[T_2] &= \mathbb{E}(\boldsymbol{\epsilon}^{\boldsymbol{D}} \mathcal{P}_{\mathcal{Z}} \boldsymbol{\epsilon}^{\boldsymbol{Y}}) = 0, \\
\mathbb{E}[T_3] &= \mathbb{E}(\delta^D \mathbf{U}^T \mathcal{P}_{\mathcal{Z}} \boldsymbol{\epsilon}^{\boldsymbol{Y}}) = 0, \\
\mathbb{E}[T_4] &= \mathbb{E}(\delta^Y \boldsymbol{\gamma}^T \mathcal{Z}^T \mathbf{U}) = 0, \\
\mathbb{E}[T_5] &= \mathbb{E}(\delta^Y \boldsymbol{\epsilon}^{\boldsymbol{D}} \mathcal{P}_{\mathcal{Z}} \mathbf{U}) = 0.
\end{aligned} \tag{3.4.15}$$

De plus,

$$\mathbb{E}[T_6] = \mathbb{E}(\delta^D \delta^Y \mathbf{U}^T \mathcal{P}_{\mathcal{Z}} \mathbf{U}) = \delta^D \delta^Y \mathbb{E}(\mathbf{U}^T \mathcal{P}_{\mathcal{Z}} \mathbf{U}),$$

et cette espérance se calcule comme suit :

$$\begin{aligned}
\mathbb{E}(\mathbf{U}^T \mathcal{P}_{\mathcal{Z}} \mathbf{U}) &= \mathbb{E}\Big[\mathbb{E}((\mathbf{U}^T \mathcal{P}_{\mathcal{Z}} \mathbf{U}) | \mathcal{Z})\Big] \\
&= \mathbb{E}\Big[tr(\mathcal{P}_{\mathcal{Z}} var(\mathbf{U})) + \mathbb{E}(\mathbf{U}) \mathcal{P}_{\mathcal{Z}} \mathbb{E}(\mathbf{U})\Big] \\
&= \mathbb{E}[tr(\mathcal{P}_{\mathcal{Z}} \sigma_U^2 \mathbf{I})] \\
&= L\sigma_U^2,
\end{aligned} \tag{3.4.16}$$

une constante, où L est le nombre de colonnes de la matrice $\mathcal{Z}$ qui contient les observations centrées de la matrice des instruments $\mathbf{Z}$. Donc le biais de $\hat{\beta}^D$ en (3.4.14) s'amenuise avec n.

Nous pouvons ainsi comprendre que, dans des grands échantillons, l'estimateur 2SLS contourne le problème du biais de confusion auquel l'estimateur des moindres carrés ordinaires n'est pas robuste.

3.4.5. Impact de l'utilisation d'instruments invalides sur l'estimateur 2SLS

Nous allons expliquer la limite de l'estimateur 2SLS pour estimer l'effet causal d'une exposition $\mathbf{D}$ sur une issue $\mathbf{Y}$ en randomisation mendélienne lorsque un ou plusieurs des SNPs utilisés a un effet pléiotropique impliquant le phénotype issue. En d'autres termes, nous entendons par pléiotropie qu'au moins un des SNPs de $\mathbf{Z}$ a une association avec $\mathbf{Y}$ qui ne passe pas uniquement par $\mathbf{D}$. Cette relation est illustrée par la présence du terme $\mathbf{Z}\boldsymbol{\alpha}$ dans le modèle d'estimation de l'issue en (3.4.3). Notons que les SNPs qui n'ont pas d'effet pléiotropique avec $\mathbf{Y}$ auront une valeur α égale à 0.

Rappelons que le modèle pour l'issue qui inclut l'effet pléiotropique des SNPs est

$$\mathbf{Y} = \mu^Y \mathbf{1} + \mathbf{D}\beta^D + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}^{\boldsymbol{Y}} + \mathbf{U}\delta^Y, \tag{3.4.17}$$

c'est-à-dire que nous avons additionné $\mathbf{Z}\boldsymbol{\alpha}$ à $\mathbf{Y}$. Ce $\mathbf{Z}\boldsymbol{\alpha}$ est additionné à la parenthèse de droite de (3.4.11) de telle sorte qu'à partir de l'équation (3.4.14) l'estimateur 2SLS de

l'effet causal de $\mathbf{D}$ sur $\mathbf{Y}$ pour le modèle avec instruments invalides peut s'écrire de la façon suivante :

$$\begin{aligned}\hat{\beta}^D &= \beta^D + O_p(n^{-1})[\sum_{i=1}^{6} T_i + \boldsymbol{\gamma}^T \mathcal{Z}^T \mathcal{P}_{\mathcal{Z}} \mathcal{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}^{\boldsymbol{D}^T} \mathcal{P}_{\mathcal{Z}} \mathcal{Z}\boldsymbol{\alpha} + \delta^D \mathbf{U}^T \mathcal{P}_{\mathcal{Z}} \mathcal{Z}\boldsymbol{\alpha}] \\ &= \beta^D + O_p(n^{-1})[\sum_{i=1}^{6} T_i + T_7 + T_8 + T_9] \\ &= \beta^D + O_p(n^{-1}) + O_p(1)\end{aligned} \tag{3.4.18}$$

puisque

$$\begin{aligned}\mathbb{E}[T_7] &= \mathbb{E}(\boldsymbol{\gamma}^T \mathcal{Z}^T \mathcal{P}_{\mathcal{Z}} \mathcal{Z}\boldsymbol{\alpha}) = \mathbb{E}(\boldsymbol{\gamma}^T \mathcal{Z}^T \mathcal{Z}\boldsymbol{\alpha}) = O_p(1), \\ \mathbb{E}[T_8] &= \mathbb{E}(\boldsymbol{\epsilon}^{\boldsymbol{D}^T} \mathcal{P}_{\mathcal{Z}} \mathcal{Z}\boldsymbol{\alpha}) = \mathbb{E}(\boldsymbol{\epsilon}^{\boldsymbol{D}^T} \mathcal{Z}\boldsymbol{\alpha}) = 0, \\ \mathbb{E}[T_9] &= \mathbb{E}(\delta^D \mathbf{U}^T \mathcal{P}_{\mathcal{Z}} \mathcal{Z}\boldsymbol{\alpha}) = \mathbb{E}(\delta^D \mathbf{U}^T \mathcal{Z}\boldsymbol{\alpha}) = 0,\end{aligned} \tag{3.4.19}$$

et parce que $(\mathcal{D}^T \mathcal{P}_{\mathcal{Z}} \mathcal{D})^{-1} = O_p(n^{-1})$, $\mathcal{D}^T \mathcal{Z}\boldsymbol{\alpha} = O_p(n)$ et que nous avons l'indépendance entre le terme d'erreur de l'exposition $\boldsymbol{e}^{\boldsymbol{D}}$ et la matrice d'instruments $\mathbf{Z}$, ainsi que l'indépendance entre la variable de confusion $\mathbf{U}$ et $\mathbf{Z}$, en plus des suppositions que $\mathbb{E}[\boldsymbol{e}^{\boldsymbol{D}}] = 0$ et $\mathbb{E}[\mathbf{U}] = 0$. Donc le biais ne s'amenuise pas en augmentant n [**9**].

Ceci implique que lorsque nous nous retrouvons dans le cas où certains SNPs pourraient avoir un effet pléiotropique impliquant l'issue, il faut avoir recours à une méthode autre que la méthode 2SLS. Ce résultat est la motivation principale pour utiliser une autre méthode d'estimation qui sera à la fois robuste au biais engendré par une variable de confusion non mesurée et à la présence du phénomène de pléiotropie dans les variables du vrai modèle causal. Une telle méthode est présentée dans la section suivante.

3.5. Estimateur sisVIVE

Face à la difficulté de trouver des intruments valides selon (A1)-(A3), des chercheurs ont investigué des méthodes d'estimation dans le cas où certains instruments seraient invalides (p.e. [**23**] et [**5**]). L'estimateur sisVIVE (*some invalid some valid instrumental variables estimator*) proposé par Hyunseung Kang et al. en 2014 en est un exemple et sera élaboré plus en détail dans cette section.

Dans la section qui suit, nous rappelons d'abord l'estimateur LASSO classique, puisqu'il est utilisé dans l'algorithme de l'estimateur sisVIVE, puis nous présentons la contribution de l'estimateur sisVIVE au problème d'estimation de l'effet causal d'une exposition $\mathbf{D}$ sur une issue $\mathbf{Y}$ en randomisation mendélienne. Enfin, nous présentons les limites de l'estimateur sisVIVE.

3.5.1. LASSO classique et validation croisée K-fold

Introduisons d'abord la notation qui sera utilisée pour expliquer les estimateurs LASSO et éventuellement sisVIVE. Supposons que nous avons le modèle de régression linéaire multiple,

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3.5.1}$$

En estimant ce modèle, nous chercherons à réduire la taille des coefficients $\boldsymbol{\beta}$ lorsque ceux-ci sont trop grands. Ainsi, la constante n'est pas d'intérêt pour le problème d'estimation. Nous pouvons donc réécrire le modèle ci-haut en utilisant la notation pour les données centrées par leur moyenne présentée précédemment :

$$\mathcal{Y} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3.5.2}$$

Pour estimer ce modèle, nous allons devoir minimiser la somme des résidus au carré. Avant de procéder, définissons d'abord les normes $||\cdot||_1$ et $||\cdot||_2$ qui seront utilisées pour expliquer le problème de minimisation pour l'estimateur LASSO et pour l'estimateur sisVIVE. Soit par exemple le vecteur $\mathbf{V} = (V_1, \ldots, V_n)^T$. Alors, nous définissons les normes $||\cdot||_1$ et $||\cdot||_2$ comme suit :

$$||\mathbf{V}||_1 = \sum_{i=1}^{n} |V_i| \tag{3.5.3}$$

$$||\mathbf{V}||_2 = \left[\sum_{i=1}^{n} V_i^2\right]^{\frac{1}{2}}. \tag{3.5.4}$$

Nous pouvons ensuite écrire le problème de minimisation de la somme des résidus au carré :

$$\begin{aligned} ||\mathcal{Y} - \mathcal{X}\boldsymbol{\beta}||_2^2 &= ||\mathcal{Y} + \mathcal{P}_{\mathcal{X}}\mathcal{Y} - \mathcal{P}_{\mathcal{X}}\mathcal{Y} - \mathcal{X}\boldsymbol{\beta}||_2^2 \\ &= ||(\mathbf{I} - \mathcal{P}_{\mathcal{X}})\mathcal{Y} + \mathcal{P}_{\mathcal{X}}(\mathcal{Y} - \mathcal{X}\boldsymbol{\beta})||_2^2 \\ &= ||(\mathbf{I} - \mathcal{P}_{\mathcal{X}})\mathcal{Y}||_2^2 + ||\mathcal{P}_{\mathcal{X}}(\mathcal{Y} - \mathcal{X}\boldsymbol{\beta})||_2^2, \end{aligned} \tag{3.5.5}$$

puisque $(\mathbf{I} - \mathcal{P}_{\mathcal{X}})\mathcal{P}_{\mathcal{X}} = 0$, $\mathcal{Y}^T(\mathbf{I} - \mathcal{P}_{\mathcal{X}})\mathcal{P}_{\mathcal{X}}(\mathcal{Y} - \mathcal{X}\boldsymbol{\beta}) = 0$. Ainsi,

$$\arg\min_{\boldsymbol{\beta}} ||\mathcal{Y} - \mathcal{X}\boldsymbol{\beta}||_2^2 = \arg\min_{\boldsymbol{\beta}} ||\mathcal{P}_{\mathcal{X}}(\mathcal{Y} - \mathcal{X}\boldsymbol{\beta})||_2^2. \tag{3.5.6}$$

Nous allons exprimer la somme des résidus au carré des estimateurs LASSO et sisVIVE sous cette forme dérivée pour la suite.

Maintenant, le *least absolute shrinkage and selection operator* (LASSO) [**38**] s'utilise dans le contexte de la régression linéraire multiple pour réduire la dimension d'un modèle et améliorer sa prédiction de la variable dépendante du modèle [**17**]. L'estimateur LASSO

peut s'écrire de la façon suivante :

$$\arg\min_{\boldsymbol{\beta}} ||\mathcal{P}_{\mathcal{X}}(\mathcal{Y} - \mathcal{X}\boldsymbol{\beta})||_2^2 \quad s.à. \quad ||\boldsymbol{\beta}||_1 \leq t, \tag{3.5.7}$$

où $\boldsymbol{\beta} = (\beta_1,...,\beta_p)^T$ et $t \geq 0$ est une constante. L'idée générale est de minimiser la somme des résidus au carré sous une contrainte imposée sur la norme $\boldsymbol{l}_1 = ||\cdot||_1$ calculée sur tous les paramètres du modèle. Certains coefficients seront pénalisés au point d'être réduits à zéro et ainsi exclus du modèle. Nous pouvons ainsi constater que la norme $\boldsymbol{l}_1$ fait en même temps une sélection de variables et qu'il est important que les variables indépendantes soient sur des échelles comparables. Pour plus de détails voir [**38**].

Nous avons mentionné qu'une des utilités du LASSO est d'améliorer la prédiction de la variable dépendante, $\mathcal{Y}$. Pour voir comment cela peut être fait, considérons d'abord que le modèle pour $\mathcal{Y}$ peut s'écrire comme suit :

$$\mathcal{Y}_i = \eta(\mathcal{X}_i) + \epsilon_i, \tag{3.5.8}$$

où $\mathbb{E}(\boldsymbol{\epsilon}) = 0$, $var(\boldsymbol{\epsilon}) = \sigma^2$ et $\eta(\mathcal{X}_i) = \mathcal{X}_i^T\boldsymbol{\beta}$ est la fonction que nous voulons prédire avec la plus faible erreur de prévision possible. Soit $\hat{\eta}_t$ une estimation fixe de η qui dépend du choix de t, le paramètre de régularisation.

Nous pouvons alors exprimer l'erreur de prévision moyenne du modèle pour $\mathcal{Y}$ comme suit :

$$PE = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\{\mathcal{Y}_i' - \hat{\eta}_t(\mathcal{X}_i)\}^2\Big], \tag{3.5.9}$$

où $\mathcal{Y}_i' = \mathcal{X}_i\boldsymbol{\beta} + \epsilon_i'$ est un nouvel échantillon, indépendant de $\mathcal{Y}_i$, basé sur les observations originales $\mathcal{X}_i$ [**13**]. Ainsi, l'objectif devient de sélectionner le paramètre de régularisation t qui minimise l'erreur PE. Si nous avions un échantillon indépendant $\mathcal{Y}_i'$, nous pourrions estimer l'erreur de prévision du modèle par :

$$e(\hat{\eta}_t) = \frac{1}{n}\sum_{i=1}^{n}\{\mathcal{Y}_i' - \hat{\eta}_t(\mathcal{X}_i)\}^2. \tag{3.5.10}$$

Toutefois, nous n'avons pas un tel échantillon à notre disposition. Par contre, nous pouvons obtenir des échantillons indépendants à partir de notre échantillon original. En effet, nous pouvons d'abord partitionner notre échantillon original en K parties essentiellement égales. Ensuite, nous pourrons en réserver une partie k pour la validation du modèle et consacrer les $K-1$ autres parties à l'entraînement du modèle. L'entraînement du modèle se fait en calculant $\hat{\eta}_t^{-k}(\mathcal{X}_i)$ basé sur les données issues des $K-1$ autres parties. La validation du modèle se fait en comparant la prédiction des données à valider $\mathcal{Y}_i$ issues de la partie k, qui sont alors indépendantes de celles utilisées pour l'entraînement, par rapport aux prédictions calculées sur les données d'entraînement $\hat{\eta}_t^{-k}(\mathcal{X}_i)$. Nous obtiendrons alors une

estimation de l'erreur de prévision $e_k(t)$. Puis, nous pourrons itérer sur $k = 1, \ldots, K$ pour une grille de valeurs possibles pour t. Cette procédure, que nous allons détailler par la suite, s'appelle la validation croisée K-fold.

L'algorithme pour choisir le paramètre de régularisation t optimal est le suivant [**39**] :

(1) Diviser l'échantillon $s = (\mathcal{Y}_i, \mathcal{X}_i)$, $i = 1, \ldots, n$, en K parties essentiellement égales. Notons ces parties $M_1, \ldots, M_K$.

(2) Pour $k = 1, \ldots, K$, entraîner le modèle sur $(\mathcal{Y}_i, \mathcal{X}_i)$, $i \notin M_k$ et valider le modèle avec $(\mathcal{Y}_i, \mathcal{X}_i)$, $i \in M_k$. Si M_{-k} sont les données avec la k^e partie exclue, alors pour chaque t,

(a) Calculer $\hat{\eta}_t^{-k}$ à partir des données dans M_{-k}.

(b) Estimer l'erreur de prévision dans la partie M_k :

$$e_k(t) = \sum_{i \in M_k} \{\mathcal{Y}_i - \hat{\eta}_t^{-k}(\mathcal{X}_i)\}^2. \tag{3.5.11}$$

(3) Pour chaque t, calculer la moyenne des estimations e_k sur les K parties:

$$CV(t) = \frac{1}{K} \sum_{k=1}^{K} e_k(t). \tag{3.5.12}$$

(4) Le t qui minimise l'erreur de prévision est alors:

$$t_{opt} = \arg\min_{\boldsymbol{t}} CV(t). \tag{3.5.13}$$

La valeur de t obtenue à l'étape (4) est le paramètre de régularisation utilisé dans l'estimation de l'effet d'intérêt $\boldsymbol{\beta}$.

3.5.2. Estimateur sisVIVE

L'estimateur sisVIVE prend le modèle 2SLS avec la modélisation de l'invalidité des instruments,

$$\mathcal{Y} = \mathcal{Z}\boldsymbol{\alpha} + \mathcal{D}\beta^D + \boldsymbol{\epsilon} \tag{3.5.14}$$

et fait la sélection des coefficients du paramètre de validité des instruments $\boldsymbol{\alpha}$. Tout comme l'estimateur LASSO peut s'écrire selon (3.5.7), l'estimateur sisVIVE peut s'exprimer de la façon suivante :

$$(\hat{\boldsymbol{\alpha}}, \hat{\beta}) = \arg\min_{\boldsymbol{\alpha}, \beta^D} ||\mathcal{P}_{\mathcal{Z}}(\mathcal{Y} - \mathcal{Z}\boldsymbol{\alpha} - \mathcal{D}\beta^D)||_2^2 \quad s.à. \quad ||\boldsymbol{\alpha}||_1 \leq t, \tag{3.5.15}$$

Incidemment, l'estimateur sisVIVE correspond à un estimateur 2SLS avec une pénalisation du type $\boldsymbol{l}_1$ sur les coefficients des instruments. Il est important de remarquer que sisVIVE diffère de l'estimateur LASSO classique parce que le paramètre β^D n'est pas pénalisé alors que le LASSO pénalise tous les paramètres du modèle. La pénalisation de type LASSO est

appropriée dans le contexte de la randomisation mendélienne : les instruments prennent tous des valeurs de 0, 1 ou 2 puisque ce sont des SNPs, donc ces variables sont sur des échelles comparables.

Tel que vu dans l'estimateur LASSO, la pénalisation utilisée par sisVIVE permet de réduire certains coefficients α_j au point d'en exclure certains du modèle en imposant un seuil sur la norme $\boldsymbol{l}_1$ du vecteur $\boldsymbol{\alpha}$. Une définition de la contrainte sur les coefficients de $\boldsymbol{\alpha}$ est :

$$\sum_j |\alpha_j| = |\alpha_1| + \ldots + |\alpha_L| \leq t, \tag{3.5.16}$$

pour un seuil $t \geq 0$. Soit $\hat{\alpha}_j^0$ l'estimation du coefficient α_j par la méthode des moindres carrés. Alors, si nous définissons $t_0 = \sum_j |\hat{\alpha}_j^0|$, tout paramètre $t < t_0$ réduit ou égale à 0 certains coefficients dans l'inéquation (3.5.16). Le problème du choix de t est résolu par la validation croisée K-fold que nous avons expliqué précédemment.

Kang et al. (2014) ont établit que (1) l'estimateur sisVIVE admet une solution unique lorsque moins de 50% des instruments sont invalides, et (2) par des simulations, l'estimateur sisVIVE peut estimer l'effet causal lorsque moins de 50% des instruments sont invalides. En effet, dans les simulations où moins de 50% des instruments sont invalides, la performance de sisVIVE se rapproche de celle du 2SLS basé uniquement sur les instruments valides, ce qui est possible de faire dans une simulation, mais qui serait impossible de faire avec un vrai jeu de données. Pour plus de détails sur les simulations et le théorème d'unicité de la solution, voir [**23**].

3.5.3. Limites de l'estimateur sisVIVE

À notre connaissance, aucune méthode pour calculer des intervalles de confiance n'a été proposée pour l'estimateur sisVIVE, entre autres, parce que sa distribution asymptotique n'est pas connue. Ceci motive l'exploration du bootstrap pour calculer des intervalles de confiance pour l'estimateur sisVIVE. Accomplir cette tâche est la contribution principale de ce mémoire à la littérature.

La technique de la randomisation mendélienne, en plus de reposer sur des suppositions non vérifiables dans les données, présente de nombreux défis. Tel qu'indiqué dans les chapitres précédents, le défi le plus important à relever dans la randomisation mendélienne est la gestion de la présence de l'effet pléotropique d'un gène, qui est un phénomène génétique, parce qu'elle implique que l'instrument génétique est invalide. Ensuite, nous avons démontré que l'estimateur 2SLS est susceptible d'engendrer un biais dans le cas où au moins un des instruments utilisés a un effet pléiotropique impliquant l'issue dans le modèle, tandis que l'estimateur sisVIVE semble être robuste au biais engendré par l'utilisation de SNPs pléiotropiques si la proportion de SNPs pléiotropiques est inférieure

ou égale à 0,5. Néanmoins, l'estimateur sisVIVE reste susceptible à d'autres sources de biais. Enfin, nous avons souligné une des limites importantes de l'estimateur sisVIVE, soit l'absence d'un estimateur pour sa variance afin de mesurer sa variabilité et plus encore, construire un intervalle de confiance.

Chapitre 4

L'APPROXIMATION BOOTSTRAP POUR L'ALGORITHME SISVIVE

Rappelons que l'estimateur sisVIVE est particulièrement pertinent pour les études de randomisation mendélienne puisqu'il pardonne, jusqu'à un seuil de 50%, l'invalidité d'instruments qui survient souvent dans l'estimation de l'effet causal d'une exposition sur une issue. Cependant, à notre connaissance, il n'y a aucun intervalle de confiance analytique proposé pour cet estimateur dans la littérature. L'intervalle de confiance nous permet non seulement de déterminer si l'effet causal est non nul, c'est-à-dire statistiquement significatif, mais aussi de mesurer la précision de son estimation. Par exemple, pour un seuil de 5%, l'intervalle de confiance de niveau 95% représente l'intervalle dans lequel le vrai effet causal devrait se retrouver, 19 fois sur 20. Ceci implique que plus l'intervalle est petit, plus notre estimation est précise.

Une raison possible pour l'absence d'intervalles de confiance pour sisVIVE est le fait que la distribution asymptotique de l'estimateur est inconnue. Dans ce chapitre, nous présentons une méthode bootstrap dans le but de calculer des intervalles de confiance pour sisVIVE.

Afin d'introduire le bootstrap non paramétrique, nous décrivons d'abord son application pour l'estimation de la variance et des quantiles de la distribution d'une moyenne. Par la suite, nous traduisons l'approche dans le contexte de sisVIVE. Finalement, nous présentons un algorithme de bootstrap adapté au problème d'inférence de sisVIVE. Le code **R** pour mettre en œuvre le bootstrap est disponible dans l'annexe A.

4.1. Illustration du bootstrap par l'exemple de la moyenne

Le bootstrap est une méthode pour estimer la distribution de l'estimateur d'un paramètre du « monde réel » dans un contexte où un échantillon de données $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de vecteurs indépendants et identiquement distribués de dimension p est issu d'une fonction

de répartition $F \in \mathbb{R}^p$ inconnue. Si la fonction F était connue, il suffirait alors de simuler des échantillons à partir de cette dernière pour obtenir une approximation de la variance $var_F(\hat{\theta})$. Or, puisque F est inconnue elle doit être estimée, et dans le cas du bootstrap non paramétrique, la fonction de répartition expérimentale $\hat{F}_n$ sert à estimer F. Ainsi, en procédant par des simulations d'échantillons obtenus à partir de $\hat{F}_n$ nous parvenons à faire de l'inférence pour le paramètre d'intérêt θ sur la base de son estimateur $\hat{\theta}$, par exemple, en estimant $var_F(\hat{\theta})$ dans le « monde bootstrap » (p.e. [**13**], [**26**]).

4.1.1. Dérivation des intervalles de confiance bootstrap pour la moyenne

Nous allons utiliser l'exemple de la moyenne pour illustrer le bootstrap dans un cas simple. Supposons, dans un premier temps, que nous avons un échantillon de variables indépendamment et identiquement distribuées $X_1, \ldots, X_n \sim F \in \mathbb{R}$ connue et que nous voulions faire de l'inférence sur la moyenne des observations X_i, soit

$$\theta = \mathbb{E}_F(X), \tag{4.1.1}$$

que nous estimons comme suit :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{4.1.2}$$

où $\bar{X}$ est l'estimateur du paramètre d'intérêt et n est la taille de l'échantillon. Lorsque F est connue, nous pouvons approximer la distribution de $\hat{\theta}$ sous F par simulation :

`Pour` $i = 1, \ldots, B$ `:`

`1. Générer` $X_1, \ldots, X_n$ `i.i.d.` F`,`

`2. Calculer` $\hat{\theta}_i$ `à partir des données` $X_1, \ldots, X_n$`.`

La distribution expérimentale des estimateurs $\hat{\theta}_i$ est une approximation de la distribution de $\hat{\theta}$. Ainsi, nous pouvons approximer la variance de l'estimation de la moyenne sous F comme suit

$$var_F(\hat{\theta}) \approx \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i - \bar{\hat{\theta}})^2, \tag{4.1.3}$$

où $\bar{\hat{\theta}} = B^{-1}\sum_{i=1}^{B}\hat{\theta}_i$ est l'approximation de la moyenne de F.

Si F était connue, nous pourrions construire des intervalles de confiance basés sur les quantiles de la distribution centrée de l'estimateur, c'est-à-dire de la quantité $\hat{\theta} - \theta$. À cette fin, définissons la fonction de répartition de $\hat{\theta} - \theta$ basée sur un échantillon de taille n provenant de F par :

$$J_n(x; F) = P_F\{\hat{\theta} - \theta \leq x\}. \tag{4.1.4}$$

La variance de J_n est alors égale à la variance de $\hat{\theta}$, soit $var_F(\hat{\theta})$. Définissons également les statistiques d'ordre des estimés $\hat{\theta}_i$ par $\hat{\theta}_{(1)} \leq \hat{\theta}_{(2)} \leq \ldots \leq \hat{\theta}_{(B)}$. Alors, les quantiles de J_n pour une probabilité ρ sont approximés de la façon suivante :

$$J_n^{-1}(\rho; F) \approx \hat{\theta}_{(B\rho)} - \theta, \tag{4.1.5}$$

où $(B\rho)$ signifie la statistique d'ordre $B\rho$. Ainsi, un intervalle de confiance exact de niveau $(1-\rho)$ pour la moyenne θ basé sur la statistique de test $\hat{\theta} - \theta$ est obtenu par,

$$\begin{aligned} 1 - \rho &= P_F\{J_n^{-1}(\rho/2; F) \leq \hat{\theta} - \theta \leq J_n^{-1}(1-\rho/2; F)\} \\ &= P_F\{\hat{\theta} - J_n^{-1}(1-\rho/2; F) \leq \theta \leq \hat{\theta} - J_n^{-1}(\rho/2; F)\}, \end{aligned} \tag{4.1.6}$$

donc l'intervalle s'écrit :

$$[\hat{\theta} - J_n^{-1}(1-\rho/2; F), \qquad \hat{\theta} - J_n^{-1}(\rho/2; F)]. \tag{4.1.7}$$

Nous pourrions alors procéder à l'inférence pour la moyenne θ basé sur son estimateur $\hat{\theta}$.

Toutefois, dans bien des situations, la distribution F n'est pas connue, de telle sorte que cette approche ne peut pas être utilisée. Par contre, si nous estimons la distribution F et procédons par simulation à partir d'échantillons provenant de cette estimation, nous pourrons alors utiliser cette approche. C'est ce que nous appelons le bootstrap. Ainsi, nous allons d'abord estimer F de façon non paramétrique par la fonction de répartition expérimentale,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x), \tag{4.1.8}$$

qui est sans biais pour F. Ensuite, nous pouvons procéder par simulation comme nous l'aurions fait avec F :

`Pour` $i = 1, \ldots, B$:

1. `Générer` $X_1^*, \ldots, X_n^*$ `i.i.d.` $\hat{F}_n$,

2. `Calculer` $\hat{\theta}_i^*$ `à partir des données` $X_1^*, \ldots, X_n^*$,

où * signifie rééchantillonné à partir de l'échantillon original. En effet, nous pouvons distinguer les observations originales $X_1, \ldots, X_n$ qui ont été observées des observations bootstrap $X_1^*, \ldots, X_n^*$ qui ont été générées sur l'ordinateur. Ainsi, la distribution des $\hat{\theta}_i^*$ est la distribution de l'estimateur $\hat{\theta}$, mais lorsque les observations proviennent de la distribution expérimentale $\hat{F}_n$ plutôt que de la vraie distribution inconnue F. Remarquons que chaque échantillon bootstrap $X_1^*, \ldots, X_n^*$ i.i.d. $\hat{F}_n$ correspond à un échantillon aléatoire simple avec remise tiré des données originales $X_1, \ldots, X_n$. Pour comprendre ce fait, notons que $\hat{F}_n$ attribue des probabilités de $1/n$ également à chacune des observations des données

originales $X_1, \ldots, X_n$. Chaque observation X_i^* est échantillonnée aléatoirement de façon indépendante à partir de ces données. Ceci implique que les échantillons bootstrap sont des échantillons aléatoires simples tirés avec remise des données originales. Cette procédure est ce que nous appelons le bootstrap non paramétrique [**10**].

Pour estimer la variance de la moyenne, nous procédons comme nous l'aurions fait à partir d'estimateurs calculés sur des échantillons provenant de F, mais en utilisant plutôt des échantillons bootstrap provenant de $\hat{F}_n$. À partir des échantillons bootstrap $X_1^*, \ldots, X_n^*$ nous calculons les réplications bootstrap $\hat{\theta}_i^*$, $\bar{X}_i^*$, puis nous estimons la variance par :

$$\begin{aligned} \hat{var}_F(\hat{\theta}) &= var_{\hat{F}_n}(\hat{\theta}^*) \\ &= \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_i^{\,*} - \overline{\hat{\theta}^*})^2, \end{aligned} \tag{4.1.9}$$

où $\overline{\hat{\theta}^*}$ est la moyenne des estimés bootstrap $\hat{\theta}_i^*$.

Notons que θ est la valeur du paramètre lorsque la distribution est F, soit, dans le « monde réel », et que lorsque la distribution est $\hat{F}_n$ dans le « monde bootstrap » le paramètre coincide avec l'estimateur de la moyenne. Nous avons donc $\theta_n^* = \hat{\theta}_n$ au tableau 4.1. Considérons $\theta = \mathbb{E}_F[X]$. Alors,

$$\theta^* = \mathbb{E}_{\hat{F}_n}[X^*] = \sum_{i=1}^{n} \frac{1}{n} X_i = \bar{X} = \hat{\theta} \tag{4.1.10}$$

pour notre exemple de la moyenne [**10**]. De façon générale, nous utilisons $\theta^* = \hat{\theta}$, soit la valeur de l'estimateur calculé sur l'échantillon original.

Si F était connue, nous pourrions construire des intervalles de confiance basés sur les quantiles de la distribution centrée de l'estimateur, c'est-à-dire de la quantité $\hat{\theta} - \theta$. Puisque F est inconnue, définissons plutôt la fonction de répartition de $\hat{\theta}^* - \hat{\theta}$ basée sur un échantillon de taille n provenant de $\hat{F}_n$ par

$$J_n(x; \hat{F}_n) = P_{\hat{F}_n}\{\hat{\theta}^* - \hat{\theta} \leq x\}. \tag{4.1.11}$$

Jusqu'à présent, nous disposons d'un estimateur bootstrap de la variance de $\hat{\theta}$ duquel nous pourrions, dès lors, construire un intervalle de confiance basé sur la normalité présumée de l'estimateur. Par contre, si nous définissons en plus un estimateur bootstrap des quantiles de la distribution de l'estimateur nous pourrions alors construire d'autres types d'intervalles de confiance bootstrap. Rappelons que nous avons défini la fonction de répartition de $\hat{\theta}^* - \hat{\theta}$ par $J_n(x; \hat{F}_n)$ en (4.1.11). Les quantiles de la fonction de répartition $J_n(\cdot; F)$ sont estimés par

$$J_n^{-1}(\rho; \hat{F}_n) \approx \hat{\theta}^*_{(B\rho)} - \hat{\theta}, \tag{4.1.12}$$

pour une probabilité ρ.

TABLEAU 4.1. Composantes pour l'inférence d'un paramètre dans le monde réel et leur estimation dans le monde bootstrap. La colonne de droite présente l'analogue des éléments de la colonne de gauche.

	monde réel	monde bootstrap
Fonction de répartition des données réelles	F	$\hat{F}_n$
Données réelles	$\mathbf{X} \sim F$	$\mathbf{X}^* \sim \hat{F}_n$
Fonction de répartition expérimentale	$\hat{F}_n$	$\hat{F}_n^*$
Paramètre d'intérêt	θ	$\theta_n^* = \hat{\theta}_n$
Estimation	$\hat{\theta}_n$	$\hat{\theta}_n^*$
Variance	$var_F(\hat{\theta})$	$var_{\hat{F}_n}(\hat{\theta}^*)$

Nous avons maintenant les éléments nécessaires à la construction de différents types d'intervalles de confiance bootstrap. D'abord, supposons que la fonction de répartition de l'estimateur de la moyenne suit une loi normale. Alors, nous pouvons construire un intervalle de confiance en utilisant simplement les quantiles de la loi normale et l'estimateur bootstrap de la variance de $\hat{\theta}$. Ainsi, un intervalle de confiance bootstrap normal de niveau approximatif $(1-\rho)$ s'écrit :

$$\Big[\hat{\theta} - z_{(1-\rho/2)}\sqrt{var_{\hat{F}_n}(\hat{\theta}^*)}, \qquad \hat{\theta} - z_{\rho/2}\sqrt{var_{\hat{F}_n}(\hat{\theta}^*)}\Big]. \tag{4.1.13}$$

Ensuite, si nous ne voulons pas faire l'approximation normale, nous pouvons utiliser les quantiles estimés sous $\hat{F}_n$. Une première façon de faire résulte en l'intervalle de confiance bootstrap de base. Nous le calculons de la même façon que l'intervalle de confiance exact, mais en utilisant les quantiles estimés $J_n^{-1}(\rho,\hat{F}_n)$,

$$\Big[\hat{\theta} - J_n^{-1}(1-\rho/2;\hat{F}_n), \qquad \hat{\theta} - J_n^{-1}(\rho/2;\hat{F}_n)\Big], \tag{4.1.14}$$

ce qui implique,

$$\Big[\hat{\theta} - (\hat{\theta}^*_{(B-B\rho/2)} - \hat{\theta}), \qquad \hat{\theta} - (\hat{\theta}^*_{(B\rho/2)} - \hat{\theta})\Big], \tag{4.1.15}$$

où le quantile du haut de la distribution des $\hat{\theta}^*_{(i)} - \hat{\theta}$ est soustrait à gauche, et le quantile du bas est soustrait à droite. Notons que l'intervalle est approximatif parce qu'il se base sur la fonction de répartition expérimentale $\hat{F}_n$ en lieu de F. Donc la probabilité de couverture n'est pas exactement $(1-\rho)$.

Une deuxième façon de faire résulte en l'intervalle de confiance bootstrap percentile. Nous le calculons en additionnant le quantile du haut à droite et le quantile du bas à gauche comme suit :

$$\Big[\hat{\theta} + J_n^{-1}(\rho/2;\hat{F}_n), \qquad \hat{\theta} + J_n^{-1}(1-\rho/2;\hat{F}_n)\Big], \tag{4.1.16}$$

ce qui implique,

$$\left[\hat{\theta}^*_{(B\rho/2)}, \qquad \hat{\theta}^*_{(B-B\rho/2)}\right]. \tag{4.1.17}$$

Remarquons que ceci correspond à prendre les statistiques bootstrap les plus centrales. Par conséquent, lorsque la fonction de répartition $J_n(\cdot; F)$ est asymétrique dans les cas d'échantillons finis, nous pouvons nous attendre à ce que l'intervalle de confiance bootstrap percentile diffère grandement de l'intervalle de confiance bootstrap de base. Par contre, en supposant la convergence de l'estimateur $\hat{\theta}$ et celle de la statistique de test centrée, $\hat{\theta} - \theta$ vers une distribution symétrique autour de zéro, alors les quantiles obtenus par $\hat{\theta}^*_{(B\rho)} - \hat{\theta}$ convergent eux aussi vers une distribution symétrique autour de zéro. Ainsi, les intervalles de confiance bootstrap de niveau $(1 - \rho)$ approximatif construits de ces deux façons différentes sont asymptotiquement équivalents.

4.2. Bootstrap pour l'algorithme sisVIVE

Dans cette section, nous décrivons l'approche bootstrap pour sisVIVE et présentons l'algorithme adapté pour calculer (1) des intervalles de confiance normal, de base, et percentile pour l'effet causal d'une exposition D sur une issue Y et (2) la probabilité de sélection d'un instrument Z par sisVIVE en randomisation mendélienne.

4.2.1. Dérivation des intervalles de confiance pour l'estimateur sisVIVE et de la probabilité de sélection d'un instrument invalide

Dans le contexte de la randomisation mendélienne nous avons un échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim F \in \mathbb{R}^p$ i.i.d., où $p = 2 + L$ et L est le nombre de SNPs utilisés comme instruments. Par exemple, les quantités qui forment $\mathbf{X}_1$ sont $(Y_1, D_1, Z_{1,1}, \ldots, Z_{1,L})$. Dans un premier temps, nous souhaitons calculer des intervalles de confiance bootstrap pour β estimé par l'estimateur sisVIVE. Suivant la procédure élaborée dans la section précédente, étant donné que F est inconnue, nous allons d'abord l'estimer de façon non paramétrique par la fonction de répartition expérimentale des données originales $\hat{F}_n$, soit la distribution qui donne un poids $1/n$ à chacun des n vecteurs X_i. Ensuite, nous pouvons procéder par simulation.

`Pour` $i = 1, \ldots, B$ `:`

1. `Générer` $\mathbf{X}^*_1, \ldots, \mathbf{X}^*_n$ `i.i.d.` $\hat{F}_n$`, ce qui est équivalent à tirer un vecteur avec remise parmi les` n `vecteurs` $\mathbf{X}_1, \ldots, \mathbf{X}_n$`.`
2. `Calculer, en appliquant l'algorithme de sisVIVE aux données` $\mathbf{X}^*_1, \ldots, \mathbf{X}^*_n$`, l'estimation bootstrap du paramètre de l'effet causal` $\hat{\beta}^*_i$`.`

Par la simulation précédente, nous obtenons une distribution d'estimateurs bootstrap $\hat{\beta}^*_i$ qui estime la distribution de l'estimateur sisVIVE $\hat{\beta}$. Ainsi, pour estimer la variance de

$\hat{\beta}$, nous appliquons l'estimateur de type s^2 en (4.1.9) avec $\hat{\theta}_i^* = \hat{\beta}_i^*$ comme suit :

$$var_{\hat{F}_n}(\hat{\beta}^*) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\beta}_i^* - \overline{\hat{\beta}^*})^2, \tag{4.2.1}$$

où $\overline{\hat{\beta}^*} = B^{-1}\sum_{i=1}^{B}\hat{\beta}_i^*$.

Nous pouvons obtenir les quantiles de la fonction de répartition expérimentale de $\hat{\beta}^* - \hat{\beta}$ basée sur un échantillon de taille n provenant de $\hat{F}_n$, pour une probabilité ρ comme suit :

$$J_n^{-1}(\rho; \hat{F}_n) \approx \hat{\beta}^*_{(B\rho)} - \hat{\beta}. \tag{4.2.2}$$

Ainsi, nous pouvons construire différents types d'intervalles de confiance de niveau approximatif $(1-\rho)$ pour l'estimateur sisVIVE tel qu'illustré au tableau 4.2. Ces intervalles pourraient grandement différer selon que la distribution des réplications bootstrap $\hat{\beta}_i^*$ est symétrique ou non.

TABLEAU 4.2. Intervalles de confiance bootstrap de niveau approximatif $(1-\rho)$ pour l'estimateur sisVIVE avec $B = 1000$ et $\rho = 0{,}05$.

Type	IC 95%
normal	$\left[\hat{\beta} - 1{,}96\sqrt{var_{\hat{F}_n}(\hat{\beta}^*)},\quad \hat{\beta} + 1{,}96\sqrt{var_{\hat{F}_n}(\hat{\beta}^*)}\right]$
base	$\left[\hat{\beta} - (\hat{\beta}^*_{(975)} - \hat{\beta}),\quad \hat{\beta} - (\hat{\beta}^*_{(25)} - \hat{\beta})\right]$
percentile	$\left[\hat{\beta}^*_{(25)},\quad \hat{\beta}^*_{(975)}\right]$

Dans un deuxième temps, nous voulons estimer par le bootstrap la probabilité que l'algorithme sisVIVE identifie parmi les L instruments ceux qui sont invalides. Notons que d'un échantillon à l'autre, la sélection des instruments invalides pourrait changer. Étant donné que l'algorithme sisVIVE définit un instrument invalide par un effet non nul de l'instrument sur l'issue, pour un modèle de génération des observations donné, nous pouvons facilement définir la probabilité que le j^e instrument soit déclaré invalide. Par exemple, la probabilité que l'instrument j soit invalide est notée par

$$P_F\{\hat{\alpha}_j \neq 0\}, \tag{4.2.3}$$

où $\hat{\alpha}_j$ est l'estimation de l'effet du j^e instrument sur l'issue. Une estimation de cette probabilité serait

$$P_{\hat{F}_n}\{\hat{\alpha}_j^* \neq 0\} \approx \frac{1}{B}\sum_{i=1}^{B} I\{\hat{\alpha}_{ij}^* \neq 0\}, \tag{4.2.4}$$

où $\hat{\alpha}_{ij}^*$ est l'estimation de l'effet de l'instrument j sur l'issue dans le i^e échantillon bootstrap. Puisque les estimations $\hat{\alpha}_{ij}^*$ sont un produit dérivé de la procédure bootstrap appliquée à sisVIVE, nous pouvons facilement calculer l'estimation de la probabilité de sélection. Cette estimation pourrait être une autre indicatrice de la fiabilité du résultat de l'algorithme

sisVIVE : par exemple, si $P_{\hat{F}_n}\{\hat{\alpha}_j^* \neq 0\} \neq 0$ pour plus que $L/2$ instruments j, nous pourrions nous attendre à ce que $\hat{\beta}$ de sisVIVE soit une mauvaise estimation du vrai effet causal de l'exposition D sur l'issue Y.

4.2.2. Algorithme pour estimer la variance de l'estimateur sisVIVE et estimer la probabilité de sélection d'un instrument invalide

Nous présentons l'algorithme suivant pour le bootstrap de sisVIVE.

0. Définir la matrice des données originales:

$$\mathcal{X} = \begin{bmatrix} Y_1 & D_1 & Z_{1,1} & \cdots & Z_{L,1} \\ Y_2 & D_2 & Z_{1,2} & \cdots & Z_{L,2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ Y_n & D_n & Z_{1,n} & \cdots & Z_{L,n} \end{bmatrix}.$$

La fonction de répartition expérimentale $\hat{F}_n$ attribue une probabilité $1/n$ aux lignes de $\mathcal{X}$.

Pour $i = 1, \ldots, B$:

1. Construire $\mathcal{X}_i^*$, les matrices résultantes de $\mathcal{X}$, dont les j^e rangées sont $\mathbf{X}_{ij}^*$ en choisissant n lignes avec remise à partir de la matrice $\mathcal{X}$.

2. Appliquer l'algorithme sisVIVE aux données bootstrap $\mathbf{X}_{ij}^*$, $j = 1, \ldots, n$, pour calculer:
   a) $\hat{\alpha}_{ij}^*$, les estimations bootstrap des paramètres d'invalidité,
   b) $\hat{\beta}_i^*$, l'estimation bootstrap du paramètre de l'effet causal.

3. Calculer la variance des estimations bootstrap $\hat{\beta}_i^*$ par

$$\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\beta}_i^* - \overline{\hat{\beta}}^*)^2.$$

4. Calculer les quantiles de probabilité $\rho/2$ et $1-\rho/2$ par les statistiques ordonnées: par exemple, pour $B = 1000$ et $\rho = 0{,}05$, le quantile

du bas est $\hat{\beta}^*_{(25)}$ et le quantile du haut est $\hat{\beta}^*_{(975)}$.

5. Construire les intervalles de confiance de niveau approximatif $(1-\rho)$ de base, percentile, et normal, tels que détaillés au tableau 4.2.

Nous avons vu que le bootstrap nous permet de construire des intervalles de confiance pour un paramètre β lorsqu'il n'y a aucune solution analytique au problème d'inférence, comme c'est le cas avec l'estimateur sisVIVE. Nous avons présenté un algorithme pour estimer la fonction de répartition inconnue de l'échantillon aléatoire des observations de l'exposition, de l'issue et des SNPs et tirer des réalisations de cette estimation. Ensuite, à partir de ces échantillons nous obtenons une distribution d'estimateurs bootstrap dont nous pouvons alors calculer la variance et les quantiles nécessaires pour construire divers types d'intervalles de confiance. De plus, nous avons suggéré une méthode pour estimer la probabilité que sisVIVE sélectionne un instrument comme étant invalide dans le monde bootstrap. Dans le chapitre suivant, ces différentes mesures d'inférence seront confrontées dans une étude de simulation.

Chapitre 5

SIMULATIONS : ÉVALUATION DE LA PERFORMANCE DE L'ALGORITHME BOOTSTRAP POUR L'ESTIMATEUR SISVIVE

L'objet de cette section est d'investiguer l'approche du bootstrap non paramétrique dans le but de construire des intervalles de confiance pour l'estimateur sisVIVE. Nous voulons évaluer le comportement du bootstrap en fonction (1) d'un pourcentage croissant d'instruments invalides en raison de la pléiotropie et (2) de la taille d'échantillon en utilisant des jeux de données de tailles différentes.

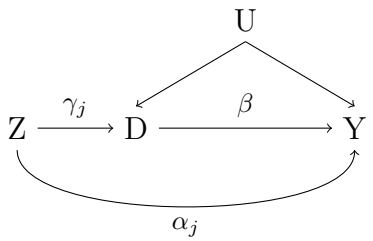
5.1. Contexte des simulations

Dans cette section, nous présentons le contexte des simulations. D'abord, nous posons le modèle à partir duquel les données sont simulées. Ensuite, nous expliquons le choix des paramètres et enfin détaillons les scénarios des simulations.

5.1.1. Modèle simulé et choix des paramètres

Nous allons poursuivre avec la notation introduite au chapitre 3. Les variables et les relations présentes dans le modèle à simuler sont illustrées par le DAG à la figure 5.1. Nous

Figure 5.1. DAG illustrant les relations entre les variables simulées.



ne suivons pas une représentation conventionnelle pour les DAGs. Les effets sont ajoutés au DAG seulement pour clarifier le lien entre ce dernier et le modèle statistique.

Nous simulons $N = 1000$ échantillons de n individus ayant des valeurs pour les variables U, Y, D, et Z. D'abord, nous simulons Z qui est une matrice de $L = 10$ SNPs. Ces observations sont simulées selon la loi de Hardy-Weinberg énoncée à la section 1.3.4. Par simplicité, nous avons fixé la valeur de MAF, c'est-à-dire, la fréquence de l'allèle de risque ($q = MAF$), pour tous les SNPs à 0,3 ce qui implique les probabilités Hardy-Weinberg suivantes : $p^2 = 0{,}49$, $2pq = 0{,}42$ et $q^2 = 0{,}09$. Nous simulons des échantillons de n individus issus d'une même population ayant chacun 10 SNPs qui prennent des valeurs de 0, 1 ou 2. Nous obtenons ainsi $Z_{ij} \sim HWE(\text{MAF}=0{,}3)$, où $j = 1, \ldots, 10$, $i = 1, \ldots, n$. En d'autres termes, les Z_{ij} sont indépendamment et identiquement distribués selon la loi

$$P\{Z_{ij} = k\} = \begin{cases} p^2, & k = 0 \\ 2pq, & k = 1 \\ q^2, & k = 2 \end{cases} . \tag{5.1.1}$$

Le code **R** des fonctions utilisées pour simuler les SNPs selon la loi HWE est disponible à l'annexe A.1.1.

Ensuite, nous simulons une variable de confusion U_i et les phénotypes exposition (l'indice de masse corporelle) D_i, et issue (la pression artérielle) Y_i, où $i = 1,...,n$, selon la séquence de modèles de régressions suivante :

$$\begin{aligned} U_i &= \epsilon_i^U, \\ D_i &= \pi^D + \sum_{j=1}^{L} \gamma_j Z_{ij} + U_i + \epsilon_i^D, \\ Y_i &= \pi^Y + \sum_{j=1}^{L} \alpha_j Z_{ij} + \beta D_i + U_i + \epsilon_i^Y, \end{aligned} \tag{5.1.2}$$

où ϵ_i^U, ϵ_i^D, $\epsilon_i^Y \sim N(0,1)$. Ici, U est une variable de confusion de l'association entre D et Y. C'est une variable de confusion non mesurée parce qu'elle est utilisée pour générer les données, mais exclue de l'analyse. La pléiotropie est induite en faisant dépendre Y sur Z hors du chemin qui passe par D, ce qui correspond aux α_j non zéro dans la troisième équation de (5.1.2). Nous fixons le vecteur de paramètre de l'effet de chaque SNP de Z sur l'exposition D à $\gamma = (-1,-1,1,1,-1,-1,1,1,-1,1)$. Le nombre de SNPs invalides en tant qu'instruments est représenté par s. Le paramètre d'invalidité α_j comportera des valeurs de 0 et 1 pour des SNPs valides et invalides, respectivement. Ainsi, s est égal au nombre de α_j différents de zéro. Nous utilisons comme effet causal de D sur Y la valeur $\beta = 1$, comme dans les simulations de sisVIVE [**23**]. Enfin, nous fixons la constante $\pi^D = 22$, soit

la moyenne de l'indice de masse corporelle lorsque Z et U sont égales à zéro, ainsi que la constante $\pi^Y = 115$, soit la moyenne de la pression artérielle lorsque Z, D et U sont égales à zéro. Ces valeurs sont acceptables parce que les effets sur l'exposition et l'issue sont relativement petits. En particulier, pour une population plus agée, une pression artérielle systolique de 115 mmHg est considérée relativement faible [**16**].

Toutes les simulations du modèle sisVIVE sont faites à l'aide du package `sisVIVE`[**22**] en **R**.

5.1.2. Scénarios des simulations

Nos scénarios varient selon la taille de l'échantillon ainsi que le nombre d'instruments (SNPs) invalides (voir le tableau 5.1). Dans chaque bloc de scénarios, nous considérons trois tailles d'échantillon : $n = 500$, ce qui est en général trop petit pour une analyse de randomisation mendélienne, $n = 2000$ et $n = 8000$. Ensuite, nous choisissons trois valeurs possibles pour s, soit 0, 3 et 7 correspondant à aucun SNP invalide, 30% de SNPs invalides et 70% de SNPs invalides, respectivement, puisque nous avons $L = 10$ SNPs. Nous pourrons ainsi étudier les cas où (1) sisVIVE et 2SLS devraient produire des résultats identiques (bloc A, $s = 0$), (2) sisVIVE gère l'invalidité des SNPs tandis que 2SLS échoue (bloc B, $s = 3$), (3) sisVIVE et 2SLS sont tous les deux théoriquement invalides (bloc C, $s = 7$), la raison pour sisVIVE étant que si la proportion de SNPs invalides dépasse 50%, il n'y a pas de garantie que la solution obtenue soit unique.

TABLEAU 5.1. Description des scénarios des simulations. Le nombre d'instruments invalides est s parmi $L = 10$, la taille de l'échantillon est n.

Bloc		s	n
A	A1	0	500
	A2	0	2000
	A3	0	8000
B	B1	3	500
	B2	3	2000
	B3	3	8000
C	C1	7	500
	C2	7	2000
	C3	7	8000

5.2. Résultats

Nous présentons les résultats de la performance des méthodes d'estimations sisVIVE et 2SLS ainsi que de l'algorithme bootstrap avec $B = 1000$ échantillons bootstrap pour l'estimation de la variance, de la probabilité d'invalidité d'un SNP et du calcul d'intervalles de confiance pour sisVIVE seulement. Nous mesurons la performance des méthodes avec les outils suivants.

— statistiques descriptives (moyenne, écart-type, minimum, maximum) des estimations $\hat{\beta}$ par les méthodes sisVIVE et 2SLS.

— biais d'estimation pour les méthodes sisVIVE et 2SLS,

$$Biais = \frac{1}{N}\sum_{k=1}^{N}\hat{\beta}_k - \beta, \qquad k = 1,\ldots,N. \tag{5.2.1}$$

— probabilité qu'un SNP soit déclaré invalide par sisVIVE et sa moyenne bootstrap : La probabilité que le j^e SNP soit déclaré invalide par sisVIVE est

$$P\{\hat{\alpha}_j \neq 0\} = \frac{1}{N}\sum_{k=1}^{N} I\{\hat{\alpha}_{jk} \neq 0\}, \qquad j = 1,\ldots,L \tag{5.2.2}$$

où $\hat{\alpha}_{jk}$ est l'estimation de la composante j du vecteur α pour le k^e échantillon simulé. La moyenne de la probabilité bootstrap que le j^e SNP soit déclaré invalide est

$$\mathbb{E}\Big[P_{\hat{F}_n}\{\hat{\alpha}^*_j \neq 0\}\Big] = \frac{1}{N}\Big(\sum_{k=1}^{N}\frac{1}{B}\sum_{l=1}^{B} I\{\hat{\alpha}^*_{jkl} \neq 0\}\Big), \qquad j = 1,\ldots,L \tag{5.2.3}$$

où $\hat{\alpha}^*_{jkl}$ est l'estimation de la composante j du vecteur $\hat{\alpha}$ pour le l^e échantillon bootstrap simulé à partir du k^e échantillon simulé.

— biais relatif de l'estimation de la variance de $\hat{\beta}$ estimé par sisVIVE,

$$biais\ relatif = \frac{\frac{1}{N}\sum_{k=1}^{N} var_{\hat{F}_n}(\hat{\beta^*}_k) - var(\hat{\beta})}{var(\hat{\beta})}, \tag{5.2.4}$$

où $var(\hat{\beta}) = \frac{1}{N-1}\sum_{k=1}^{N}(\hat{\beta}_k - \bar{\hat{\beta}})^2$ est la variance des estimations de sisVIVE.

— longueur moyenne d'un intervalle de confiance (IC) pour β basé sur l'estimateur sisVIVE :

$$longueur\ moyenne\ = \frac{1}{N}\sum_{k=1}^{N}\Delta_k, \tag{5.2.5}$$

où $\Delta_k = borne\ sup_k - borne\ inf_k$ et $IC_k = [borne\ inf_k\ , borne\ sup_k]$ est l'intervalle de confiance pour le k^e échantillon.

— probabilité de couverture (ρ), c'est-à-dire la proportion sur le nombre de simulations dans un scénario donné des intervalles de confiance contenant β,

$$\rho = \frac{1}{N} \sum_{k=1}^{N} I\{\beta \in IC_k\}. \tag{5.2.6}$$

5.2.1. Estimations avec sisVIVE et 2SLS

Pour chacun des scénarios détaillés à la section précédente, nous présentons les statistiques descriptives des estimations de l'effet causal avec (1) sisVIVE ($\hat{\beta}_{sisVIVE}$) et (2) 2SLS ($\hat{\beta}_{2SLS}$).

TABLEAU 5.2. Description des estimations de l'effet causal $\hat{\beta}_{sisVIVE}$.

Bloc		n	s	moyenne	écart-type	minimum	maximum	biais [a]
A	A1	500	0	1,00	0,0305	0,91	1,09	0,00
	A2	2000	0	1,00	0,0151	0,96	1,06	0,00
	A3	8000	0	1,00	0,0079	0,97	1,02	0,00
B	B1	500	3	0,98	0,0391	0,84	1,11	-0,02
	B2	2000	3	0,99	0,0195	0,92	1,04	-0,01
	B3	8000	3	0,99	0,0096	0,96	1,02	-0,01
C	C1	500	7	0,95	0,0690	0,68	1,13	-0,05
	C2	2000	7	0,98	0,0453	0,87	1,96	-0,02
	C3	8000	7	0,99	0,0461	0,01	1,06	-0,01

a. Le biais est calculé selon l'équation (5.2.1).

TABLEAU 5.3. Description des estimations de l'effet causal $\hat{\beta}_{2SLS}$.

Bloc		n	s	moyenne	écart-type	minimum	maximum	biais [a]
A	A1	500	0	1,00	0,0305	0,91	1,09	0,00
	A2	2000	0	1,00	0,0151	0,96	1,06	0,00
	A3	8000	0	1,00	0,0079	0,97	1,02	0,00
B	B1	500	3	0,91	0,0430	0,76	1,03	-0,09
	B2	2000	3	0,90	0,0220	0,81	0,97	-0,10
	B3	8000	3	0,90	0,0110	0,86	0,93	-0,10
C	C1	500	7	0,90	0,0538	0,72	1,09	-0,10
	C2	2000	7	0,90	0,0281	0,81	1,01	-0,10
	C3	8000	7	0,90	0,0137	0,86	0,96	-0,10

a. Le biais est calculé selon l'équation (5.2.1).

Selon les tableaux 5.2 et 5.3, le biais de sisVIVE est toujours moindre que celui de 2SLS, ce qui suggère que sisVIVE fait mieux que la méthode conventionnelle en terme du biais d'estimation. Nous pouvons remarquer au tableau 5.2 que les cas où 70% des instruments sont invalides ne démontrent pas une très forte augmentation du biais relativement aux deux autres cas et que le biais relatif est d'au plus 5%. Les résultats suggèrent que dans nos simulations, sisVIVE semble être tout aussi robuste à l'invalidité avec 70% d'instruments invalides qu'avec 30% d'instruments invalides pourvu que l'échantillon soit suffisamment grand (voir les cas où la taille est $n = 8000$).

Concernant la distribution de $\hat{\beta}_{sisVIVE}$, dans les cas où 30% des instruments sont invalides, nous observons que la sous-estimation de l'effet diminue avec l'augmentation de la taille d'échantillon. Pour les cas où 70% des instruments sont invalides, nous observons que la sous-estimation de l'effet est un peu plus importante que dans les cas où 30% des instruments sont invalides. De plus, nous remarquons la présence de valeurs plus grandes, mais toutefois cliniquement plausibles, dans la distribution de $\hat{\beta}_{sisVIVE}$: pour C2 nous avons une sur-estimation importante (valeur de 1,96) et pour C3 nous avons deux sous-estimations importantes (valeurs de 0,01 et 0,02). Concernant la distribution de $\hat{\beta}_{2SLS}$, qu'il y ait 30% ou 70% de SNPs invalides et que la taille soit de 500, 2000 ou 8000, le biais est essentiellement toujours le même, soit un biais relatif [1] de 10%, alors que le pire biais relatif de sisVIVE est de 5% pour n=500, 2% pour n=2000 et 1% pour n=8000. Contrairement à ce que nous observons pour sisVIVE, nous n'observons pas de valeurs extrêmes pour la méthode 2SLS.

Une autre particularité importante des résultats du tableau 5.2 est qu'à l'exception du cas de sisVIVE avec 70% d'invalidité, tant sisVIVE que 2SLS ont une variance qui semble inversement proportionnelle à la taille de l'échantillon alors que lorsqu'il y a 70% d'invalidité, l'écart-type de la distribution diminue peu en fonction de la taille d'échantillon et a même augmenté dans notre simulation avec sisVIVE lorsque la taille des échantillons est passée de 2000 à 8000.

5.2.2. Estimations bootstrap de la probabilité d'invalidité d'un SNP selon sisVIVE

L'estimateur sisVIVE fait une sélection des instruments valides par le LASSO (voir l'équation (3.6.13)) et nous vérifions les performances de cette sélection dans l'application du bootstrap au paramètre d'invalidité estimé par sisVIVE.

Le tableau 5.4 présente les estimations de la probabilité qu'un SNP soit invalide selon (1) l'algorithme de sisVIVE (rangées « estimée », basées sur l'équation (5.2.2)) et (2) le bootstrap appliqué à sisVIVE (rangées « bootstrap », basées sur l'équation (5.2.3)). Nous

1. le biais sur la valeur du paramètre

observons dans ce tableau que sisVIVE détecte les s instruments invalides, peu importe la valeur de s. De même, pour tous les échantillons bootstrap, la probabilité estimée d'être invalide est de 1 pour les s instruments invalides. L'estimateur semble se tromper quelques fois en identifiant des instruments comme invalides alors qu'ils ne le sont pas. De plus, le nombre de ces « faux positifs » augmente plus la proportion de SNPs invalides augmente. Pour les cas où aucun instrument n'est invalide, il n'y a aucun faux positif dans l'estimation par sisVIVE, mais pour l'estimation bootstrap sisVIVE se trompe 3-4% des fois. Pour les cas où 30% des instruments sont invalides, l'estimation par sisVIVE se trompe 4-6% des fois et pour l'estimation bootstrap sisVIVE se trompe 16-18% des fois. Enfin, pour les cas où 70% des instruments sont invalides, la probabilité estimée augmente à 14-20% pour l'estimation par sisVIVE et 23-28% pour l'estimation bootstrap.

TABLEAU 5.4. Estimation de l'invalidité des SNPs : proportions pour les $N = 1000$ échantillons et chacun des $B = 1000$ échantillons bootstrap.

Bloc	Scénario	n	s	probabilité [a]	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
A				vraie	0	0	0	0	0	0	0	0	0	0
	A1	500	0	estimée	0	0	0	0	0	0	0	0	0	0
		500	0	bootstrap	0,04	0,03	0,04	0,03	0,03	0,03	0,03	0,04	0,04	0,03
	A2	2000	0	estimée	0	0	0	0	0	0	0	0	0	0
		2000	0	bootstrap	0,03	0,04	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
	A3	8000	0	estimée	0	0	0	0	0	0	0	0	0	0
		8000	0	bootstrap	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
B				vraie	1	1	1	0	0	0	0	0	0	0
	B1	500	3	estimée	1	1	1	0,05	0,06	0,05	0,06	0,04	0,05	0,04
		500	3	bootstrap	1	1	1	0,17	0,17	0,17	0,18	0,17	0,17	0,17
	B2	2000	3	estimée	1	1	1	0,04	0,04	0,04	0,04	0,04	0,05	0,05
		2000	3	bootstrap	1	1	1	0,16	0,16	0,16	0,17	0,17	0,18	0,17
	B3	8000	3	estimée	1	1	1	0,05	0,05	0,04	0,04	0,04	0,04	0,05
		8000	3	bootstrap	1	1	1	0,17	0,17	0,16	0,16	0,17	0,16	0,17
C			7	vraie	1	1	1	1	1	1	1	0	0	0
	C1	500	7	estimée	1	1	1	1	1	1	1	0,17	0,17	0,20
		500	7	bootstrap	1	1	1	1	1	1	1	0,27	0,27	0,28
	C2	2000	7	estimée	1	1	1	1	1	1	1	0,15	0,15	0,15
		2000	7	bootstrap	1	1	1	1	1	1	1	0,26	0,25	0,26
	C3	8000	7	estimée	1	1	1	1	1	1	1	0,16	0,14	0,15
		8000	7	bootstrap	1	1	1	1	1	1	1	0,24	0,23	0,24

a. La probabilité estimée est calculée selon l'équation (5.2.2) et la probabilité bootstrap est calculée selon l'équation (5.2.3).

5.2.3. Estimations bootstrap de la variance et intervalles de confiance pour sisVIVE

Le tableau 5.5 présente les statistiques descriptives des estimations de la variance par bootstrap (avec B=1000 réplications bootstrap). Nous y rapportons également le biais relatif de l'estimateur bootstrap de la variance de $\hat{\beta}_{sisVIVE}$ basé sur l'équation (5.2.4).

Au tableau 5.5, certains scénarios présentent des estimations de $var_{\hat{F}_n}(\hat{\beta}^*_{sisVIVE})$ extrêmes de l'ordre $\geq 1 \times 10^{19}$. Nous avons donc aussi présenté ces scénarios sans les valeurs extrêmes, parce qu'elles fournissent des estimations de la variance qui sont tellement grandes qu'elles seraient simplement mises de côté si un jeu de données donnait de tels résultats. Nous distinguons l'échantillon tronqué de l'échantillon original par l'affixe (**). Remarquons que le nombre de valeurs extrêmes diminue plus la taille de l'échantillon augmente et le phénomène est davantage répandu dans les cas où 70% des instruments sont invalides. Parce que ces valeurs sont très surprenantes, nous avisons le chercheur ou la chercheuse d'éviter d'utiliser sisVIVE dans ces cas. Bien que nous observons que la méthode d'estimation de la variance fonctionne mieux dans les échantillons tronqués que dans les échantillons originaux, nous ne pouvons conclure à partir des échantillons tronqués que notre méthode fonctionne.

Dans le cas où tous les instruments sont valides, le biais relatif est très faible pour tous les échantillons (de l'ordre de 1×10^{-2}). Nous avons vu précédemment que la variance de sisVIVE à estimer a tendance à diminuer plus la taille d'échantillon augmente (voir tableau 5.2). Ici, nous observons pour le cas où 30% des instruments sont invalides que la qualité de l'estimateur bootstrap de cette variance diminue lorsque la taille de l'échantillon augmente. En effet, pour ce cas, le biais relatif de l'estimateur bootstrap de la variance est de 0,14 (B1), puis 0,18 (B2) et ensuite $6{,}24 \times 10^{21}$ (B3), où la valeur en B3 est très grande parce qu'un échantillon de ce scénario a mené à une estimation bootstrap de la variance qui est extrême. Pour le cas où 70% des instruments sont invalides, le biais relatif reflète les estimations extrêmes de la variance et est donc extrêmement large pour les trois tailles d'échantillon considérées. Notons que dans les échantillons tronqués, le biais relatif augmente avec la taille de l'échantillon : d'abord 0,20 (C1**), puis 0,70 (C2**) et ensuite 2,08 (C3**). Toutefois, tel que mentionné précédemment, nous ne recommanderions pas l'utilisation de sisVIVE dans les cas où des estimations extrêmes sont produites, notamment dans le cas où 70% des instruments sont invalides.

Les tableaux 5.6, 5.7 et 5.8 présentent les résultats concernant les intervalles de confiance bootstrap de niveau 95% de type percentile, de base, et normal pour sisVIVE. Nous y rapportons la longueur moyenne et la probabilité de couverture (ρ). Notons que puisque ρ est estimé à partir de 1000 intervalles de confiance, un test d'hypothèses de niveau 5% que la

TABLEAU 5.5. Description des estimations bootstrap de la variance de $\hat{\beta}_{sisVIVE}$.

Bloc		n	s	moyenne	minimum	maximum	biais relatif [a]
A	A1	500	0	9,79e-04	6,42e-04	1,61e-03	5,34e-02
	A2	2000	0	2,45e-04	1,89e-04	3,55e-04	7,04e-02
	A3	8000	0	6,10e-05	4,96e-05	8,06e-05	-2,41e-02
B	B1	500	3	1,74e-03	9,01e-04	2,87e-03	1,40e-01
	B2	2000	3	4,52e-04	2,98e-04	2,41e-02	1,84e-01
	B3	8000	3	6,29e17	7,99e-05	6,29e20	6,74e21
	B3 (** 1)	8000	3	5,42e-03	7,99e-05	9,26e-01	5,71e01
C	C1	500	7	3,19e23	2,43e-03	6,84e25	6,69e25
	C1 (** 55)	500	7	5,73e-03	2,43e-03	1,99e-02	2,00e-01
	C2	2000	7	1,67e22	7,14e-04	1,42e25	8,14e24
	C2 (** 40)	2000	7	3,53e-03	7,14e-04	1,86e-01	6,98e-01
	C3	8000	7	1,61e22	2,07e-04	8,32e24	7,55e24
	C3 (** 23)	8000	7	6,69e-03	2,07e-04	3,97e-01	2,08e00

a. Le biais relatif est calculé selon l'équation (5.2.4).

vraie probabilité de couverture est de 95% rejette cette hypothèse si la couverture estimée est inférieure à 0,936 ou supérieure à 0,964.

TABLEAU 5.6. Description des intervalles de confiance bootstrap de niveau 95% pour $\hat{\beta}_{sisVIVE}$ dans les cas où tous les instruments sont valides.

Bloc		n	type	longueur moyenne [a]	ρ [b]
A	A1	500	percentile	0,122	0,940
$s = 0$		500	base	0,122	0,946
		500	normal	0,122	0,944
	A2	2000	percentile	0,061	0,946
		2000	base	0,061	0,952
		2000	normal	0,061	0,952
	A3	8000	percentile	0,030	0,945
		8000	base	0,030	0,948
		8000	normal	0,031	0,948

a. La longueur moyenne est calculée selon l'équation (5.2.5).
b. La couverture est calculée selon l'équation (5.2.6).

D'abord, dans les cas où tous les instruments sont valides, la longueur moyenne diminue avec la taille de l'échantillon n et la probabilité de couverture est proche de 0,95. Le type d'intervalle, soit percentile, de base et normal, n'influence pas les résultats parce que

TABLEAU 5.7. Description des intervalles de confiance bootstrap de niveau 95% pour $\hat{\beta}_{sisVIVE}$ dans les cas où 30% des instruments sont invalides.

Bloc		n	type	longueur moyenne [a]	ρ [b]
B	B1	500	percentile	0,163	0,946
$s = 3$		500	base	0,163	0,944
		500	normal	0,163	0,948
	B2	2000	percentile	0,080	0,901
		2000	base	0,080	0,917
		2000	normal	0,081	0,918
	B3	8000	percentile	0,040	0,921
		8000	base	0,040	0,924
		8000	normal	9,830e07	0,967
	B3 (** 1)	8000	percentile	0,040	0,921
		8000	base	0,040	0,924
		8000	normal	0,141	0,967

a. La longueur moyenne est calculée selon l'équation (5.2.5).
b. La couverture est calculée selon l'équation (5.2.6).

les probabilités de couverture ρ ne sont pas significativement différentes au seuil de 5% et les longueurs moyennes sont essentiellement identiques. Ensuite, dans les cas où 30% des instruments sont invalides, la longueur moyenne est un peu plus grande que dans le cas précédent et diminue généralement avec n, sauf dans le scénario B3 où l'intervalle de confiance normal est largement imprécis. Ceci s'explique par le fait que la méthode normale se base sur l'estimateur de variance bootstrap et qu'un des échantillons bootstrap cause une estimation extrême de la variance de sisVIVE (de l'ordre de 1×10^{20}, voir le tableau 5.5). En revanche, les intervalles de confiance percentile et de base ne sont pas affectés : la longueur de B3 et B3(**) est identique (voir le tableau 5.7). Ceci est dû au fait que ces intervalles se basent sur les quantiles 2,5% et 97,5% et que l'unique estimation extrême de la variance de sisVIVE n'a aucun effet sur le calcul de ces derniers. Il n'y a pas de tendance claire pour la couverture, qui semble d'abord diminuer de B1 à B2, puis augmenter de B2 à B3 pour chacun des trois types d'intervalle. Dans ce scénario, la probabilité de couverture de 0,95 est atteinte dans le petit échantillon ($n = 500$), puis se détériore de façon significative pour l'échantillon de taille moyenne ($n = 2000$) pour enfin s'améliorer dans le plus grand échantillon ($n = 8000$) où ρ n'est plus aussi loin de 0,95.

Enfin, dans les cas où 70% des instruments sont invalides les intervalles sont beaucoup plus longs que dans les cas précédents, mais continuent à diminuer plus la taille de l'échantillon augmente. Les intervalles de confiance de type percentile, de base et normal

TABLEAU 5.8. Description des intervalles de confiance bootstrap de niveau 95% pour $\hat{\beta}_{sisVIVE}$ dans les cas où 70% des instruments sont invalides.

Bloc		n	type	longueur moyenne [a]	ρ [b]
C	C1	500	percentile	0,295	0,903
$s = 7$		500	base	0,295	0,970
		500	normal	3,703e11	0,953
	C1 (** 55)	500	percentile	0,293	0,901
		500	base	0,293	0,970
		500	normal	0,293	0,950
	C2	2000	percentile	0,138	0,889
		2000	base	0,138	0,940
		2000	normal	4,324e10	0,972
	C2 (** 40)	2000	percentile	0,137	0,886
		2000	base	0,137	0,941
		2000	normal	0,195	0,971
	C3	8000	percentile	0,068	0,918
		8000	base	0,068	0,946
		8000	normal	3,532e10	0,989
	C3 (** 23)	8000	percentile	0,068	0,918
		8000	base	0,068	0,947
		8000	normal	0,212	0,989

a. La longueur moyenne est calculée selon l'équation (5.2.5).
b. La couverture est calculée selon l'équation (5.2.6).

donnent des résultats plus variables pour ces cas. La couverture est de moins de 0,95 pour les intervalles percentile, se rapproche de 0,95 pour les intervalles de base et est au-delà de 0,95 pour les intervalles normal. Pour les intervalles de type normal, cela s'explique par le fait que l'estimation de variance bootstrap est trop grande. Les intervalles de confiance percentile et de base ne sont pas affectés par la variance plus grande parce que l'estimation des quantiles 2,5% et 97,5% bootstrap n'est pas affectée par la présence d'estimations extrêmes de la variance de sisVIVE. En effet, la proportion des estimations $\hat{\beta}_{sisVIVE}$ plus extrêmes qui mènent à une estimation de la variance bootstrap qui est plus extrême est très faible (voir le tableau 5.5).

De façon générale, dans les cas où tous les instruments sont valides, tous les types d'intervalles ont une bonne couverture. Dans les cas où 30% et 70% des instruments sont invalides, l'intervalle normal a une moins bonne couverture en raison des estimations extrêmes de la variance. Pour ce qui est des intervalles percentile et de base, leur couverture est semblable pour le cas de 30% d'instruments invalides, mais l'intervalle de base a une

meilleure couverture pour le cas de 70% d'instruments invalides où, mis à part l'échantillon de taille $n = 500$, la couverture n'est pas statistiquement différente de 95%.

Les simulations nous ont permis de vérifier (1) la contribution de sisVIVE à l'estimation de l'effet causal d'une exposition $\mathbf{D}$ sur une issue $\mathbf{Y}$ en randomisation mendélienne, (2) la possibilité d'utiliser la méthode bootstrap pour calculer des intervalles de confiance pour sisVIVE et (3) le comportement du bootstrap pour estimer la probabilité qu'un SNP soit sélectionné comme étant invalide par sisVIVE. D'abord, sisVIVE gère le biais engendré par l'utilisation de SNPs invalides lorsque leur proportion est inférieure à 50%. Au-delà de ce seuil, nous observons que, pour les trois scénarios considérés, sisVIVE semble être aussi robuste à l'invalidité des SNPs pourvu que l'échantillon soit suffisamment grand ($n \geq 8000$), contrairement à ce qui est décrit et observé par Kang et al. (2014). En effet, leurs résultats indiquaient plutôt que la proportion d'instruments invalides doit être inférieure à 50%. Ensuite, pour faire l'inférence pour l'effet causal estimé par sisVIVE, nous recommandons les intervalles de confiance calculés par la méthode de base parce que ce dernier est robuste au biais de l'estimateur bootstrap de la variance de sisVIVE engendré par l'utilisation de SNPs invalides et par les valeurs parfois extrêmes qu'il peut produire. Enfin, à partir des résultats du bootstrap appliqué au paramètre d'invalidité $\boldsymbol{\alpha}$, nous observons que l'estimation de la probabilité d'invalidité d'un SNP selon sisVIVE comporte un biais positif alors que la probabilité estimée à partir de l'algorithme sisVIVE se rapproche davantage de la vérité. Dans le chapitre suivant, nous illustrons notre méthode dans une application aux données de la Biobanque de l'Institut de Cardiologie de Montréal.

Chapitre 6

APPLICATION : ÉTUDE DE L'EFFET CAUSAL DE L'OBÉSITÉ SUR LA PRESSION ARTÉRIELLE

Nous souhaitons illustrer notre méthode d'inférence par bootstrap pour l'estimateur sisVIVE dans un contexte réel en randomisation mendélienne. Pour ce faire, nous appliquons cette méthode à l'analyse de l'effet causal de l'obésité sur la pression artérielle. Nous utilisons 10 SNPs associés à l'indice de masse corporelle comme instruments dans l'estimation de l'effet avec les méthodes 2SLS et sisVIVE.

Dans un premier temps, nous devons construire une base de données pour l'approche par la randomisation mendélienne. Tel que vu dans les chapitres précédents, celle-ci doit inclure l'exposition (l'indice de la masse corporelle, ci-après l'IMC), l'issue (la pression artérielle systolique, ci-après PAS), des variables de confusion potentielles pour l'effet de l'IMC sur la PAS (e.x. l'âge, le sexe, etc.) ainsi que des SNPs qui sont fortement associés à l'IMC.

Pour notre application, ces données proviennent de la Biobanque de l'Institut de Cardiologie de Montréal que nous présentons dans la prochaine section, parce qu'elle contient des données génétiques (certaines sont mesurées directement sur la puce de l'exome de la compagnie Illumina, d'autres sont imputées), cliniques et descriptives. Pour les données génétiques, nous nous intéressons uniquement aux SNPs qui sont mesurés directement sur la puce de génotypage, plutôt que les SNPs imputés, la raison étant que les méthodes d'imputation sont au-delà des objectifs de ce mémoire.

Afin de choisir quels SNPs devraient servir d'instruments, nous avons utilisé les résultats d'une méta-analyse. Nous avons choisi la méta-analyse publiée en 2018 par Turcot et al. [**40**] parce qu'elle est la plus récente qui rapporte les SNPs associés à l'IMC dont le génotypage a été fait avec la puce de l'exome, soit la même puce que celle utilisée dans la Biobanque. Ainsi, nous savions que les SNPs rapportés dans cette méta-analyse seraient directement mesurés, plutôt qu'imputés, dans les données de la Biobanque. Bien que la méta-analyse [**40**] se concentre sur l'identification de SNPs plus rares, c'est-à-dire les SNPs

avec une fréquence d'allèle rare (MAF) inférieure à 5%, elle rapporte également une liste de 150 SNPs communs ($MAF \geq 5\%$) dont nous avons pu nous servir pour sélectionner nos SNPs. Nous allons présenter l'algorithme de sélection des 10 SNPs communs et fortement associés à l'IMC dans les prochaines sections.

Dans un deuxième temps, nous appliquons les méthodes d'estimation 2SLS et sisVIVE à notre base de données et notre méthodologie d'inférence par bootstrap pour sisVIVE. Les résultats de l'analyse pour l'échantillon de la Biobanque seront présentés et la discussion de ces derniers fera la clôture ce chapitre.

6.1. Biobanque de l'Institut de Cardiologie de Montréal

Avant de procéder à l'explication de la construction de la base de données, nous présentons d'abord le projet Biobanque de l'Institut de Cardiologie de Montréal duquel les données proviennent et la puce de l'exome Illumina à partir de laquelle le génotypage a été fait.

6.1.1. Projet Biobanque

Affilié à l'Université de Montréal, l'Institut de Cardiologie de Montréal est reconnu mondialement pour sa recherche de pointe en médecine cardiovasculaire [**3**]. En 2007, cette institution a mis sur pied un projet de cohorte hospitalière qui a pour objectif d'investiguer la génétique des maladies cardiovasculaires. Sous la direction de Jean-Claude Tardif, M.D., la Biobanque de l'Institut de Cardiologie de Montréal recueille, entre autres, des données médicales, nutritionnelles, psychosociales, ainsi que des données sur les habitudes de vie de ses participants [**2**].

Les données de la Biobanque sont recueillies par le biais d'un questionnaire administré par un infirmier ou une infirmière de l'équipe de recherche lors de la première visite du participant et le suivi se fait tous les quatre ans. Tous les participants au projet Biobanque ont accepté d'être génotypés dès leur recrutement dans l'étude génétique et le génotypage se fait selon l'ordre de recrutement dans l'étude. Par exemple, les premiers 11 000 participants sont génotypés, suivi d'un autre ensemble de participants, et ainsi de suite.

6.1.2. Illumina Exome Chip

Les participants de la Biobanque sont génotypés avec la puce de l'exome Illumina. Fabriquée par la compagnie Illumina, cette puce de génotypage a un haut débit et, contrairement aux puces traditionnelles de criblage du génome, elle a une couverture génomique enrichie dans les régions du génome codant pour des protéines. Elle permet de tester si

certains SNPs peuvent être responsables de variations codant pour des protéines, et ainsi être responsables de variations inter-individuelles dans les phénotypes associés à ces gènes, tels ceux liés au cholestérol à lipoprotéines de haute densité (HDL).

6.2. Méta-analyse d'études pangénomiques

Nous expliquons maintenant ce qu'est la méta-analyse d'études pangénomiques et présentons la méta-analyse utilisée pour sélectionner les 10 SNPs communs les plus fortement associés à l'IMC dans la Biobanque.

De façon générale, une méta-analyse d'étude à l'échelle du génome permet d'estimer les effets de différents SNPs sur un phénotype d'intérêt à partir des résultats de diverses études pangénomiques (ci-après, GWAS). Chaque GWAS produit un coefficient pour l'estimation de l'effet du SNP sur le phénotype. Les statistiques sommaires (entre autres, le coefficient estimé et l'erreur standard) de chaque SNP provenant de chaque étude sont ensuite combinées à l'aide d'un modèle dans la méta-analyse afin d'obtenir l'effet final. Nous pouvons ainsi faire une sélection des SNPs les plus fortement associés avec un phénotype sur la base, par exemple, du coefficient de l'effet estimé rapporté par la méta-analyse et/ou de sa valeur-p.

À notre connaissance, l'analyse de Turcot et al. (2018) est la plus récente méta-analyse portant sur la génétique de l'IMC dont les SNPs sont répertoriés sur une puce de l'exome. Cette méta-analyse recueille les résultats d'un total de 125 GWAS ($n_{max} = 718\,734$ individus), parmi lesquelles on retrouve les cohortes deCODE ($n_{max} = 72\,613$) et UK Biobank ($n_{max} = 119\,613$ [1]). La significativité statistique de 246 328 SNPs rares et communs a été testée. Ceux qui atteignaient le seuil de significativité statistique de $p < 2 \times 10^{-7}$ ont été considérés comme étant fortement associés à l'IMC. La combinaison des résultats des études s'est faite avec une méta-analyse à effets fixes. Le tableau S4 [**40**] présente, entre autres, le coefficient de l'effet estimé sur l'IMC et sa valeur-p en méta-analyse pour 150 SNPs communs ($MAF \geq 5\%$).

6.3. Construction de la base de données

Tel que mentionné précédemment, nous utilisons les données de la Biobanque qui contient des données descriptives, cliniques et génétiques à partir desquelles nous pouvons effectuer une analyse par la technique de la randomisation mendélienne. Dans les sections qui suivent, nous décrivons d'abord le traitement des données génétiques et la sélection des 10 SNPs communs les plus fortement associés à l'IMC ; ensuite nous décrivons le traitement

1. Il s'agit de la taille rapportée pour la version provisoire de la cohorte. Plus d'individus ont été inclus dans le UK Biobank depuis le début de la méta-analyse.

des données descriptives et cliniques et présentons la description de notre échantillon final de 10 455 participants.

6.3.1. Données génétiques

Nous présentons la description des 10 SNPs utilisés dans l'analyse et l'algorithme de sélection de ces derniers fait avec la méta-analyse [**40**].

6.3.1.1. *Description des 10 SNPs communs les plus fortement associés à l'indice de masse corporelle*

Nous avons sélectionné 10 SNPs qui sont des variants génétiques communs présents pour chacun des deux sexes et qui ne sont pas en déséquilibre de liaison, c'est-à-dire que nous ne choisirons pas de SNPs fortement corrélés entre eux. Le tableau 6.1 présente les effets de chaque SNP sur l'IMC dans la méta-analyse et dans l'échantillon de la Biobanque. Le coefficient estimé $\hat{\beta}$ rapporté pour la Biobanque est estimé dans un modèle de régression linéaire simple de l'IMC (standardisé pas sa moyenne et son écart-type) sur le SNP. Au tableau 6.1 nous rapportons également le chromosome sur lequel le SNP est situé, le nom du SNP, le gène le plus proche du SNP associé à la mutation et la fréquence de l'allèle d'effet [2] dans la méta-analyse et dans la Biobanque.

6.3.1.2. *Algorithme de sélection de 10 SNPs fortement associés à l'indice de masse corporelle*

Notre procédure de sélection de 10 SNPs communs et fortement associés avec l'IMC est la suivante.

Étape 1. La méta-analyse [**40**] identifie 150 SNPs communs qui ont une valeur-p inférieure à 2×10^{-7}. Sur la base de la description et des statistiques sommaires rapportées pour les 150 SNPs de la méta-analyse :

(i) Par convention, nous nous sommes limités à ceux dont la valeur-p était inférieure à 1×10^{-8}.

(ii) Parce que nous ne considérons que les chromosomes 1 à 22, nous avons éliminé le SNP rs11539157 de notre sélection qui se trouve sur le chromosome 23.

(iii) Nous avons éliminé 3 autres SNPs, soit rs2946994, rs2280843 et rs6065, parce que l'effet rapporté est spécifique à un seul des deux sexes.

(iv) Nous avons éliminé 23 autres SNPs (lignes 63 à 86 du tableau S4 de [**40**]) pour lesquels la méta-analyse n'a pas clairement spécifié si les SNPs pouvaient être

2. L'allèle d'effet ne correspond pas nécessairement à l'allèle rare. L'allèle rare est par définition celui dont la fréquence est plus faible dans une population, tandis que l'allèle d'effet est celui dont la présence dans la génétique de l'individu a pour effet d'augmenter son IMC.

TABLEAU 6.1. Description des 10 SNPs communs associés à l'IMC.

			Méta-analyse				Biobanque			
Chr	**SNP**	**Gène**	**FAE** [a] **(%)**	$\hat{\beta}$ [b]	**ES** [c]	**Valeur-p** [d]	**FAE (%)**	$\hat{\beta}$	**ES**	**Valeur-p**
2	rs11676272	*ADCY3*	48,6	0,029	0,002	3,84e-59	46,9	0,026	0,014	0,062
3	rs295322	*RASA2*	7,4	0,028	0,004	3,81e-14	6,8	0,001	0,028	0,965
4	rs34811474	*ANAPC4*	79,0	0,024	0,002	2,91e-30	78,9	0,006	0,017	0,723
4	rs13107325	*SLC39A8*	6,0	0,050	0,004	5,34e-40	8,1	0,044	0,025	0,087
6	rs9469913	*UHRF1BP1*	15,8	0,025	0,003	5,37e-16	16,6	0,034	0,018	0,063
11	rs6265	*BDNF*	81,6	0,041	0,002	2,75e-68	80,0	0,047	0,017	0,007
11	rs1064608	*MTCH2*	33,7	0,024	0,002	1,28e-31	34,4	0,028	0,015	0,055
16	rs2904880	*CD19*	32,7	0,028	0,002	1,10e-35	29,9	0,023	0,015	0,138
16	rs3213758	*RPGRIP1L*	94,3	0,025	0,004	5,50e-11	94,9	0,040	0,032	0,205

a. FAE signifie la fréquence de l'allèle d'effet.
b. Les effets $\hat{\beta}$ sont calculés dans un modèle de régression linéaire simple où l'IMC est standardisé (moyenne=0, écart-type=1).
c. ES signifie l'erreur standard.
d. La valeur-p est calculée à partir d'un test Z en bilatéral pour le coefficient de la régression linéaire de l'IMC (standardisé) sur le SNP.

considérés.

Suite à cette élimination préliminaire, il reste 117 SNPs.

Étape 2. Nous considérons maintenant les données de la Biobanque. Puisque nous voulons des SNPs communs, nous enlevons le SNP rs62623713 dont la fréquence de l'allèle d'effet est de 4% dans la Biobanque, ce qui est inférieur à 5%. Il reste alors 116 SNPs.

Étape 3. Étant donné que nous voulons les SNPs les plus fortement associés à l'IMC, nous classons les 116 SNPs restants par ordre décroissant des coefficients $\hat{\beta}$ dans la méta-analyse. Nous conservons les 20 premiers SNPs sur la liste à partir desquels nous retiendrons les 10 SNPs les plus forts suite à une élimination des SNPs qui présentent des erreurs d'encodage dans la base de données de la Biobanque et des SNPs qui sont corrélés entre eux.

Étape 4. Pour sélectionner les 10 SNPs les plus forts :

(i) Parce que nous voulons nous assurer que les effets dans la Biobanque sont rapportés pour le même allèle que celui dans la méta-analyse, nous éliminons le SNP rs34149579 dont la fréquence d'allèle d'effet entre la Biobanque et la méta-analyse diffère trop (43% dans la Biobanque et 95% dans la méta-analyse). Il reste alors 19

SNPs.

(ii) Puisque nous voulons que les effets des SNPs dans la Biobanque aillent dans le même sens que ceux de la méta-analyse, nous éliminons les SNPs rs2820312 et rs2250377 dont les coefficients $\hat{\beta}$ dans la Biobanque sont négatifs, parce que ça implique qu'il y a un changement de signe par rapport au $\hat{\beta}$ de la méta-analyse. Il reste alors 17 SNPs.

(iii) Étant donné que nous voulons que nos 10 SNPs soient non corrélés, nous calculons la matrice de corrélation des SNPs par chromosome. Si le coefficient de corrélation $r^2 \geq 0{,}18$, nous éliminons le SNP qui a le plus petit $\hat{\beta}$ dans la méta-analyse. Dans notre cas, après avoir calculé les coefficients de corrélation pour les SNPs situés sur le même chromosome [3], nous avons observé que les 10 meilleurs SNPs n'étaient pas corrélés entre eux.

(iv) Alors, parmi les 17 SNPs restants, nous avons sélectionné les 10 premiers.

6.3.2. Données descriptives et cliniques

Tel que mentionné au chapitre 2, les SNPs utilisés comme instruments doivent satisfaire trois suppositions pour être considérés comme étant valides. Bien que les suppositions (A2) et (A3) ne soient pas vérifiables dans le jeu de données réelles, nous pouvons prendre certaines précautions pour minimiser la possibilité de ne pas respecter la supposition (A2), c'est-à-dire, qu'il n'existe pas de cause commune non mesurée entre un SNP et le phénotype issue. En effet, nous pouvons sélectionner notre échantillon de telle sorte à ce que nous minimisons la possibilité que l'échantillon soit stratifié par une variable telle que l'origine ethnique, ce qui mènerait à une violation de la supposition (A2). Pour ce faire, nous avons inclus seulement des participants d'ascendance européenne [4] dans notre échantillon ($n = 11\ 292$). De plus, les données de la Biobanque sont soumises à un contrôle de qualité dont l'un des critères est qu'uniquement les participants issus de familles distinctes soient étudiés. Après cette première sélection, la construction de la base de données s'est poursuivie comme suit.

En premier lieu, nous avons construit une base de données transversales qui contient l'exposition et l'issue, soit l'IMC et la PAS, ainsi que les variables de confusion potentielles suivantes : âge, sexe, fréquence de consommation d'alcool et le statut tabagique.

En deuxième lieu, parce que les participants de la Biobanque sont d'abord des patients de l'Institut de Cardiologie de Montréal, il est possible qu'un ensemble d'entre eux prenne

3. Les SNPs situés sur le même chromosome peuvent être corrélés entre eux et cette corrélation est plus probable lorsque les SNPs sont proches l'un de l'autre sur la séquence d'ADN.

4. Nous avons considéré les Caucasiens parce qu'ils forment la plus grande proportion de la Biobanque.

des médicaments antihypertenseurs. La prise de tels médicaments a un effet sur la PAS et certains de ces médicaments peuvent avoir un effet sur l'IMC. Par exemple, certains agents bêta-bloquants sont associés avec des gains pondéraux dans les premiers mois d'utilisation [**32**]. Pour ces raisons, la prise de médicaments antihypertenseurs peut être une variable de confusion potentielle, dépendamment toutefois des médicaments étudiés et des individus. Nous avons donc construit une variable indiquant si un médicament ayant un effet spécifiquement sur la PAS a été prescrit à l'entrée dans la cohorte. Parmi les médicaments servant à traiter les troubles cardiovasculaires, nous avons indiqué si au moins un des sept médicaments suivants qui agissent sur la PAS a été prescrit : les inhibiteurs calciques, les agents agissant sur le système rénine-angiotensine, les agents bêta-bloquants, les diurétiques, les vasodilatateurs périphériques, les vasoprotecteurs, et les antihypertenseurs.

La description des participants inclus dans l'analyse est présentée au tableau 6.2. Les participants ayant des données incomplètes, c'est-à-dire au moins une observation manquante pour une variable de la base de données finale, n'ont pas été inclus dans l'analyse.

TABLEAU 6.2. Mesures cliniques et descriptives de l'échantillon des participants de la Biobanque ($n = 10\ 455$).

Mesure	
Âge, années	68,3 ± 11,0
Indice de masse corporelle, kg/m^2	28,6 ± 5,3
Pression artérielle systolique, mmHg	125,9 ± 16,8
Homme, %	59
Consommation d'alcool, %	
Jamais	12
Occasions spéciales seulement	27
1-2 fois par mois	5
1-2 fois par semaine	18
3-4 fois par semaine	14
Chaque jour ou presque	24
Statut tabagique, cigarette, %	
Fumeur actuel	9
Ancien fumeur	55
N'a jamais fumé	35
Prise de médicaments agissant sur la PAS, %	84

6.3.3. Données manquantes

Nous avions des données cliniques et descriptives pour $n = 11\ 990$ participants (européens et non-européens) et les données génétiques de $n = 11\ 292$ européens. Après la fusion de ces deux bases de données, nous avions les données génétiques de nos 10 SNPs ainsi que les données cliniques et descriptives des variables du tableau 6.2 pour $n = 10\ 671$ participants européens. De ces 10 671 participants, 216 avaient des données manquantes pour au moins une des variables de l'analyse. Ainsi, nous avons une base de données complètes pour $n = 10\ 455$ participants européens de la Biobanque.

6.4. Méthodes

Nous présentons une analyse pour l'effet de l'IMC (exposition) sur la PAS (issue) avec trois modèles : un modèle conventionnel (OLS) et deux modèles basés sur l'instrumentation, soit 2SLS et sisVIVE.

Contrairement aux estimateurs OLS et 2SLS, où les fonctions du logiciel **R** qui les mettent en œuvre permettent l'ajustement pour d'autres variables, les fonctions du package `sisVIVE` en **R** ne nous permettent pas d'ajouter des variables autres que l'issue, l'exposition et les instruments [**22**]. Donc, l'ajustement pour les variables de confusion potentielles doit se faire en utilisant les résidus de la régression de l'IMC sur ses covariables et de la régression de la PAS sur ses covariables. Nous prenons la même approche pour l'estimation des modèles par OLS et 2SLS afin de préserver la comparabilité des résultats. Nous avons vérifié que la méthode d'ajustement par les résidus mène aux mêmes résultats que l'ajustement conventionnel dans le modèle 2SLS.

Pour chacun des modèles, nous ajustons pour les variables suivantes : l'âge, le sexe, la fréquence de consommation d'alcool, le statut tabagique et l'indicatrice de la prise de médicaments agissant sur la PAS.

Nous utilisons une sélection de 10 SNPs qui sont fortement associés à l'IMC dans la méta-analyse [**40**]. Nous calculons la force de nos instruments dans un modèle de régression linéaire multiple de l'IMC sur les 10 SNPs et rapportons la statistique F. Cette statistique est une mesure de la force globale des instruments, c'est-à-dire, la force du signal allant des instruments à l'IMC. La convention est qu'une valeur de la statistique F inférieure à 10 indique que les instruments sont faibles et que nous pouvons encourir un biais dit « de faiblesse ».

Le traitement des données et l'analyse ont été faits à l'aide des logiciels **R** (version 3.2.4) et **PLINK** (v1.07) [**31**][**30**].

6.5. Résultats

D'abord, nous présentons les résultats de l'estimation de la validité de chacun des 10 SNPs comme instrument pour l'IMC avec les estimations bootstrap de la probabilité que ces SNPs soient invalides selon sisVIVE. Ensuite, nous présentons les résultats de l'estimation de l'effet de l'IMC sur la PAS par les trois modèles (OLS, 2SLS et sisVIVE). Enfin, nous présentons les intervalles de confiance bootstrap pour l'effet estimé par sisVIVE calculés selon trois méthodes (percentile, base et normal). Nous illustrons ces intervalles dans l'histogramme de la densité des estimations bootstrap de l'effet de sisVIVE. Nous présentons également les résultats pour les modèles sans ajustement pour les covariables dans le but de vérifier l'effet d'ajuster pour les variables de confusion potentielles.

Chacun des 10 SNPs est associé à l'IMC ce qui suggère que la supposition (A1) est respectée dans les données, bien que l'association entre les SNPs et l'IMC soit plutôt faible : $F = 3{,}01$ pour le modèle avec ajustement pour les covariables et $F = 3{,}11$ pour le modèle sans ajustement.

Le tableau 6.3 présente les estimations du paramètre d'invalidité des SNPs, $\hat{\boldsymbol{\alpha}}$, dans le modèle de sisVIVE ainsi que les estimations bootstrap de la probabilité qu'un SNP soit estimé comme étant invalide par sisVIVE, basé sur $B = 1000$ échantillons bootstrap. Ces proportions sont rapportées dans les colonnes $Pr\{\hat{\boldsymbol{\alpha}}^* \neq 0\}$ pour les modèles avec et sans ajustement pour les variables de confusion potentielles.

Tableau 6.3. Estimation par sisVIVE du paramètre α et estimation bootstrap ($B = 1000$) de la probabilité qu'un SNP soit signalé invalide.

			Avec ajustement		**Sans ajustement**	
Chr	**SNP**	**Gène**	$\hat{\boldsymbol{\alpha}}$	$Pr\{\hat{\boldsymbol{\alpha}}^* \neq 0\}$	$\hat{\boldsymbol{\alpha}}$	$Pr\{\hat{\boldsymbol{\alpha}}^* \neq 0\}$
2	rs11676272	*ADCY3*	0,000	0,124	0,000	0,160
3	rs295322	*RASA2*	0,000	0,020	0,000	0,013
4	rs34811474	*ANAPC4*	0,000	0,029	0,000	0,021
4	rs13107325	*SLC39A8*	0,000	0,036	0,000	0,046
6	rs9469913	*UHRF1BP1*	0,000	0,048	0,000	0,063
11	rs6265	*BDNF*	0,000	0,116	0,000	0,108
11	rs1064608	*MTCH2*	0,000	0,066	0,000	0,091
16	rs2904880	*CD19*	0,000	0,041	0,000	0,035
16	rs3213758	*RPGRIP1L*	0,000	0,127	0,000	0,141
19	rs1800437	*GIPR*	0,000	0,086	0,000	0,103

Nous pouvons remarquer au tableau 6.3 que sisVIVE estime que tous les SNPs sont des instruments valides autant pour le modèle avec ajustement que pour le modèle sans

ajustement ($\hat{\alpha}_j = 0$). Quant aux estimations bootstrap de la probabilité qu'un SNP soit invalide, nous constatons qu'il y a quelques fois où sisVIVE estime qu'un SNP est invalide, la proportion maximale étant de 12,7%, pour le SNP rs3213758 situé sur le gène *RPGRIP1L* pour le modèle avec ajustement (la proportion pour ce SNP dans le modèle sans ajustement est de 14,1%) et 16,0% pour le SNP rs11676272 situé sur le gène *ADCY3* selon le modèle sans ajustement (la proportion pour ce SNP dans le modèle avec ajustement est de 12,4%).

Le tableau 6.4 présente l'effet de l'IMC sur la PAS estimé selon les trois modèles (OLS, 2SLS et sisVIVE). Pour le modèle avec ajustement, selon la méthode OLS, il existe un effet positif et statistiquement significatif de l'IMC sur la PAS. La méthode estime cet effet très précisément. Quant aux modèles 2SLS et sisVIVE, nous remarquons en premier que les deux méthodes mènent à la même estimation. Tel qu'expliqué au chapitre 3, étant donné que tous les SNPs ont été désignés comme étant valides selon l'algorithme de sisVIVE, l'estimation produite par sisVIVE est identique à l'estimation produite par 2SLS. Seule l'estimation de la variabilité diffère entre les deux méthodes. Cependant, ceci s'explique par le fait que deux méthodes différentes sont utilisées pour estimer l'erreur standard : pour 2SLS, c'est l'erreur standard résiduelle du modèle 2SLS divisée par la somme des carrés totale de l'IMC et le coefficient R^2 de la régression linéaire multiple de l'IMC sur les SNPs, alors que pour sisVIVE c'est la racine carrée de l'estimation bootstrap de la variance. L'effet estimé de l'IMC sur la PAS est négatif et statistiquement non significatif. Les résultats sont essentiellement les mêmes pour le cas sans ajustement pour les covariables.

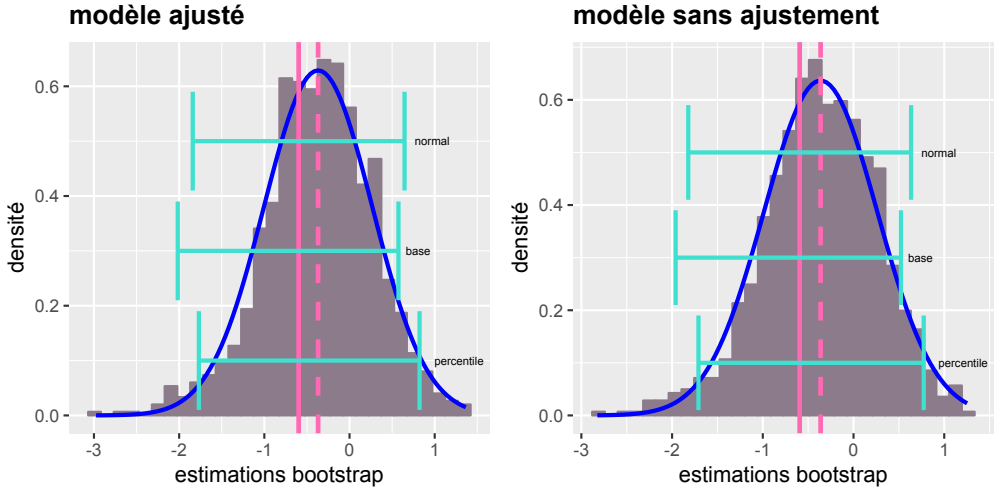
TABLEAU 6.4. Estimations de l'effet causal de l'IMC sur la PAS, avec et sans ajustement pour les covariables.

		$\hat{\beta}$	**ES**[a]	**valeur-p**[b]
avec ajustement	OLS	0,45	0,03	$\leq$ 2e-16
	2SLS	-0,60	0,61	0,325
	sisVIVE	-0,60	0,63	0,347
sans ajustement	OLS	0,46	0,03	$\leq$ 2e-16
	2SLS	-0,59	0,59	0,316
	sisVIVE	-0,59	0,63	0,344

a. L'erreur standard est estimée différemment selon le modèle. Pour OLS et 2SLS, c'est simplement l'élément (2,2) de la matrice de la covariance de l'estimation par OLS et l'erreur standard résiduelle du modèle 2SLS divisée par la somme des carrés totale de l'IMC et le coefficient R^2 de la régression linéaire multiple de l'IMC sur les SNPs, respectivement. Pour sisVIVE, c'est la racine carrée de l'estimation bootstrap de la variance.
b. La valeur-p est calculée à partir d'un test Z en bilatéral.

La figure 6.1 présente l'histogramme de la densité des estimations bootstrap de sisVIVE, $\hat{\beta}^*_{sisVIVE}$, avec leur moyenne basée sur $B = 1000$ échantillons bootstrap (ligne pointillée), l'estimation de l'effet par sisVIVE ($\hat{\beta}_{sisVIVE}$, ligne pleine) et les trois types d'intervalle de confiance bootstrap pour les modèles avec et sans ajustement pour les covariables. Nous constatons que la distribution des estimations bootstrap de sisVIVE comporte des valeurs plus extrêmes ($\hat{\beta}^*_{502} = -2{,}98$ et $\hat{\beta}^*_{608} = -2{,}73$ pour le modèle ajusté, $\hat{\beta}^*_{502} = -2{,}82$ et $\hat{\beta}^*_{17} = -2{,}60$ pour le modèle sans ajustement). La taille de ces valeurs est cliniquement plausible, mais le signe négatif ne l'est pas. Nous illustrons également le biais de l'estimation par bootstrap (la différence entre la ligne pleine et la ligne pointillée) qui est non négligeable (biais = 0,23 pour chacun des modèles). Dans cette même figure, nous pouvons constater que tous les intervalles de confiance mènent à la conclusion que l'effet estimé par sisVIVE est statistiquement non significatif. Le tableau 6.5 présente ces intervalles.

FIGURE 6.1. Histogrammes des estimations bootstrap ($B = 1000$) de sisVIVE, $\hat{\beta}^*_{sisVIVE}$, avec les intervalles de confiance bootstrap de niveau 95% pour l'estimation de l'effet causal de l'IMC sur la PAS par sisVIVE.



La ligne pleine correspond à l'estimation $\hat{\beta}_{sisVIVE}$ et la ligne pointillée correspond à la moyenne des estimations bootstrap $\hat{\beta}^*_{sisVIVE}$. Le biais est la différence entre ces deux lignes.

Les bornes diffèrent selon le type d'intervalle et il en va de même pour la longueur. Pour le modèle ajusté, celle-ci est autour de 2,58 pour les intervalles de base et percentile, qui sont de la même longueur par construction, alors que l'intervalle normal a une longueur de 2,49. Pour le modèle sans ajustement, la longueur est de 2,48 pour les intervalles de base et percentile, tandis que l'intervalle normal a une longueur de 2,45. Enfin, nous remarquons

TABLEAU 6.5. Estimation et intervalles de confiance bootstrap ($B = 1000$) de niveau 95% pour l'effet causal de l'IMC sur la PAS estimé par sisVIVE.

	$\hat{\beta}_{sisVIVE}$	**percentile**	**base**	**normal**
avec ajustement	-0,60	[-1,77 ; 0,82]	[-2,02 ; 0,57]	[-1,84 ; 0,65]
sans ajustement	-0,59	[-1,71 ; 0,77]	[-1,96 ; 0,52]	[-1,82 ; 0,63]

que tous les intervalles contiennent l'effet estimé par le modèle OLS ($\hat{\beta}_{OLS} = 0{,}45$ pour le modèle ajusté et $\hat{\beta}_{OLS} = 0{,}46$ pour le modèle sans ajustement).

6.6. DISCUSSION

L'analyse précédente a pour but de déterminer si l'obésité cause un changement dans la pression artérielle en utilisant l'estimateur sisVIVE en randomisation mendélienne. Nous avons sélectionné 10 SNPs associés à l'IMC à partir d'une méta-analyse répertoriant des SNPs sur la puce de l'exome [**40**]. Nous avons utilisé ces SNPs dans le modèle sisVIVE pour estimer l'effet de l'IMC sur la PAS. Rappelons que nous utilisons sisVIVE pour deux raisons : (1) parce qu'il est plus robuste à l'utilisation de certains SNPs invalides (e.x. par la pléiotropie) que la méthode conventionnelle 2SLS, et (2) parce qu'il permet d'obtenir une estimation de l'invalidité des SNPs. Nous avons appliqué l'algorithme bootstrap pour sisVIVE présenté au chapitre 4 pour calculer différents types d'intervalles de confiance et pour estimer la probabilité que les SNPs soient estimés comme étant invalides.

Dans un premier temps, nous avons trouvé que tous les 10 SNPs ont été estimés comme étant des instruments valides pour l'IMC par la méthode sisVIVE. Rappelons que dans le cas où tous les instruments sont estimés valides, le modèle sisVIVE produit une estimation de l'effet qui est identique à celle du modèle 2SLS et ce que tous les instruments soient réellement valides ou non. Incidemment, ce résultat est appuyé par le fait que les effets estimés par la méthode 2SLS et par la méthode sisVIVE sont exactement les mêmes.

Dans un deuxième temps, selon la littérature médicale, une augmentation de l'IMC est associée avec une augmentation de la PAS de façon linéaire [**42**]. Cependant, selon nos modèles 2SLS et sisVIVE, nous avons estimé un effet négatif non significatif de l'IMC sur la PAS. Les intervalles de confiance bootstrap pour sisVIVE sont très larges et l'effet positif et significatif estimé par OLS est contenu dans tous les intervalles de confiance bootstrap. Ainsi, d'un côté, avec la méthode OLS, nous observons que l'IMC a l'effet d'augmenter la PAS et l'estimation produite est très précise. Si le modèle OLS est effectivement biaisé en raison de la confusion, alors la méthode rapporte un effet biaisé avec peu d'incertitude, ce qui n'est pas souhaitable. De l'autre côté, avec les méthodes par les variables instrumentales, nous observons qu'il n'y a pas évidence d'effet de l'IMC sur la PAS parce que les

intervalles sont trop larges, mais puisqu'ils sont assez larges, il se peut qu'ils contiennent le vrai effet de l'IMC sur la PAS sans toutefois avoir la précision nécessaire pour l'estimer de façon statistiquement significative. Ce manque de précision pourrait être attribuable, entre autres, au fait de ne pas avoir des instruments suffisamment forts pour l'IMC, ce que nous discutons subséquemment.

Dans un troisième temps, notre sélection de SNPs mène à des instruments faibles pour l'IMC. Selon les résultats de l'étude par simulations de Kang et al. (2014), l'estimateur sisVIVE est susceptible au biais engendré par la faiblesse des instruments, tout comme l'estimateur 2SLS. Dans le cas de 2SLS, plus il y a d'instruments faibles, plus le biais devient important [**7**]. Une des raisons explicant la faiblesse des instruments est que nous avons utilisé seulement les SNPs répertoriés sur la puce de l'exome étant donné que les participants de la Biobanque sont génotypés avec cette puce. Nous étions alors limités dans la sélection des SNPs parce que nous tenions à utiliser les résultats d'une méta-analyse de SNPs répertoriés sur la puce de l'exome. Ainsi, la liste des SNPs sur laquelle notre sélection se base n'inclut pas tous les SNPs que la littérature a établi comme étant fortement associés avec l'obésité par le phénotype de l'IMC. Par exemple, nous n'avons aucun SNP situé sur le gène FTO, dont l'effet sur l'obésité est reconnu [**14**]. Nous pourrions utiliser des SNPs situés sur une puce plus standard que celle de l'exome, auquel cas nous devrions utiliser des données génétiques imputées, mais c'est au-delà de l'objectif du mémoire.

Enfin, bien que nous trouvons qu'il n'y a pas évidence qu'une augmentation de l'IMC cause une augmentation de la PAS selon les méthodes 2SLS et sisVIVE, ce résultat peut tout aussi bien être imputable à la faiblesse des instruments dans nos analyses. Puisque la littérature dit plutôt que l'obésité a l'effet d'augmenter la pression artérielle, si nous nous tenions à nos résultats sur la significativité de l'effet seulement, notre conclusion serait surprenante. Or, nos résultats ne contredisent pas forcément la littérature, étant donné que les intervalles de confiance pour l'effet estimé par sisVIVE sont très larges, allant jusqu'à inclure des valeurs positives qui sont plus plausibles pour cet effet.

Une limite importante à souligner est le fait que nous tentons de contrôler pour l'effet de la prise des médicaments agissant sur la PAS en ajustant pour une variable indiquant s'il y a prise de tels médicaments ou non, ce qui revient essentiellement à introduire une constante dans les modèles. Cette méthode a déjà été utilisée dans la littérature (p.e. [**27**], où seulement 9,3% de l'échantillon prend des médicaments agissant sur la pression artérielle). Il est possible que pour notre échantillon, cet ajustement soit insuffisant pour expliquer l'effet de la prise des médicaments, d'autant plus que 84% des individus dans notre échantillon prennent des médicaments agissant sur la PAS.

Cet exemple illustre notre méthode d'inférence pour sisVIVE dans un contexte d'instruments faibles. Cependant, ce dernier ne fait pas partie des scénarios des simulations

que nous avons considéré, parce que c'est au-delà des objectifs du mémoire. Nous avons néanmoins trouvé que les intervalles de confiance bootstrap calculés pour l'effet de sis-VIVE donnaient des résultats plutôt fiables, quoique peu probants, tenant en compte le biais de faiblesse.

Chapitre 7

CONCLUSION

Dans ce mémoire, nous avons présenté une méthode par bootstrap pour calculer des intervalles de confiance pour un effet causal estimé par la méthode sisVIVE [**23**] en randomisation mendélienne. Nous avons d'abord expliqué comment la randomisation mendélienne utilise l'assignation aléatoire des gènes qui se fait à la naissance, ce qui équilibre la répartition des variables de confusion dans les groupes de l'exposition. Ceci permet de gérer la confusion présente dans les données issues d'études observationnelles et ainsi pouvoir inférer la causalité de l'exposition sur l'issue.

Nous avons illustré, à l'aide de DAGs et d'exemples, les trois suppositions qu'un SNP doit satisfaire pour être un instrument valide en randomisation mendélienne et ainsi contribuer à l'estimation de l'effet causal. Ces suppositions sont :

(A1) le SNP est corrélé avec l'exposition,

(A2) le SNP et l'issue n'ont pas une cause commune qui précède l'assignation du SNP à l'individu dans le temps,

(A3) le SNP n'a pas d'effet pléiotropique sur l'issue, donc il n'affecte pas l'issue par un autre chemin que celui de l'exposition.

Nous avons par la suite présenté deux estimateurs à instruments multiples qui combinent les effets de plusieurs SNPs dans un modèle d'estimation en randomisation mendélienne. Nous avons expliqué la contribution et les limites de chacun de ces deux estimateurs. Dans un premier temps, l'estimateur 2SLS est fréquemment utilisé et repose sur la supposition que tous les SNPs sont valides selon (A1)-(A3), ce qui est impossible à vérifier en pratique. Dans un deuxième temps, l'estimateur sisVIVE trouve une solution unique au problème d'estimation du type LASSO si le nombre de SNPs invalides utilisés ne dépasse pas le seuil de 50%. Des simulations démontrent que lorsque ce seuil est respecté, sisVIVE a une performance supérieure à 2SLS en terme de biais d'estimation. Cependant, sisVIVE ne

fournit qu'une estimation ponctuelle et aucune méthode de construction d'intervalle de confiance n'a été proposée.

Pour remédier à ce problème, nous avons présenté le bootstrap comme solution pour calculer des intervalles de confiance pour l'estimation d'un effet causal par sisVIVE. Nous avons expliqué comment calculer des intervalles de confiance bootstrap percentiles, de base et normal à partir de l'échantillon original des données. Nous avons également suggéré d'utiliser l'estimation du paramètre d'invalidité, qui est un produit secondaire de l'estimation avec sisVIVE, pour estimer par la méthode bootstrap la probabilité qu'un SNP soit sélectionné comme étant invalide.

Dans notre étude, nous avons simulé, pour trois tailles d'échantillon différentes, un cas où tous les SNPs sont valides selon (A1)-(A3), un cas où 30% des SNPs sont invalides en raison d'un effet pléiotropique sur l'issue et un cas où 70% des SNPs sont invalides par la pléiotropie. Par notre analyse de la performance du bootstrap pour sisVIVE, nous concluons que pour les cas où les SNPs sont tous valides, la méthode d'inférence par bootstrap produit des intervalles de confiance qui atteignent la couverture désirée de 95%. Des trois types d'intervalles de confiance considérés, l'intervalle de confiance bootstrap de base offre la meilleure couverture dans toutes les tailles d'échantillon. Conjointement au fait que l'estimation bootstrap de la probabilité d'invalidité d'un SNP identifie avec faible erreur qu'aucun SNP n'est invalide, nos résultats suggèrent qu'il est possible de faire l'inférence pour l'estimation produite par sisVIVE aussitôt que l'algorithme signale que les instruments sont valides.

Pour les cas où certains SNPs sont invalides, nous devons conclure différemment selon que le seuil de 50% est respecté ou est dépassé. Nos simulations soutiennent la contribution principale de sisVIVE, soit que la méthode parvient à estimer l'effet causal avec un biais très faible lorsque certains SNPs sont invalides et plus de 50% sont valides. Dans ces cas, notre méthodologie d'inférence par bootstrap est aussi applicable, mais la couverture de 95% est atteinte seulement dans le petit échantillon. D'autres simulations seraient nécessaires pour déterminer le comportement de notre méthodologie dans ce cas particulier.

Lorsque plus de 50% des SNPs sont invalides par la pléiotropie, nos simulations révèlent que l'estimation par sisVIVE est effectivement biaisée, mais ce biais devient plus petit dans de grands échantillons. Un problème pour ces cas est que la méthode sisVIVE peut parfois produire des estimations extrêmes de l'effet causal qui sont tout de même cliniquement plausibles. De plus, pour ce qui est de notre méthodologie bootstrap, la méthode peut parfois (au plus 5,5% des fois dans le petit échantillon) produire des estimations extrêmes de la variance, comme dans l'estimation de l'effet. Pour l'estimation avec sisVIVE, parce qu'il y a des fois où les valeurs extrêmes seront plausibles, d'autres fois non, et qu'il n'y a pas de seuils clairs, cela reviendra donc au chercheur ou a la chercheuse de faire un

tri judicieux. Par contre, pour les estimations extrêmes de la variance, celles-ci sont loin d'être cliniquement plausibles (de l'ordre de 1×10^{19} et plus) et seraient facilement mises de côté si un jeu de données réelles retrouvait ce même résultat. Enfin, nous trouvons que dans les grands échantillons, les intervalles de confiance bootstrap de base parviennent à atteindre une couverture de 95%, ce qui nous rassure de l'utilité potentielle de la méthode.

Nous avons également étudié notre méthodologie dans un jeu de données réelles provenant de la Biobanque de l'Institut de Cardiologie de Montréal. Dans l'analyse, nous avons voulu estimer l'effet causal de l'indice de masse corporelle sur la pression artérielle systolique, mais nous sommes arrivés à des résultats peu probants en raison de la faible force d'association entre les SNPs et l'indice de masse corporelle. De plus, la méthode bootstrap a mené à des intervalles de confiance très larges et une estimation ponctuelle incertaine. Nous nous sommes limités à l'utilisation de SNPs situés sur l'exome parce que c'est la puce utilisée pour le génotypage des participants de la Biobanque, mais en utilisant des SNPs imputés nous pourrions possiblement inclure des SNPs ayant une prévision plus forte de l'indice de masse corporelle : par exemple, les SNPs du gène FTO. Ceci pourrait améliorer la force globale des instruments et améliorer l'estimation par les méthodes 2SLS et sisVIVE.

Nous croyons que ce mémoire fait un premier pas dans l'exploration du bootstrap pour calculer des intervalles de confiance pour sisVIVE. En effet, nous avons obtenu quelques résultats encourageants, particulièrement pour les cas où tous les instruments sont valides. Cependant, il est important de rappeler que, étant donné les scénarios considérés dans nos simulations, ce que nous avons conclu de notre étude de la méthodologie d'inférence par bootstrap pour sisVIVE s'applique aux cas où les SNPs ont une association forte avec l'exposition et ont une force relativement égale entre eux. Puisque nous ne pouvions pas couvrir l'ensemble des scénarios possibles, d'autres intérêts seraient l'étude des cas où il y a faiblesse des instruments, ces derniers étant problématiques en théorie et fréquents en pratique (p.e. [**7**]).

Bibliographie

[1] Scitable Nature, definitions. `http://www.nature.com/scitable`, 2001. [En ligne ; accédé 8 Décembre, 2016].

[2] La Biobanque de la Cohorte hospitalière de l'ICM. `https://www.icm-mhi.org/fr/recherche/infrastructures-services/biobanque/quest-ce-que-biobanque`, 2018. [En ligne ; accédé 19 Mars, 2018].

[3] L'Institut de Cardiologie de Montréal, à propos du centre de recherche. `https://www.icm-mhi.org/fr/recherche/propos-centre-recherche`, 2018. [En ligne ; accédé 19 Mars, 2018].

[4] Science Direct Topics : The Hardy-Weinberg Principle,. `https://www.sciencedirect.com/topics/neuroscience/hardy-weinberg-principle`, 2018. [En ligne ; accédé 9 Août, 2018].

[5] Jack Bowden, George Davey Smith et Stephen Burgess : Mendelian randomization with invalid instruments : effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2) :512–525, 2015.

[6] Stephen Burgess et Simon G Thompson : *Mendelian randomization : methods for using genetic variants in causal estimation*. CRC Press, 2015.

[7] Stephen Burgess, Simon G Thompson *et al.* : Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, 40(3) :755–764, 2011.

[8] David Clayton et Michael Hills : *Statistical Models in Epidemiology*, volume 161. Oxford University Press, 1993. ISBN : 0198522215, 9780198522218.

[9] Russell Davidson et James G MacKinnon : *Econometric Theory and Methods*, volume 5. Oxford University Press, 2004.

[10] Anthony Christopher Davison et David Victor Hinkley : *Bootstrap Methods and their Application*, volume 1. Cambridge University Press, 1997.

[11] Richard Doll et A Bradford Hill : Smoking and carcinoma of the lung. *British Medical Journal*, 2(4682) :739, 1950.

[12] Tamara Dubowitz, Madhumita Ghosh-Dastidar, Christine Eibner, Mary E Slaughter, Meenakshi Fernandes, Eric A Whitsel, Chloe E Bird, Adria Jewell, Karen L

Margolis, Wenjun Li *et al.* : The Women's Health Initiative : the food environment, neighborhood socioeconomic status, BMI, and blood pressure. *Obesity*, 20(4) :862–871, 2012.

[13] Bradley Efron et Robert J. Tibshirani : *An Introduction to the Bootstrap*. Chapman & Hall, 1994.

[14] Katherine A Fawcett et Inês Barroso : The genetics of obesity : FTO leads the way. *Trends in Genetics*, 26(6) :266–274, 2010.

[15] Ronald A Fisher : Lung cancer and cigarettes ? *Nature*, 182(4628) :108, 1958.

[16] Oscar H Franco, Anna Peeters, Luc Bonneux et Chris De Laet : Blood pressure in adulthood and life expectancy with cardiovascular disease in men and women : life course analysis. *Hypertension*, 46(2) :280–286, 2005.

[17] Jerome Friedman, Trevor Hastie et Robert Tibshirani : *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics, 2001.

[18] M Maria Glymour, Eric J Tchetgen Tchetgen et James M Robins : Credible Mendelian randomization studies : approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4) :332–339, 2012.

[19] David Hall : SNP. `http://www.nature.com/scitable/content/snp-4815`, 2007. [En ligne ; accédé 8 Décembre, 2016].

[20] Farhad Hormozdiari, Gleb Kichaev, Wen-Yun Yang, Bogdan Pasaniuc et Eleazar Eskin : Identification of causal genes for complex traits. *Bioinformatics*, 31(12) :i206–i213, 2015.

[21] Mark M Iles, Matthew H Law, Simon N Stacey, Jiali Han, Shenying Fang, Ruth Pfeiffer, Mark Harland, Stuart MacGregor, John C Taylor, Katja K Aben *et al.* : A variant in FTO shows association with melanoma risk not due to BMI. *Nature Genetics*, 45(4) :428–432, 2013.

[22] Hyunseung Kang : *sisVIVE : Some Invalid Some Valid Instrumental Variables Estimator*, 2017. R package version 1.4.

[23] Hyunseung Kang, Anru Zhang, T Tony Cai et Dylan S Small : Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513) :132–144, 2016.

[24] Mika Kivimäki, Markus Jokela, Mark Hamer, John Geddes, Klaus Ebmeier, Meena Kumari, Archana Singh-Manoux, Aroon Hingorani et G. David Batty : Examining overweight and obesity as risk factors for common mental disorders using fat mass and obesity-associated (FTO) genotype-instrumented analysis the Whitehall II Study, 1985–2004. *American Journal of Epidemiology*, 173(4) :421–429, 2011.

[25] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson et George Davey Smith : Mendelian randomization : using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8) :1133–1163, 2008.

[26] Christian Léger, Dimitris N Politis et Joseph P Romano : Bootstrap technology and applications. *Technometrics*, 34(4) :378–398, 1992.

[27] Donald M Lyall, Carlos Celis-Morales, Joey Ward, Stamatina Iliodromiti, Jana J Anderson, Jason MR Gill, Daniel J Smith, Uduakobong Efanga Ntuk, Daniel F Mackay, Michael V Holmes *et al.* : Association of body mass index with cardiometabolic disease in the UK Biobank : a Mendelian randomization study. *JAMA Cardiology*, 2(8) :882–889, 2017.

[28] Elie Mouhayar et Abdulla Salahudeen : Hypertension in cancer patients. *Texas Heart Institute Journal*, 38(3) :263, 2011.

[29] Tom M Palmer, Debbie A Lawlor, Roger M Harbord, Nuala A Sheehan, Jon H Tobias, Nicholas J Timpson, George Davey Smith et Jonathan AC Sterne : Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research*, 21(3) :223–242, 2012.

[30] Shaun Purcell, Neale B, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender *et al.* : PLINK : a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81, 2007.

[31] Purcell, Shaun : PLINK v1.07. `http://pngu.mgh.harvard.edu/purcell/plink/`.

[32] Arya M. Sharma, Tobias Pischon, Sandra Hardt, Iris Kunz et Friedrich C. Luft : Hypothesis : β-adrenergic receptor blockers and weight gain. *Hypertension*, 37(2) :250–254, 2001.

[33] George Davey Smith et Shah Ebrahim : Mendelian randomization : prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1) :30–42, 2004.

[34] Frank W Stearns : One hundred years of pleiotropy : a retrospective. *Genetics*, 186(3) :767–773, 2010.

[35] Fikru Tesfaye, Peter Byass et Stig Wall : Population based prevalence of high blood pressure among adults in Addis Ababa : uncovering a silent epidemic. *BMC Cardiovascular Disorders*, 9(1) :39, 2009.

[36] Henri Theil : *Principles of Econometrics*. Wiley, 1971.

[37] Henri Theil : Estimation and simultaneous correlation in complete equation systems. *In Henri Theils Contributions to Economics and Econometrics*, pages 65–107. 1992.

[38] Robert Tibshirani : Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[39] Ryan Tibshirani : Lecture notes in Data Mining. model selection and validation 1 : Cross-validation, March 2013.

[40] Valérie Turcot, Yingchang Lu, Heather M Highland, Claudia Schurmann, Anne E Justice, Rebecca S Fine, Jonathan P Bradfield, Tõnu Esko, Ayush Giri, Mariaelisa Graff *et al.* : Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nature Genetics*, 50(1) :26, 2018.

[41] Sanford Weisberg : *Applied Linear Regression*, volume 528. John Wiley & Sons, 2005.

[42] G Whitlock, S Lewington *et al.* : Body-mass index and cause-specific mortality in 900 000 adults : collaborative analyses of 57 prospective studies. *The Lancet*, 373(9669) :1083–1096, 2009.

[43] Philip Green Wright : *The tariff on animal and vegetable oils*. Investigations in international commercial policies. The Macmillan Company, 1928.

Annexe A

CODE R

A.1. Fonctions

A.1.1. Générer des SNPs en équilibre de Hardy-Weinberg

```
#SNP
function(n, MAF,index){
  sample(0:2, size=n, prob=HWG(n, MAF[index]), replace=T)
}

#HWG
function(n, MAF){
  q <- MAF                    # matrice de MAFs pour L differents SNPs
  q2 <- q*q
  p <- 1-q
  p2 <- p*p
  return(cbind(p2, 2*p*q,q2)) # matrice des probabilités HW pour L differents SNPs
  }
```

A.1.2. Calculer les estimateurs sisVIVE dans l'algorithme bootstrap

```
# sisVIVE.fun2
function(DATA,g,index){
  G<-g+2
  Y<-as.vector(DATA[index,1])
  D<-as.vector(DATA[index,2])
  Z<-as.matrix(DATA[index,3:G])

  out<-cv.sisVIVE(Y,D,Z,K=10)
```

```
  return( c(out$beta, out$alpha) )
}
```

A.1.3. Calculer l'estimateur 2SLS dans l'algorithme bootstrap

```
# tsls.fun2
function(DATA, g, index){
  G<-g+2

  Y<-as.vector(DATA[index,1])
  D<-as.vector(DATA[index,2])
  Z<-as.matrix(DATA[index,3:G])

  out<-summary(ivreg(Y~D|Z))

  return(out$coef[2,1])
}
```

A.1.4. Identifier les SNPs comme étant invalides

```
# isZero
function(x){
  ifelse(x==0,0, 1)
  }
```

A.2. Algorithmes pour l'inférence par le bootstrap

A.2.1. Estimer la variance de sisVIVE et calculer les quantiles 2,5% et 97,5% de la distribution des estimations bootstrap

```
B<-1000
out.boot<-boot(cbind(Y,D,Z), statistic=sisVIVE.fun2, g=L, R=B, stype="i")
betas<-out.boot$t[,1]
var.boot<-var(betas)
IC.perc95<-as.vector( quantile(betas, c(0.025, 0.975)) ) # intervalle percentile
```

A.2.2. Estimer les probabilités d'invalidité

```
alphas<-out.boot$t[,2:(L+1)]
invalid.boot<-matrix(apply(apply(alphas, 2, isZero), 2, sum),nrow=1)
```

Annexe B

DÉVELOPPEMENTS SUPPLÉMENTAIRES

B.1. Développements supplémentaires

B.1.1. Développement du modèle d'estimation de l'effet causal avec une cause commune omise pour le cas des instruments invalides

Puisque $\mathbf{U}$ est non mesurée, son effet sera absorbé dans le terme d'erreur. D'abord, pour le modèle de l'exposition, nous faisons le développement suivant :

$$\begin{aligned}\mathbf{D} &= \pi^D \mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\boldsymbol{D}} + \mathbf{U}\delta^D \\ &= (\pi^D + \mathbb{E}(\mathbf{U}\delta^D))\mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{\boldsymbol{D}} + \mathbf{U}\delta^D - \mathbb{E}(\mathbf{U}\delta^D) \\ &= c^D \mathbf{1} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}^{\mathbf{D}},\end{aligned} \tag{B.1.1}$$

où $c^D = \pi^D + \mathbb{E}(\mathbf{U}\delta^D)$ et $\mathbf{e}^{\mathbf{D}} = \boldsymbol{\epsilon}^{\boldsymbol{D}} + \mathbf{U}\delta^D - \mathbb{E}(\mathbf{U}\delta^D)$. Ensuite, pour le modèle de l'issue, nous faisons le développement similaire :

$$\begin{aligned}\mathbf{Y} &= \pi^Y \mathbf{1} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{D}\beta + \boldsymbol{\epsilon}^{\boldsymbol{Y}} + \mathbf{U}\delta^Y \\ &= (\pi^Y + \mathbb{E}(\mathbf{U}\delta^Y))\mathbf{1} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{D}\beta + \boldsymbol{\epsilon}^{\boldsymbol{Y}} + \mathbf{U}\delta^Y - \mathbb{E}(\mathbf{U}\delta^Y) \\ &= c^Y \mathbf{1} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{D}\beta + \mathbf{e}^{\mathbf{Y}},\end{aligned} \tag{B.1.2}$$

où $c^Y = \pi^Y + \mathbb{E}(\mathbf{U}\delta^Y)$ et $\mathbf{e}^{\mathbf{Y}} = \boldsymbol{\epsilon}^{\boldsymbol{Y}} + \mathbf{U}\delta^Y - \mathbb{E}(\mathbf{U}\delta^Y)$.

Ce cas mène aux mêmes conclusions que celui pour les instruments valides : les termes d'erreur $\mathbf{e}^{\mathbf{D}}$ et $\mathbf{e}^{\mathbf{Y}}$ satisfont $\mathbb{E}(\mathbf{e}^{\mathbf{D}}) = \mathbb{E}(\mathbf{e}^{\mathbf{Y}}) = 0$, mais puisqu'ils sont tous les deux des fonctions de $\mathbf{U}$, l'exposition $\mathbf{D}$ varie avec le terme d'erreur du modèle de l'issue, $\mathbf{e}^{\mathbf{Y}}$. Par conséquent, l'effet estimé par les moindres carrés ordinaires n'est pas une estimation de l'effet causal de l'exposition $\mathbf{D}$ sur l'issue $\mathbf{Y}$.