Développement d'algorithmes de Plasmode longitudinaux pour l'évaluation d'approches d'ajustement pour la confusion et illustration pour l'étude de l'effet d'une exposition cumulée aux stresseurs psychosociaux au travail

Mémoire

Youssra Souli

Maîtrise en statistique - avec mémoire
Maître ès sciences (M. Sc.)

Québec, Canada

Développement d'algorithmes de Plasmode longitudinaux pour l'évaluation d'approches d'ajustement pour la confusion et illustration pour l'étude de l'effet d'une exposition cumulée aux stresseurs psychosociaux au travail

Mémoire

Youssra Souli

Sous la direction de:

Denis Talbot, directeur de recherche
Xavier Trudel, codirecteur de recherche

Résumé

Le biais de confusion peut affecter tous les types d'études d'observation. Il apparaît lorsque la caractéristique étudiée est associée à un facteur de perturbation complémentaire et que ce dernier fait croire à l'existence d'une relation de cause à effet entre la caractéristique étudiée et l'issue. Des méthodes d'ajustement pour le biais de confusion, notamment les modèles structurels marginaux, peuvent être utilisées pour corriger ce type de biais. Ces modèles n'ont toutefois été utilisés qu'une seule fois pour l'étude de l'effet d'une exposition cumulative aux stresseurs psychosociaux au travail sur la pression artérielle.

L'objectif principal de ce mémoire était de comparer différents estimateurs des paramètres d'un modèle structurel marginal à des approches classiques. Nous avons considéré les estimateurs par pondération inverse de la probabilité de traitement, le calcul-g, le maximum de vraisemblance ciblé avec et sans *SuperLearner*. Ces estimateurs ont d'abord été utilisés pour estimer l'effet d'une exposition cumulée aux stresseurs psychosociaux au travail sur la pression artérielle systolique dans le cadre d'une étude de cohorte prospective de 5 ans. Cette analyse a révélé des différences significatives entre les estimateurs. Puisqu'il s'agit de données réelles, il est toutefois impossible de déterminer quelle méthode produit les résultats les plus valides.

Pour répondre à cette question, nous avons développé deux algorithmes de simulation de données longitudinales de type Plasmode, l'un utilisant des modèles paramétriques et l'autre utilisant des approches non paramétriques. Les simulations Plasmode combinent des données réelles et des données synthétiques pour étudier les propriétés dans un contexte connu, mais similaire au contexte réel.

Au vue des résultats, nous avons conclu que les modèles structurels marginaux représentent des approches pertinentes pour estimer l'effet des stresseurs psychosociaux au travail. Nous recommandons particulièrement d'utiliser la méthode de maximum de vraisemblance ciblé avec et sans *SuperLearner*. Cependant, cela nécessite un effort supplémentaire en termes d'implantation de code et de temps d'exécution.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je voudrais dans un premier temps remercier, mon directeur de recherche Monsieur Denis Talbot pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion. Son soutien et ses encouragements ont sans doute eu une grande influence sur mes travaux de recherche. Je tiens également à remercier Monsieur Xavier Trudel pour ses précieuses remarques et les échanges fructueux avec lui. Il a été d'un grand soutien dans l'élaboration de ce mémoire.

Je désire aussi remercier les professeurs de l'Université Laval, qui m'ont fourni les outils nécessaires à la réussite de mes études universitaires.

Je remercie mes très chers parents, qui ont toujours été là pour moi. Je remercie mes sœurs pour leurs encouragements. Enfin, je remercie mes amis qui ont toujours été là pour moi. Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide.

Introduction

L'inférence causale est un processus qui permet de tirer des conclusions sur des relations de cause à effet à l'aide d'observations empiriques. Bien que son processus soit généralement complexe, il est d'une grande importance et ses applications sont en évolution continue, surtout dans le domaine de la santé. Ainsi, la causalité est devenue l'une des principales préoccupations de plusieurs branches de la médecine, pour estimer l'effet d'un traitement potentiel, d'une campagne ou d'une politique sur la santé des individus et ainsi apporter des bénéfices supplémentaires au développement de la santé publique.

L'inférence causale peut généralement s'effectuer assez facilement à l'aide d'études randomisées [Hernán (2004)]. Cependant, il est parfois impossible de réaliser des études randomisées pour des raisons éthiques, financières ou logistiques. Les décideurs doivent donc souvent se baser sur les informations provenant d'études observationnelles.

Les applications des méthodes d'inférence causale, en contexte d'études observationnelles, ont contribué de façon majeure à l'avancement des connaissances dans plusieurs domaines. Par exemple :

* Explorer l'effet du tabagisme sur l'incidence du cancer du poumon, [Godtfredsen et al. (2005)], en utilisant les modèles à risques proportionnels de Cox. Les résultats de cette étude ont montré que, pour les personnes qui fument 15 cigarettes ou plus par jour, une réduction de 50% du tabagisme réduit considérablement le risque de cancer du poumon.
* Examiner l'effet de la consommation d'Aspirine sur le risque d'atteinte de cancer chez les femmes en bonne santé, [Cook et al. (2005)]. Des modèles à risques proportionnels de Cox ont été utilisés pour estimer les rapports de taux (RT) et les intervalles de confiance (IC) à 95% associés aux femmes qui prennent l'Aspirine. Il a été montré que l'utilisation d'Aspirine à faible dose (100 mg) tous les deux jours, pendant une moyenne de 10 ans de traitement, ne diminue pas le risque de cancer du sein.
* Étudier l'effet de l'allaitement sur les infections gastro-intestinales chez les nouveau-nés, [Schnitzer et al. (2014)]. En raison de la présence d'une confusion dépendante du temps, des méthodes d'estimation causale spécialisées sont utilisées, telles que l'estimation par maximum de vraisemblance ciblée avec et sans *SuperLearner*, le calcul-g, la pondération par l'inverse de probabilité de traitement. Toutes les méthodes ont convenu que l'allongement de la durée de l'allaite-

ment réduit considérablement le nombre attendu d'infections gastro-intestinales. En particulier, l'estimation par maximum de vraisemblance ciblée avec *SuperLearner* a donné les meilleurs résultats.

* Explorer l'effet de prendre la zidovudine sur la survie des hommes atteints du VIH [Hernán et al. (2000)]. Un modèle structurel marginal à risques proportionnels de Cox a été utilisé pour estimer l'effet causal de la zidovudine sur la mortalité des patients séropositifs. Le taux ajusté de mortalité pour la zidovudine était de 3,6.

Dans les études observationnelles, différents facteurs peuvent influencer à la fois le niveau d'exposition des sujets et leur réponse observée. Dans ce cas, il peut survenir un biais de confusion qui a pour origine la non comparabilité des deux groupes d'individus (exposés / non exposés).

Dans ce cadre, plusieurs méthodes en inférence causale ont été proposées pour contrôler ces biais de confusion, telles que les méthodes classiques de régression (modèles mixtes). Ces approches traditionnelles ne permettent cependant pas d'ajuster les biais de confusion lorsqu'on s'intéresse à l'effet d'une exposition cumulative, c'est-à-dire l'effet d'une collection de mesures d'exposition dans le temps.

En effet, lorsqu'on s'intéresse à une exposition cumulative, c'est-à-dire l'effet d'une exposition sur une période de temps prolongée, il existe souvent des covariables qui ont un double rôle de médiateurs et de facteurs de confusion. Ce phénomène est connu sous le nom de facteur de confusion variant dans le temps.

En résumé, ce type de facteur de confusion pose problème, car il biaise l'estimation de l'effet causal d'une exposition variant dans le temps. En effet, il faut contrôler pour les facteurs de confusion dépendants du temps, sinon l'estimateur de l'effet de l'exposition sera biaisé en raison de confusion résiduelle non contrôlée. En revanche, la correction de ces covariables par les méthodes classiques peut également engendrer un biais de sur-ajustement étant donné leur rôle de médiateur.

Dès lors, des nouvelles méthodes longitudinales ont été introduites pour prendre en compte adéquatement les facteurs de confusion variant dans le temps, prédits par l'historique de l'exposition. Notamment, les modèles structurels marginaux (MSM) [Robins et al. (2000)] sont une classe de modèles causaux visant spécifiquement l'estimation de l'effet cumulatif d'une exposition. Différents estimateurs des paramètres d'un MSM existent. Le plus simple, et le plus utilisé, est l'estimateur par la méthode de la pondération par l'inverse de probabilité de traitement (inverse probability of treatment weighting- IPTW) [Robins et al. (2000)] reposant sur les scores de propension. Une autre méthode développée pour produire des estimateurs avec de meilleurs attributs, est le calcul-g [Robins (1987)], qui est une approche de calcul par standardisation généralisée. On trouve aussi l'approche de l'estimation par maximum de vraisemblance ciblée (TMLE) [Van der Laan and Rubin (2006)] qui est une méthode produisant des estimateurs doublement robustes et efficaces pour les paramètres causaux (donnant des estimateurs convergents même si le modèle de traitement ou celui de l'issue est mal spécifié). Le TMLE se combine naturellement aux méthodes d'apprentissage automatique, comme le

SuperLearner, afin d'éviter les biais associés aux erreurs de modélisation.

Les principaux écots de ce mémoire sont de présenter les différentes méthodes causales longitudinales citées ci-dessus, et les appliquer sur une vraie cohorte de travailleurs cols blancs au Québec dont le recrutement a été réalisé en 1991-1993. L'objectif principal étant de comparer ces différentes méthodes dans le contexte de l'étude de l'effet des stresseurs psychosociaux au travail sur la pression artérielle. Afin de réaliser cette comparaison, une extension des simulations de type « Plasmode » aux données longitudinales est proposée. Les simulations Plasmode combinent des données réelles à des données synthétiques et permettent d'évaluer des méthodes dans un contexte plus réaliste que les simulations entièrement basées sur des données synthétiques.

Dans le chapitre 1 , nous introduisons les notions épidémiologiques et causales ainsi que les variables principales à l'étude qui seront utilisées tout au long de ce mémoire. Au deuxième chapitre, nous présentons les différentes approches des modèles structurels marginaux, pour les appliquer après dans le chapitre 3 sur une base de données réelle développée pour évaluer l'effet des stresseurs psychosociaux au travail sur la pression artérielle. Le dernier chapitre sera concentré sur le développement d'une nouvelle méthodologie basée sur le Plasmode pour simuler des données longitudinales. Le but est de valider les comparaisons effectuées entre les différentes approches du MSM dans le cadre d'une étude de simulation. Nous présentons ainsi nos paramètres de simulation et nos résultats obtenus sous une approche paramétrique et une approche non paramétrique.

Le mémoire se termine par une conclusion qui traite de l'interprétation des résultats obtenus dans l'étude de simulation tout en les plaçant dans un contexte élargi, ouvrant ainsi de nouvelles perspectives de recherches.

Chapitre 1

Inférence causale : concepts & définitions

1.1 Association= causalité ?

Une association signifie principalement que lorsqu'on observe la présence d'un facteur de risque donné, un certain résultat est plus susceptible de se produire [Bouyer et al. (2003)]. Elle réfère à toute relation qui peut être définie en fonction d'une distribution conjointe des variables observées.

Une causalité désigne que lorsque l'on modifie la présence d'un facteur de risque donné, un résultat est plus susceptible de se produire parce que le premier facteur l'a provoqué [Cox and Wermuth (2004)]. La « cause » d'un résultat peut être immédiate et directe, comme elle peut être éloignée et indirecte.

Souvent, il existe une amalgamation entre les notions d'association et de causalité. Toutefois, la différence est importante, car une association entre un facteur et un résultat n'entraîne pas forcément un lien de causalité, et vice-versa, l'absence d'une association ne reconduit pas impérativement à une absence de relation cause-effet [Lecoutre (2004)]. La différence entre association et causalité peut être due à différents biais, dont les biais de confusion et de sélection.

1.2 Biais de confusion et de sélection

1.2.1 Biais de confusion

Dans les études épidémiologiques, un biais de confusion est le résultat de l'existence de variables qui sont des déterminants communs entre le facteur d'exposition et le résultat étudié. De telles variables sont nommées variables confondantes.

Par exemple, si l'on s'intéresse à la relation entre l'exposition à des stresseurs psychosociaux au travail et la pression artérielle, le niveau d'éducation pourrait être un facteur confondant. En effet, les personnes avec un niveau d'éducation plus faible sont moins susceptibles d'occuper des emplois où elles disposent d'autonomie dans le choix et l'organisation de leurs tâches, menant à une situation de travail plus stressante. Par ailleurs, un faible niveau d'éducation peut mener à de moins bons résultats

de santé, par exemple par un manque de connaissances ou de ressources pour adopter les meilleures habitudes de vie. Ainsi, le niveau d'éducation est potentiellement une variable confondante dans la relation entre les stresseurs au travail et la pression artérielle.

Le biais de confusion est une erreur systématique de l'estimation d'une mesure d'effet entre l'exposition et l'issue. Il résulte du fait que les groupes exposés et non exposés ne sont pas comparables par rapport à des déterminants de la réponse. Dans l'exemple précédent, la différence entre les sujets exposés aux stresseurs au travail et ceux non exposés par rapport au niveau d'éducation pourrait mener à surestimer la relation néfaste entre l'exposition et la réponse, puisqu'une partie de l'association statistique observée est due à une troisième variable non impliquée dans la relation causale étudiée : l'éducation.

Il est possible de réduire la confusion au niveau des analyses en identifiant et en contrôlant les variables potentiellement confondantes. Dans notre exemple, il serait notamment possible d'éliminer le biais potentiel en comparant les sujets exposés et non exposés séparément pour différents niveaux d'éducation, pour faire disparaître l'association entre exposition et éducation.

1.2.2 Biais de sélection

De son côté, un biais de sélection désigne une erreur systématique provenant de la sélection des sujets qui influence leur participation dans l'étude. Cette sélection biaisée se produit quand il existe un facteur qui influence la participation des sujets dans l'étude et qui est associé à la fois à l'exposition et à l'issue.

Afin d'illustrer ce type de biais, reprenons l'exemple précédent sur les travailleurs. Lorsque l'on étudie l'effet des stresseurs psychosociaux au travail sur la pression artérielle, il est naturel de se limiter à l'étude des personnes actuellement en emploi. Comme nous l'expliquerons, cette sélection peut engendrer un biais dans l'étude de l'effet d'intérêt. Dans cet exemple, il est raisonnable de présumer qu'un âge avancé ou une exposition élevée aux stresseurs psychosociaux au travail réduisent tous les deux les chances de demeurer en emploi. On pourrait diviser les individus en quatre catégories, selon leur âge et leur exposition aux stresseurs psychosociaux au travail : 1) ceux qui ne sont pas exposés et qui n'ont pas un âge avancé, 2) ceux qui sont exposés et qui n'ont pas un âge avancé, 3) ceux qui ne sont pas exposés et qui ont un âge avancé et 4) ceux qui sont exposés et qui ont un âge avancé.

Il est logique de supposer dans cet exemple que ceux de la première catégorie sont ceux qui ont la plus grande probabilité de demeurer en emploi, ceux de la dernière catégorie sont ceux pour qui cette probabilité est la plus faible et ceux des catégories 2 et 3 ont une probabilité entre ces deux extrêmes. Une telle situation engendre une association négative entre l'âge et l'exposition aux stresseurs psychosociaux parmi les individus sélectionnés, même si une telle association n'existait pas dans la population d'origine. Par ailleurs, on peut s'attendre à ce qu'un âge avancé soit associé à une augmentation de la pression artérielle. Ainsi, puisque l'exposition aux stresseurs psychosociaux tend à être associée à un âge plus faible dans les données sélectionnées, l'exposition aux stresseurs psychosociaux pourrait

être associée négativement à la pression artérielle dans les données sélectionnées.

En résumé, le biais de sélection est similaire au biais de confusion, dans le sens où les deux résultent du fait qu'une covariable est associée à la fois à l'exposition et à l'issue. Toutefois, alors que le biais de confusion dans la relation entre l'exposition et l'issue est attribuable au fait que ces deux variables ont une cause commune, ce qui a pour effet d'engendrer une association non causale, le biais de sélection résulte de l'ajustement d'une association exposition-issue sur un tel effet commun. Ce type d'ajustement rend les estimations des associations biaisées.

1.3 Études randomisées et études observationnelles

Une approche adéquate pour éviter les biais de confusion et de sélection, et donc s'assurer que les associations observées correspondent aux liens de causalité d'intérêt, est de choisir le bon plan expérimental.

En épidémiologie, un des objectifs principaux est de déterminer les différentes causes possibles d'une maladie donnée à partir des études épidémiologiques, pour promouvoir des actions de prévention. De toute évidence, la causalité est, au moins en principe, étroitement liée à la méthodologie de l'expérience suivie. Il existe deux grands types d'études : études randomisées et études observationnelles.

Les études randomisées réfèrent aux études dont l'exposition est définie aléatoirement de telle façon que les facteurs de confusion ne risquent pas d'avoir une implication sur les résultats et que les groupes soient les plus comparables possibles.

Bien entendu, nous ne vivons pas dans un univers parfait dans lequel nous pouvons contrôler tous les facteurs externes qui peuvent engendrer un changement des autres facteurs. Donc, nous faisons recours aux essais randomisés contrôlés pour répartir également les variations entre les personnes.

En fait, les études randomisées sont considérées comme le « gold standard », car elles permettent d'éviter le biais de confusion de telle façon que les groupes de traitement seront bien équilibrés pour les facteurs connus et inconnus, assurant ainsi une estimation non biaisée de l'effet du traitement. Les biais de sélection sont également mitigés dans la mesure où la sélection des participants se fait avant l'assignation au traitement, ainsi le traitement ne peut pas être associé à la sélection.

Or, les études randomisées se heurtent à plusieurs difficultés au niveau pratique, car elles peuvent durer très longtemps et être infaisables. Par exemple, il est difficile de réaliser les études randomisées dans le domaine des stresseurs psychosociaux au travail à cause de la difficulté d'attribuer aléatoirement ce type d'exposition, ou, à l'inverse, de réaliser des interventions au plan individuel pour réduire ces stresseurs, en fonction d'une séquence aléatoire. Aussi, les études randomisées ont des limites éthiques liées à leur nature expérimentale qui nécessite de consulter les participants, afin d'identifier et de prendre en compte leurs souhaits. De plus, ces études manquent souvent de validité externe (généralisabilité), puisque les participants sont souvent très différents de ceux qu'on retrouve dans la

population générale et le contexte contrôlé de ces expériences reflète mal le contexte de la vie réelle. Pour surpasser ces limites, on fait recours aux études observationnelles.

Les études observationnelles réfèrent aux études qui analysent les résultats chez des groupes différemment exposés à une intervention donnée, ces groupes ne sont pas forcément comparables. Une limite de ces études, est qu'il est difficile d'interpréter ces relations, car elles sont à risque d'être affectées par les biais de confusion et de sélection.

Il existe toutefois des méthodes statistiques qui permettent de réduire ces biais. Dans la prochaine section, nous allons présenter formellement la notation et les hypothèses d'inférence causale qui permettent d'estimer un effet causal à partir des associations obtenues dans les études observationnelles.

1.4 Notation de l'effet causal

Supposons qu'on observe des observations longitudinales de n individus de la forme suivante :

$$O = (L_0, L_1, A_1, L_2, A_2, \ldots, L_{k-1}, A_{k-1}, Y),$$

telles que :

- Chaque individu participe à k visites.
- La variable L_0 définit l'ensemble des variables potentiellement confondantes qui ne varient pas dans le temps.
- L'ensemble $\{L_t | t = 1, \ldots, k-1\}$ représente le vecteur des covariables variant dans le temps.
- Les variables $\{A_t | t = 1, \ldots, k-1\}$ indiquent l'exposition dichotomique dépendant du temps, par exemple, $A_t = 1$ indique que le patient a été exposé au temps t et $A_t = 0$ indique que le patient n'a pas été exposé au temps t.
- L'issue Y est le résultat final évalué à la fin du protocole.
- Soit $\bar{A}_t = (A_1, \ldots, A_t)$ l'historique de l'exposition d'un sujet du premier temps jusqu'au temps t.
- Soit $\bar{a}_{k-1} = (a_1, \ldots, a_{k-1})$ les différents régimes, c'est-à-dire les valeurs possibles de $\bar{A}_{k-1}$, par exemple, $\bar{a}_{k-1} = (1, \ldots, 1)$ signifie que le sujet est exposé à une exposition à chaque temps.
- L'ensemble $\bar{\mathscr{A}}$ désigne tous les régimes possibles.
- Pour simplifier, on note $\bar{A} = \bar{A}_{k-1}$ et $\bar{L} = \bar{L}_{k-1}$ l'historique sur l'ensemble du suivi.
- Pour définir les modèles structurels marginaux, on introduit la notion d'issue potentielle définie comme étant la valeur que prendrait Y sous le régime d'exposition $\bar{a}$. L'issue potentielle est alors notée $Y^{\bar{a}}$.

L'effet causal est défini mathématiquement comme étant la différence d'espérance mathématique de l'issue entre deux régimes différents $\bar{a}$ et $\bar{a}'$:

$$E[Y^{\bar{a}}] - E[Y^{\bar{a}'}],$$

avec $\bar{a} \neq \bar{a}'$ et $\bar{a}, \bar{a}' \in \bar{\mathscr{A}}$.

1.5 Hypothèses de l'inférence causale

Le paradigme contrefactuel demande d'imaginer que, pour chaque individu, il existe autant d'issues contrefactuelles qu'il existe de différents régimes d'exposition $\bar{a}$. Au total, il existe donc 2^{k-1} issues contrefactuelles.

Le but est de déterminer comment la valeur de $Y^{\bar{a}}$ varie en fonction de $\bar{a}$. Le problème fondamental de l'inférence causale qui se pose est que, pour chaque individu i, on ne peut observer qu'une seule valeur de $Y^{\bar{a}_i}$, soit la valeur qui correspond à l'exposition répétée qui s'est réellement réalisée $\bar{a}_i$.

Les autres issues contrefactuelles peuvent être considérées comme des données manquantes. Autrement dit, nous ne pouvons pas déterminer directement l'effet de l'exposition, car il est impossible de comparer pour un même individu les résultats qui seraient survenus sous différents scénarios d'exposition répétée $\bar{a}$.

Étant donné que l'effet causal pour une certaine unité i ne peut pas être observé, nous cherchons à identifier l'effet causal moyen pour la population dans son ensemble ou pour certaines sous-populations.

Pour que l'approche contrefactuelle ait un sens et soit valide, il est nécessaire de faire quelques hypothèses. Plus précisément les hypothèses de causalité sont l'interchangeabilité conditionnelle, la positivité et la stabilité de la valeur du traitement de l'unité.

1.5.1 Interchangeabilité conditionnelle

L'hypothèse de l'interchangeabilité conditionnelle est telle que :

$$(Y^{\bar{a}} \perp\!\!\!\perp A_t | \bar{A}_{t-1} = \bar{a}_{t-1},\ \bar{L}_t = \bar{l}_t)\ \ \forall t \in 1,..,k-1.$$

Cette hypothèse réfère à une indépendance entre l'issue potentielle et l'exposition à t sachant l'historique de l'exposition et des covariables. Reprenons l'exemple des travailleurs, cette hypothèse signifie informellement qu'il n'existe pas de facteurs pronostics de la pression artérielle qui seraient aussi associés à l'exposition au temps t, à part l'exposition aux temps antérieurs et les facteurs confondants $\bar{L}_t$.

Cette condition tient dans les études observationnelles quand la probabilité de recevoir un traitement ne dépend pas des covariables non mesurées, et donc, l'exposition peut être considérée comme étant attribuée au hasard conditionnellement à l'historique observé.

1.5.2 Positivité

L'hypothèse de positivité est comme suit :

$$P(\bar{L}_t = \bar{l}_t, \bar{A}_{t-1} = \bar{a}_{t-1}) > 0 \ \ \& \ \ P(\bar{A}_t = \bar{a}_t | \bar{L}_t) > 0 \ \ \forall \bar{a}_t, \ \ t \in 1, .., k-1.$$

C'est-à-dire à tout temps t, chaque individu a des probabilités non nulles d'avoir chacun des niveaux d'exposition.

En pratique, cette hypothèse [Swanson et al. (2018)] peut être violée au cas où un traitement est déconseillé pour certains patients et donc ces patients auraient des probabilités nulles de subir le traitement. Des problèmes pratiques peuvent également survenir dans le cas où les données sont rares et les probabilités estimées de recevoir une exposition approchent zéro.

1.5.3 Hypothèse de stabilité de la valeur du traitement de l'unité (Stable Unit Treatment Value Assumption : SUTVA)

Cette hypothèse permet de définir le résultat contrefactuel [Rubin (1978)]. Elle stipule que l'issue mesurée pour un sujet ne doit pas être influencée par les expositions des autres individus. En d'autres termes, cela se traduit par l'indépendance entre l'issue potentielle pour un individu et les niveaux d'exposition des autres individus dans la population.

Pour une exposition observée $A_i = a \Rightarrow Y_i^a = Y_i^{obs}$ tel que Y_i^{obs} est l'issue du participant i .

Un exemple possible peut être de la vaccination contre une grippe. On prend par exemple un échantillon de 100 individus, dont 90 personnes sont vaccinées contre cette maladie. Si une des 10 personnes restantes est vaccinée ou pas, ça ne va rien changer, car la maladie a peu de chances de se propager dans ce cas. Au contraire, si les 90 personnes n'avaient pas été vaccinées, il aurait pu exister un effet pour les 10 personnes restantes à être vaccinées. Ainsi, la valeur du traitement pour un individu donné dépend du traitement des autres personnes dans cet exemple, contrevenant à l'hypothèse SUTVA.

Aussi, pour le cas de travailleurs énoncé dans la section précédente, il faut donc supposer que l'exposition d'un individu n'influence pas la pression artérielle d'un autre travailleur. Donc, la pression artérielle mesurée pour le participant i doit être la même, quelque soit les expositions des autres sujets.

SUTVA suppose également qu'une seule version du traitement est appliquée à toutes les unités c'est-à-dire que les résultats potentiels, pour chaque individu et sous chaque exposition possible sont bien définis et prennent une valeur unique.

On suppose donc que l'issue réellement observée est précise et cohérente à l'issue potentielle au niveau d'exposition observé. Au cas contraire, s'il existe plusieurs versions d'une même exposition et si ces différentes versions donnent lieu à des issues potentielles différentes, l'hypothèse est violée. Par exemple si on est intéressé par l'effet d'un médicament sur les sujets, et que ce médicament peut être pris soit par voie orale ou voie auriculaire, et que ces deux versions ont des effets différents sur

un même individu, alors cette hypothèse sera violée et l'effet estimé sera ambigu, ne correspondant pas à l'effet de ni l'un ni l'autre des deux versions.

Chapitre 2

Les modèles structurels marginaux

La recherche en épidémiologie est largement basée sur des études observationnelles. Dans ce type d'étude, il est difficile d'inférer des relations causales à partir des associations observées. Toutefois, plusieurs méthodes d'analyse causale ont été développées pour relever les défis propres aux études observationnelles.

En particulier, dans les études longitudinales, les modèles structurels marginaux [Robins et al. (2000)] permettent notamment de remédier au biais qui peut survenir lorsque des facteurs de confusions dépendant du temps sont affectés par l'exposition précédente. Par ailleurs, des méthodes qui combinent l'apprentissage automatique et l'inférence causale ont également été développées pour limiter les hypothèses statistiques nécessaires aux inférences.

Afin d'illustrer le problème des variables confondantes variant dans le temps, considérons une situation où l'on s'intéresse à estimer l'effet d'une exposition dépendante du temps $\{A_1, A_2\}$ sur Y en présence de variables confondantes $\{L_1, L_2\}$. Cette situation pourrait être représentée à la figure 2.1.

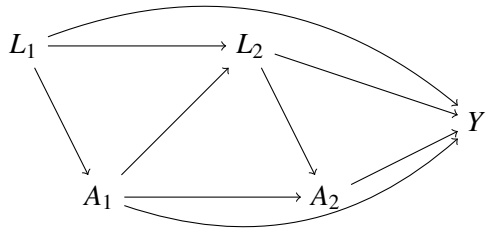


FIGURE 2.1 – Figure représentant une présence de confusion dépendante du temps

Dans cette figure, une partie de l'effet de A_1 sur Y passe par L_2 en raison du chemin $A_1 \rightarrow L_2 \rightarrow Y$. Ainsi, l'ajustement dans un modèle traditionnel pour L_2 crée un biais dans l'estimation de l'effet causal, car une partie de l'effet induit n'est pas pris en compte. Cependant, L_2 est aussi une variable confondante de l'effet de A_2 sur Y, puisque L_2 est une cause commune de ces deux variables ($A_2 \leftarrow L_2 \rightarrow Y$). Ne pas ajuster pour L_2 induit donc une association non-causale entre A_2 et Y. Les MSM

(modèles structurels marginaux) permettent de résoudre cette impasse.

Les MSM décrivent l'effet de l'exposition sur la distribution marginale des issues contrefactuelles $E(Y^{\bar{a}})$. Ils peuvent être définis de la façon générale suivante :

$$g(E(Y^{\bar{a}})) = f(a_1, \ldots, a_{k-1}),$$

telle que $g(.)$ est la fonction de lien entre $E(Y^{\bar{a}})$ et l'historique de l'exposition.

Un exemple de MSM spécifique est :

$$logit(P(Y^{\bar{a}} = 1)) = \gamma_0 + \gamma_1 \times cumul(\bar{a}),$$

où $cumul(\bar{a}) = \sum_{t=1}^{k-1} a_t$. Un tel MSM indique que le logit de probabilité de l'événement Y est une fonction linéaire de l'exposition cumulative. Le paramètre γ_1 obtenu est interprété comme étant l'effet causal de l'exposition cumulée.

Sous les hypothèses causales énoncées à la section précédente, différents estimateurs des paramètres des MSM ont été proposés, dont l'IPTW (inverse de probabilité de traitement), le calcul-g et le TMLE (estimation par maximum de vraisemblance ciblée).

2.1 MSM via IPTW

La méthode de pondération par l'inverse de probabilité de traitement est inspirée par la méthode de Horvitz (1952). Elle peut être utilisée pour estimer les effets de l'exposition cumulée. Elle se subdivise en deux grandes étapes. La première consiste au calcul des poids de traitement en fonction de l'inverse de la probabilité du traitement réellement éprouvé par chaque patient à chaque temps d'exposition, en considérant son historique de covariables. Ces poids peuvent ensuite être stabilisés à partir de l'historique du traitement en fonction des informations de base de chaque individu pour réduire la variabilité des poids.

Intuitivement, cette méthode de pondération permet de régler le problème des variables jouant le double rôle de confondants et d'intermédiaire, puisque chaque exposition n'est modélisée qu'en fonction de son historique propre.

Sous les quatre hypothèses citées dans le chapitre précédent, cette pondération [Robins et al. (2000)] crée une « pseudo-population » dans laquelle toute la confusion est éliminée et le traitement est indépendant des facteurs de confusion mesurés. Cette pseudo-population est le résultat d'affectation à chaque participant d'un poids proportionnel à sa probabilité de recevoir son propre historique d'exposition.

En pratique, Robins et al. (2000) montre que les poids peuvent être de la forme $g(A|\bar{A})/P(A|\bar{A},\bar{L})$, où $g(.)$ est une fonction quelconque. C'est-à-dire qu'ils partagent le même dénominateur et que seul le numérateur varie selon le type de poids.

Pour effectuer la pondération, on rencontre souvent dans la littérature trois types de poids : les poids standards, les poids stabilisés et les poids stabilisés marginaux, respectivement notés

$$
\begin{aligned}
w_i &= \prod_{t=1}^{k-1} \frac{1}{P(A_t = a_{t,i}|\bar{A}_{t-1} = \bar{a}_{t-1,i}, \bar{L}_t = \bar{l_{t,i}})} \\
sw_i &= \prod_{t=1}^{k-1} \frac{P(A_t = a_{t,i}|\bar{A}_{t-1} = \bar{a}_{t-1,i})}{P(A_t = a_{t,i}|\bar{A}_{t-1} = \bar{a}_{t-1,i}, \bar{L}_t = \bar{l_{t,i}})} \\
swm_i &= \prod_{t=1}^{k-1} \frac{P(A_t = a_{t,i})}{P(A_t = a_{t,i}|\bar{A}_{t-1} = \bar{a}_{t-1,i}, \bar{L}_t = \bar{l_{t,i}})}.
\end{aligned}
$$

Les probabilités de l'exposition en fonction des variables confondantes peuvent varier considérablement entre les sujets lorsqu'il existe une forte association entre ces deux ensembles de variables. Cette variabilité peut induire des valeurs extrêmement élevées des poids standards pour des sujets donnés, et l'estimateur IPTW aura une grande variance. Cependant, cette variabilité peut être atténuée en remplaçant les poids standards par les poids stabilisés. Par exemple, si l'exposition et les facteurs de confusion ne sont pas associés, $P(A_t = a_{t,i}|\bar{A}_{t-1}) = P(A_t = a_{t,i}|\bar{A}_{t-1}, \bar{L}_t)$, donc $sw_i = 1$, alors que les autres poids seront variables.

En ce qui concerne la deuxième étape, un modèle de régression associant le résultat et l'historique de l'exposition est construit à l'aide de l'échantillon pondéré, n'incluant pas les facteurs de confusion mesurés en tant que covariables. Ce modèle définissant la relation entre l'issue et l'historique d'exposition doit être bien spécifié. Les paramètres des modèles de régression obtenus, qui correspondent à ceux des MSM, peuvent être utilisés pour estimer l'effet causal moyen dans la population de l'étude.

Nous élaborons ci-dessous une ébauche de preuve que l'estimateur IPTW permet d'estimer sans biais les moyennes contrefactuelles sous les hypothèses données au chapitre précédent.

Supposons qu'on a un régime de 2 périodes, et qu'on considère le cas particulier où : $A_1 = 1$ et $A_2 = 1$.

On note :

$$f(A|L) = P(A_1|L_1, L_2)P(A_2|A_1, L_1, L_2)\cdot$$

Montrons que

$$E\left[\frac{I(A_1 = 1, A_2 = 1)}{f(A|L)}Y\right] = E(Y^{(1,1)})\cdot$$

Démonstration. On a que

$$E_Y\left[\frac{I(A_1=1,A_2=1)}{f(A|L)}Y\right] = E_Y\left[\frac{I(A_1=1,A_2=1)}{f(A|L)}Y^{(1,1)}\right] \tag{2.1}$$

$$= E_{Y|L_1}\left\{E_{L_1}\left[\frac{I(A_1=1,A_2=1)}{f(A|L)}Y^{(1,1)}|L_1\right]\right\} \tag{2.2}$$

$$= E_{Y|L_1}\left\{E_{L_1}\left[\frac{I(A_2=1)}{f(A|L)}Y^{(1,1)}|A_1=1,L_1\right]E\left(A_1|L_1\right)\right\} \tag{2.3}$$

$$\begin{aligned}= E_{Y|L_1,L_2}\Bigg(E_{L_1,L_2}\left\{E_{L_1}\left[\frac{I(A_2=1)}{f(A|L)}Y^{(1,1)}|A_1=1,L_1\right]|L_2\right\}\\ E\left(A_1=1|L_1\right)\Bigg)\end{aligned} \tag{2.4}$$

$$\begin{aligned}= E_{Y|L_1,L_2}\Bigg(E_{L_1,L_2}\left\{E_{L_1}\left[\frac{1}{f(a|L)}Y^{(1,1)}|A_2=1,A_1=1,L_1\right]|L_2\right\}\\ E\left(A_2=1|A_1=1,L_1,L_2\right)E\left(A_1=1|L_1\right)\Bigg)\end{aligned} \tag{2.5}$$

$$\begin{aligned}= E_{Y|L_1,L_2}\Bigg(\frac{1}{f(a|L)}E_{L_1,L_2}\left\{E_{L_1}\left[Y^{(1,1)}|A_2=1,A_1=1,L_1\right]|L_2\right\}\\ E\left(A_2=1|A_1=1,L_1,L_2\right)E\left(A_1=1|L_1\right)\Bigg)\end{aligned}$$

$$\begin{aligned}= E_{Y|L_1,L_2}\Bigg(\frac{1}{f(a|L)}P(A_2=1|A_1,L_1,L_2)P(A_1=1|L_1)\\ E_{L_1,L_2}\left\{E_{L_1}\left[Y^{(1,1)}|A_2=1,A_1=1,L_1\right]|L_2\right\}\Bigg)\end{aligned}$$

$$= E_{Y|L_1,L_2}\left(\frac{f(a|L)}{f(a|L)}E_{L_1,L_2}\left\{E_{L_1}\left[Y^{(1,1)}|A_2=1,A_1=1,L_1\right]|L_2\right\}\right)$$

$$= E_{Y|L_1,L_2}\left(E_{L_1,L_2}\left\{E_{L_1}\left[Y^{(1,1)}|A_2=1,A_1=1,L_1\right]|L_2\right\}\right)$$

$$= E\left(Y^{(1,1)}\right),$$

□

où :

(2.1) est obtenu selon l'hypothèse de cohérence ;

(2.2) est obtenu selon le théorème de l'espérance totale ;

(2.3) suit de $Y^{(a1,a2)} \perp\!\!\!\perp A1|L1$;

(2.4) découle du thèorème de l'espérance totale ;

(2.5) suit de $Y^{(a1,a2)} \perp\!\!\!\perp A2|A1,L2,L1$.

IPTW est une méthode intuitive pour l'estimation de l'exposition variant dans le temps. Par contre, elle peut produire des estimations instables et biaisées lorsque les probabilités de traitement approchent

la valeur 0 pour certains individus, même avec les poids stabilisés. Une approche alternative utilisée dans les articles méthodologiques est le calcul-g.

2.2 MSM via calcul-g

Le calcul-g a été introduit par Robins (1987) et constitue un estimateur alternatif des paramètres du MSM. Le "g" signifie "généralisé" car, selon les hypothèses décrites dans la section 1.5, le calcul-g nous permet d'estimer l'effet d'une exposition donnée sans introduire un biais dû à un ajustement inapproprié pour la confusion dépendante du temps.

Cette approche nécessite de modéliser l'espérance de la réponse en plus de la densité des variables confondantes pour estimer le lien de causalité. La formule utilisée pour l'estimation du paramètre d'intérêt s'écrit sous la façon suivante :

$$E(Y^{\bar{a}}) = \int_{(l_0,\ldots,l_{k-1})} E(Y|\bar{A}=\bar{a},\bar{L}=\bar{l}) \prod_{t=0}^{k-1} f(L_t = l_t|\bar{A}_{t-1}=\bar{a}_{t-1},\bar{L}_{t-1}=\bar{l}_{t-1})\, dF_{l_{k-1}} \ldots dF_{l_0}.$$

L'avantage de l'intégrale, c'est qu'avec un choix de mesure approprié, la notation d'intégrale peut aussi représenter une somme (dans le cas discret), mais le contraire n'est pas évident.

Cette formule a été proposée par Robins (1987). Elle suggère d'estimer les moyennes contrefactuelles comme une somme pondérée des espérances de la réponse conditionnellement à l'historique d'exposition et des covariables, où les poids sont une fonction de la densité des variables confondantes. L'espérance de la réponse pourrait être estimée à partir d'un modèle de régression simple. Ce dernier modèle est fréquemment appelé « modèle Q ». Pour bien estimer l'effet de l'exposition, le modèle Q doit être correctement spécifié. Une possibilité est alors d'utiliser conjointement le calcul-g avec des algorithmes d'apprentissage automatique.

Le modèle Q obtenu est employé pour prédire les issues contrefactuelles de chaque observation sous chaque régime d'exposition. Cette approche permet ainsi de résoudre le problème des données contrefactuelles manquantes décrit à la section 1.5.

Théoriquement, il y a une condition que les méthodes d'apprentissage automatiques pour estimer Q convergent à une vitesse paramétrique ($\sqrt{n}$) afin que le calcul-g possède ses propriétés standards.

Démonstration. Nous présentons ci-dessous une preuve que l'effet causal est identifiable par le calcul-g dans le cas d'une exposition avec deux temps de mesure.

Sous les hypothèses causales données au chapitre 1, on a :

$$
\begin{aligned}
E(Y^{A_1=a_1,A_2=a_2}) &= \int_{L_1} E[Y^{A_1=a_1,A_2=a_2}|L_1]f(L_1)dF_{L_1} && (1)\\
&= \int_{L_1} E[Y^{A_1=a_1,A_2=a_2}|A_1=a_1,L_1]f(L_1)dF_{L_1} && (2)\\
&= \int_{L_2}\int_{L_1} E[Y^{A_1=a_1,A_2=a_2}|L_2,A_1=a_1,L_1]f(L_2|L_1,A_1=a_1)f(L_1)\,dF_{L_2}\,dF_{L_1} && (3)\\
&= \int_{L_2}\int_{L_1} E[Y^{A_1=a_1,A_2=a_2}|A_2=a_2,L_2,A_1=a_1,L_1]f(L_2|L_1,A_1)f(L_1)\,dF_{L_2}\,dF_{L_1} && (4)\\
&= \int_{L_2}\int_{L_1} E[Y|A_2=a_2,L_2,A_1=a_1,L_1]f(L_2|L_1,A_1)f(L_1)\,dF_{L_2}\,dF_{L_1}\ , && (5)
\end{aligned}
$$

□

où :

dF_{L_1} et dF_{L_2} sont des mesures de probabilités appropriées (par exemple, une mesure de dénombrement pour des covariables discrètes).

(1) est obtenu selon la loi des probabilités totales ;

(2) suit de $Y^{A_1,A_2} \perp\!\!\!\perp A_1|L_1$:

(3) est obtenu selon la loi des probabilités totales ;

(4) suit de $Y^{A_1,A_2} \perp\!\!\!\perp A_2|L_2,A_1,L_1$;

(5) découle de $A_1=a_1,A_2=a_2 \Rightarrow Y^{A_1=a_1,A_2=a_2}=Y$

Le calcul-g peut aussi s'écrire sous la forme d'espérances itérées, ce qui suggère l'algorithme ci-dessous. Plus précisément, l'algorithme de la méthode de G-computation se construit comme suit :

Algorithme 2.2.1 : Algorithme du calcul-g

Données : $O = (L_0, L_1, A_1, L_2, A_2, \ldots, L_{k-1}, A_{k-1}, Y)$

Résultat : $\bar{Q}_0$

1. Construire un modèle de régression de l'issue Y en fonction de l'ensemble de l'exposition $\bar{A}$ et de l'historique des covariables $\bar{L}$.
2. **pour** $\bar{a} \in \mathscr{A}$ **faire**
   a) Utiliser le modèle obtenu en (1) pour prédire l'issue Y sous le régime $\bar{A} = \bar{a}$. Ce qui entraîne l'estimation contrefactuelle $\hat{Y}^{\bar{a}_{k-1}} = \bar{Q}_{k-1}$.
   b) **pour** $t = k-2, \ldots, 1$ **faire**
      i. Ajuster un modèle pour $\bar{Q}_{t+1}$ en fonction des covariables $\bar{L}_t$ et de l'historique de l'exposition $\bar{A}_t$.
      ii. Pour tous les sujets, prédire une nouvelle issue à partir du modèle obtenu sous le régime d'exposition $\bar{A}_t = \bar{a}_t$. Ce qui produit l'estimateur $\bar{Q}_t$.

      fin
   c) Le dernier estimateur obtenu sera $\bar{Q}_1$ à partir de la régression en fonction des covariables du *baseline* $\bar{L}_1$ et de $\bar{A}_1$ pour chaque exposition cumulée.
   d) Prendre la moyenne du résultat contrefactuel $\bar{Q}_1$ sur l'ensemble des observations. Cette moyenne sera notée $\bar{Q}_0$.

   fin

En appliquant cet algorithme, on va obtenir une base de données qui contient autant de moyennes contrefactuelles $\bar{Q}_0$ que de régimes possibles. Si on désire résumer l'information, on peut effectuer une régression des moyennes obtenues en fonction de leur régime d'exposition correspondant selon un modèle paramétrique.

Ensuite, nous pouvons appliquer le bootsrap pour calculer la variance et les intervalles de confiance de niveau 95% pour l'estimation de l'effet causal moyen.

2.3 La courbe d'influence

La courbe d'influence est une fonction de score pondérée qui contient toutes les informations sur la variance de l'estimateur associé. Une courbe d'influence efficace pour un paramètre donné est la courbe d'influence qui atteint la plus petite variance. Une possibilité d'atteindre une inférence semi-paramétrique [Petersen et al. (2012)] consiste à estimer les composantes de la courbe d'influence efficace (CIE), puis à les utiliser comme une équation d'estimation en la fixant à 0, et à résoudre cette équation pour le paramètre d'intérêt.

On note $\bar{g}_t$ la probabilité associée à l'obtention d'un certain historique de traitement $\bar{a}_{t-1}$ sachant $\bar{L}_{t-1}$.

Elle peut s'écrire :

$$\bar{g}_t(\bar{L}_{t-1}) = \prod_{s=0}^{t-1} P(A_s = \bar{a}_s | \bar{A}_{s-1} = \bar{a}_{s-1}, \bar{L}_s = \bar{l}_s), \forall t = 2, \ldots, k-1.$$

Donc, la courbe d'influence efficace $D^*(O)$ pour un régime $\bar{a}$ peut s'écrire récursivement [Petersen et al. (2012)] comme étant la somme des composantes suivantes :

$$\begin{aligned} D_t &= \frac{I(\bar{A}_{t-1} = \bar{a}_{t-1})}{\bar{g}_t}(\bar{Q}_{t+1} - \bar{Q}_t), \text{ pour t=k-1,...,2} \\ D_1 &= \frac{1}{\bar{g}_1}(\bar{Q}_2 - \bar{Q}_1) \\ D_0 &= (\bar{Q}_1 - \bar{Q}_0), \end{aligned} \tag{2.6}$$

avec $\bar{Q}_k = Y \cdot$

L'information de Fisher I est définie comme étant la taille d'échantillon n divisée par la variance de la courbe d'influence efficace [Gruber and Van der Laan (2010)], $I = \frac{n}{Var(D^*(O))}$. Étant donné que la variance de la courbe d'influence efficace divisée par n fois la variance d'un estimateur asymptotiquement efficace converge vers 1 lorsque la taille de l'échantillon converge vers l'infini, nous pouvons donc également considérer la variance d'un estimateur efficace du paramètre d'intérêt comme étant l'inverse de l'information I . Par conséquent, pour tout estimateur linéaire asymptotiquement et régulier (LAR) de paramètre d'intérêt, sa courbe d'influence a une variance qui est supérieure ou égale à la variance de la courbe d'influence efficace.

2.4 MSM via TMLE

Depuis les années 2000, la méthode de TMLE a été utilisée pour plusieurs applications, préalablement pour les effets d'expositions ponctuelles. Récemment, de plus en plus de travaux sont mis en œuvre pour l'application du TMLE sur les données longitudinales pour s'ajuster à la confusion variante dans le temps.

TMLE longitudinale est une méthode semi-paramétrique doublement robuste basée sur le maximum de vraisemblance. Elle demande de modéliser le traitement et l'issue. La méthode est dite doublement robuste, car l'effet causal est estimé d'une façon cohérente soit si le modèle de régression de l'issue ou celui du traitement est bien spécifié. Si les deux modèles sont estimés d'une façon cohérente, l'estimateur obtenu est efficace, c'est-à-dire que sa variance atteint la borne inférieure de la CIE.

En effet, l'estimateur TMLE converge à vitesse $\sqrt{n}$, même si le modèle pour l'issue et celui pour le traitement convergent à vitesse $n^{1/4}$ [Kennedy (2016)]. L'approche du TMLE peut produire une inférence rapide et efficace et réduire le biais à l'aide des algorithmes de l'apprentissage machine. Il est également facile de calculer les erreurs standards et des intervalles de confiance, en se basant sur la CIE.

L'algorithme de TMLE que nous décrivons ci-dessous est très similaire à celui du calcul-g. En fait, cet algorithme utilise l'estimation du calcul-g comme une estimation initiale. Cette estimation initiale est ensuite fluctuée en fonction de l'association résiduelle entre l'issue prédite et la probabilité de traitement conditionnelle aux variables confondantes. L'intuition est que si l'estimation initiale était adéquate (sans biais), la probabilité de traitement ne devrait ajouter aucune nouvelle information sur l'issue. Dans le cas contraire, la probabilité de traitement peut être utilisée pour estimer le biais résiduel. Le TMLE effectue cette fluctuation de l'estimation initiale d'une façon ingénieuse qui garantit que les équations d'estimation de la CIE non paramétrique soient solutionnées, de sorte que l'estimateur résultant dispose des propriétés de double-robustesse et d'efficacité énoncées ci-dessus.

L'algorithme du *TMLE* se décrit comme suit :

Algorithme 2.4.1 : algorithme de la méthode TMLE

Données : $O = (L_0, L_1, A_1, L_2, A_2, \ldots, L_{k-1}, A_{k-1}, Y)$

Résultat : $\bar{Q}_0^1$

pour *t=k-1,...,1* **faire**

* Mettre à jour/ Fluctuer l'estimation $\bar{Q}_t$ obtenue après chaque étape des régressions séquentielles utilisées dans la méthode du calcul-g en utilisant un sous-modèle paramétrique :

$$\bar{Q}_t^1(\varepsilon_t) = \bar{Q}_t + \varepsilon_t H_t(\bar{A}, \bar{L})_{t-1}$$

  1. Cette étape consiste à calculer de la valeur des covariables ingénieuses (*clever covariates*). Ces covariables ingénieuses sont calculées en se basant sur le produit cumulatif de l'inverse des probabilités du traitement. C'est-à-dire le produit d'une indicatrice désignant si un sujet a subi le régime des traitements d'intérêt divisée par la probabilité prévue de l'avoir fait :

$$\hat{H}_t(\bar{A}, \bar{L})_{t-1} = \frac{I(\bar{A}_{t-1} = \bar{a}_{t-1})}{\prod_{s=0}^{t-1} P(A_s = \bar{a}_s | \bar{A}_{s-1} = \bar{a}_{s-1}, \bar{L}_s = \bar{l}_s)}$$

  2. Le paramètre ε_t est estimé à partir de la régression de l'issue Y sur $H_t(\bar{A}, \bar{L})_{t-1}$, en précisant que $\bar{Q}_t$ est le *offset*.

  Il convient de souligner que la forme fonctionnelle des covariables ingénieuses est similaire à celle de la pondération par l'inverse de probabilité de traitement.

* Prendre la moyenne de l'estimation $\bar{Q}_1^1$ sur l'ensemble des observations. Cette moyenne sera notée $\bar{Q}_0^1$.

fin

La fonction de fluctuation montrée dans l'algorithme ci-dessus doit satisfaire deux conditions :

1. la fonction de fluctuation doit se réduire à la densité d'origine lorsque $\varepsilon_t = 0$;
2. la dérivée par rapport à ε_t de la fonction de perte doit s'étendre linéairement sur la courbe

d'influence efficace à $\varepsilon_t = 0$.

L'estimation de ε_t est trouvée en minimisant la moyenne empirique de la fonction de perte suivante :

$$\mathcal{L}\left[\bar{Q}_t^1(\varepsilon_t)\right] = \sum_{i=1}^{t}\left[\bar{Q}_i - \bar{Q}_i^1(\varepsilon_i)\right]^2.$$

L'inférence statistique pour TMLE peut être obtenue en calculant les erreurs-types via les estimateurs de la courbe d'influence ou par le bootstrap. Notamment, les intervalles de confiance peuvent être estimés à partir de l'estimation de la courbe d'influence de $\bar{Q}_0$, tel que :

$$\widehat{IC}(\bar{Q}_0) = \left\{\sum_{t=2}^{k-1}\hat{H}(\bar{A},\bar{L})_{t-1}\left[\bar{Q}_t^1 - \bar{Q}_{t-1}^1\right]\right\} - \bar{Q}_0^1.$$

Un intervalle de confiance asymptotiquement normal au niveau de 95% peut alors s'écrire de la façon suivante :

$$\left[\bar{Q}_0^1 \pm 1.96\sqrt{\frac{\widehat{Var}(\widehat{IC})}{n}}\right].$$

Si on utilise des méthodes paramétriques (régression simple) pour l'estimation du traitement et pour les modèles intératifs conditionnels, on s'attend à ce que ces deux composantes soient mal spécifiées.

Donc, afin de réduire les biais, d'atteindre l'efficacité de l'estimateur obtenu par l'algorithme et d'assurer une inférence statistique précise, TMLE est souvent associé aux approches de l'apprentissage automatique, en particulier la méthode de *SuperLearner* (SL).

2.5 MSM via TMLE-SL

SuperLearner [Breiman (1996)] est une méthode d'apprentissage machine basée sur le principe de la validation croisée pour trouver la combinaison idéale des différents algorithmes de prédiction selon une fonction de perte donnée. Les méthodes de prédiction appartenant aux différentes librairies sont choisies au préalable par l'utilisateur. Elles peuvent regrouper des méthodes paramétriques et non paramétriques (par exemple, la régression linéaire, les forêts aléatoires, les splines...) [Van der Laan et al. (2007)].

Cependant, les coefficients de chaque candidat n'ont pas d'interprétation significative. En effet, imaginons une situation où l'utilisateur choisit deux méthodes candidates qui fournissent des prédictions hautement corrélées. Dans ce cas les coefficients associés aux candidats seront très variables, mais l'estimateur de *SuperLearner* ne sera pas affecté par une telle variabilité.

Plus précisément, chaque méthode produit une prédiction de la validation croisée, et puis, le poids optimal est déterminé en minimisant l'erreur de cette prédiction, formulée comme étant une régression

de l'issue Y en fonction des prédictions obtenues. La combinaison optimale des poids de nos apprenants candidats est appliquée aux données originales complètes pour produire un nouvel ensemble des valeurs prédites qu'on appelle l'estimateur de *SuperLearner*.

Asymptotiquement, *SuperLearner* fonctionne aussi bien que le modèle le plus performant spécifié par l'utilisateur en termes de minimisation de la fonction de perte de la validation croisée [Van der Laan et al. (2007)]. Cette méthodologie est utilisée pour la prédiction à partir des méthodes présélectionnées de chaque librairie. Elle peut aussi servir à obtenir les paramètres de l'effet causal. Ainsi, il a été démontré que *SuperLearner* réduit le biais d'une mauvaise spécification du modèle d'intérêt. Pour mettre en pratique cette approche, on peut utiliser la fonction *SuperLearner* sous le package R **SuperLearner** [Polley et al. (2011)].

Chapitre 3

Application

En épidémiologie occupationnelle, des études prospectives suggèrent un effet délétère des stresseurs psychosociaux au travail sur la santé cardiovasculaire. Ces études utilisent parfois une mesure cumulée de l'exposition aux stresseurs psychosociaux au travail, pour prévenir la sous-estimation potentielle liée à l'utilisation d'une mesure unique d'exposition. Toutefois, cet effet cumulé n'a presque jamais été estimé à partir des méthodes causales présentées dans le précédent chapitre. Des analyses classiques de l'issue en fonction de l'exposition ajustée soit pour les covariables initiales, soit pour les covariables initiales et celles variant dans le temps, sont majoritairement utilisées.

À notre connaissance, une seule étude antérieure dans ce domaine de recherche a eu recours aux méthodes causales [Sall et al. (2019)]. Dans cette étude, les paramètres de MSM sont estimés par la méthode de la pondération par l'inverse de probabilité de traitement. Les estimations obtenues par IPTW ne sont valides que si les hypothèses sont respectées. Parmi ces hypothèses figure la bonne spécification du modèle causal. C'est ainsi qu'un test est développé dans cet article afin de vérifier, à partir des données, l'hypothèse selon laquelle le modèle considéré est bien spécifié. Ce test se base sur l'égalité des estimateurs des effets causaux obtenus par une pondération avec des poids ordinaires et par une pondération avec des poids stabilisés. Dans tous les cas étudiés, ce test ne rejette jamais l'hypothèse nulle que le modèle structurel est correctement spécifié au delà du niveau de 5% lorsque celle-ci est vraie.

Comme nous l'avons déjà présenté, les analyses classiques sont théoriquement inappropriées s'il existe des variables qui ont le double rôle de confondant et d'intermédiaire de la relation entre l'exposition aux stresseurs psychosociaux au travail et les issues de santé cardiovasculaire. Tel que suggéré dans l'article de Sall et al. (2019), cette situation est probable lorsqu'on examine l'effet de ces stresseurs sur la pression artérielle des travailleurs.

Toutefois, la présence et l'ampleur du biais attribuable au phénomène de confusion dépendante du temps n'ont jamais été empiriquement évaluées dans ce domaine de recherche. Considérant la complexité des approches causales présentées dans le précédent chapitre, il est important d'évaluer la pertinence réelle de les utiliser dans ce contexte particulier, et, de façon plus générale, dans le domaine

de l'épidémiologie occupationnelle.

Dans ce chapitre, les données d'une étude de cohorte longitudinale de 5 ans sur les cols blancs de la ville de Québec (Canada) sont utilisées pour illustrer et évaluer les méthodes proposées dans le chapitre 2, afin d'estimer l'effet causal de l'exposition cumulée aux stresseurs psychosociaux au travail sur la pression artérielle.

3.1 Les Maladies Cardiovasculaires

Les maladies cardiovasculaires (MCV) regroupent un certain nombre de troubles qui affectent les vaisseaux sanguins et le cœur, tel que l'hypertension artérielle (élévation de la tension) [Claes and Jacobs (2007)]. Ils provoquent la plupart des maladies et des décès dans les pays idustrialisés [World Health Organization (2011)].

Une élévation de la pression artérielle peut être attribuable à des facteurs de risque tel l'âge [Lawes et al. (2005)], l'obésité [Whelton et al. (2002)], un régime riche en sodium [Campbell et al. (1999)], la consommation d'alcool [Appel (2003)] et la sédentarité [Campbell et al. (1999)]. Des études épidémiologiques ont également montré que des facteurs psychosociaux, incluant les stresseurs psychosociaux au travail, peuvent contribuer à l'élévation de la pression artérielle [Markovitz et al. (2004)].

3.2 Le profil de la cohorte

L'étude PROspective de Québec (PROQ) sur le travail et la santé a été menée auprès de 9.000 travailleurs pour la compréhension de la relation cause-effet entre les facteurs de stress au travail et les MCV [Trudel et al. (2016a)]. Une cohorte prospective imbriquée dans cette cohorte plus vaste a été initiée en 2000-2004, avec deux points de suivi 3 ans et 5 ans plus tard (2004-2006 et 2006-2009). Cette cohorte imbriquée est constituée de l'ensemble des travailleurs cols blancs de trois entreprises publiques du secteur de l'assurance à Québec, Canada (taux de participation : 85%).

Cette cohorte a été mise sur pied pour examiner la relation entre les stresseurs psychosociaux au travail et la pression artérielle ambulatoire (PA), prise durant la journée de travail [Brisson et al. (1994), Brisson et al. (2001)]. Cette dernière est composée de cadres supérieurs, de professionnels, de techniciens et d'employés de bureau. Les travailleurs ont rempli un questionnaire auto-déclaré sur les caractéristiques du travail et les facteurs de risque de la PA.

3.2.1 La variable réponse

Au dernier temps de mesure, la PA a été calculée par un personnel formé en utilisant l'appareil *Spacelabs* (mesures ambulatoires) pour améliorer la précision et la validité des mesures prises. Les mesures ont été prises durant une journée normale de travail, à chaque 15 minutes. Les moyennes de PA systolique et diastolique ont été estimées.

Les mesures ambulatoires de la PA sont connues pour éviter l'erreur d'observation (ce que l'on appelle « *white-coat effect* »). Ils fournissent également une meilleure précision en captant les fluctuations de la PA liées à la vie quotidienne et permettent de capturer l'hypertension « masquée », définie comme une augmentation quotidienne de la PA ambulatoire journalière ($\geq 135/85$ mmHg) face à la PA normale de bureau ($\leq 140/90$ mmHg) [Gilbert-Ouimet et al. (2014)].

3.2.2 La variable d'exposition

Les stresseurs psychosociaux au travail ont été mesurés selon un modèle théorique reconnu, le modèle du déséquilibre effort-reconnaissance (DER). Ce modèle postule qu'un état de déséquilibre entre les efforts déployés dans le cadre du travail (par exemple, les heures supplémentaires, la pression du temps, etc.) et la reconnaissance reçue en contrepartie (par exemple, le salaire), a des conséquences délétères sur la santé [Trudel et al. (2016b)]. Les efforts et la reconnaissance ont été évalués aux trois points de suivi à l'aide d'un instrument d'auto-évaluation validé. Le rapport entre les scores d'effort et de reconnaissance obtenus à partir de cet instrument a été calculé. Le rapport effort/récompense a été dichotomisé, tel qu'un rapport supérieur à 1 indique une exposition au DER [Siegrist (1996)]. Ensuite, l'exposition répétée au DER a été divisée en cinq catégories : jamais exposé (0, 0, 0), exposition intermittente (0, 1, 0 ou 1, 0, 1), exposition qui a cessé au cours du suivi (1, 0, 0 ou 1, 1, 0), exposition *onset* (0, 1, 1 ou 0, 0, 1), exposition chronique (1, 1, 1).

3.2.3 Les covariables

Les variables de confusion incluent le sexe, l'âge au départ, le niveau de scolarité (moins que le collège, le collège terminé, l'université terminée), le tabagisme (actuel ou non-fumeur), la consommation d'alcool ($<$1 verre/ semaine, 1-5 verres / semaine, $\geq$ 6 verres / semaine), les antécédents cardiovasculaires familiaux, le style de vie sédentaire (activité physique $<$1 / semaine ou $\geq$1 / semaine), l'indice de masse corporelle ($<$18,5, 18,5-25, $\geq$25kg/m^2). Les trois premières covariables ne variaient pas dans le temps, tandis que toutes les autres sont dépendantes du temps. Ces covariables ont été sélectionnées a priori, car ce sont des facteurs qui affectent la pression artérielle [Chobanian et al. (2003), Vargas et al. (2000)] et qui sont également potentiellement associées à une exposition au DER [Kouvonen et al. (2006)].

3.2.4 L'application sur la cohorte

Parmi les travailleurs de la cohorte, l'échantillon utilisé dans les analyses comprenait les personnes qui ont participé aux trois examens de suivi, qui avaient la mesure d'exposition au DER aux trois points d'examen de suivi, qui avaient des mesures de la PA au dernier temps de suivi et qui n'étaient pas enceintes lors du dernier examen de suivi, qui n'avaient aucune valeur manquante pour les covariables et travaillaient au moins 21 heures par semaine. L'échantillon final se compose de 1 576 travailleurs, dont 925 femmes et 651 hommes.

Nous avons identifié l'existence possible d'une confusion variante dans le temps. En fait, en plus d'être des facteurs de confusion possibles, certaines covariables sélectionnées peuvent également être affectées par l'exposition à des stresseurs psychosociaux au travail (par exemple, le tabagisme, la consommation d'alcool [Siegrist and Rödel (2006)]). Autrement dit, ces facteurs peuvent agir comme des variables médiatrices qui interviennent sur le chemin causal reliant l'exposition au DER à la PA. Les MSM sont donc une méthode appropriée pour évaluer l'impact d'une exposition répétée au DER sur la PA.

Ces données ont été analysées à l'aide des méthodes de MSM mentionnées dans le chapitre 2 : IPTW, calcul-G, TMLE, TMLE avec *SuperLearner* pour estimer l'issue contrefactuelle en fonction de l'exposition répétée au DER. De plus, nous avons implémenté des méthodes classiques de régression de l'issue sur les variables initiales et sur l'ensemble des variables confondantes et des expositions à tout temps de suivi.

Pour IPTW, nous avons d'abord ajusté trois modèles d'exposition, un pour chaque période de suivi. Nous avons ajusté un modèle de régression logistique où la variable dépendante était l'exposition à la période de suivi considérée, et les variables indépendantes étaient toutes les variables confondantes mesurées à la période de suivi et aux périodes précédentes, ainsi que les variables d'exposition aux périodes précédentes. À partir des résultats de ces modèles, nous avons calculés les poids IPTW tel que décrit dans la section 2.1. Finalement, nous avons ajusté une régression par équations d'estimation généralisées de la PA en fonction de l'exposition répétée sur les données pondérées en utilisant un estimateur robuste de la variance. Cet estimateur robuste de la variance permet d'obtenir des inférences conservatrices [Hernán et al. (2001)].

En ce qui concerne le calcul-g, nous avons commencé par l'ajustement d'un modèle de régression de la PA en fonction de l'exposition cumulée et de l'historique des covariables. Ensuite, nous avons fixé le régime d'exposition à un scénario donné, par exemple $\bar{a} = (0,0,0)$. Le modèle obtenu à la première étape est utilisé pour prédire la PA sous ce scénario. Cette prédiction est notée $\bar{Q}_3$. Ensuite, nous avons ajusté un modèle de l'estimation contrefactuelle $\bar{Q}_3$ en fonction de l'exposition répétée et des covariables aux périodes précédentes. À partir de ce modèle, nous avons obtenu la prédiction notée $\bar{Q}_2$ sous le régime $\bar{a}_2 = (0,0)$. De même, nous avons modélisé $\bar{Q}_2$ en fonction des variables aux périodes précédentes pour obtenir l'estimateur $\bar{Q}_1$ sous l'exposition $\bar{a}_1 = 0$. Ensuite, nous avons pris la moyenne du résultat $\bar{Q}_1$ sur l'ensemble des observations. Cette moyenne est notée $\bar{Q}_0$. Ces étapes ont été répétées pour les huit scénarios possibles, comme c'est indiqué dans la section 2.2. Et puis, nous avons estimé des modèles de régression de nos huit moyennes obtenues sur un modèle paramétrique, tel que la variable dépendante est le vecteur des moyennes contrefactuelles obtenues et la variable indépendante est la catégorie de l'exposition répétée.

Pour l'approche de TMLE, nous avons ajusté trois modèles d'exposition, un pour chaque période de suivi, tel que la variable dépendante est l'exposition au temps de suivi et les variables indépendantes sont celles de l'exposition aux périodes précédentes et les covariables mesurées au temps de suivi et

aux périodes antérieures. Ces modèles ont été utilisés pour trouver les trois prédictions $(\bar{g}_1, \bar{g}_2, \bar{g}_3)$. Ensuite, un modèle de la PA en fonction des variables de l'exposition ainsi que l'historique des covariables a été ajusté. Après cette étape, tel que décrit dans la section 2.4, le reste est similaire à la méthode du calcul-g, sauf qu'une fluctuation est appliquée sur chaque estimateur obtenu à chaque étape. Par exemple, pour $\bar{Q}_3$, nous avons ajusté un modèle de régression de la PA en fonction de l'inverse des probabilités obtenues $(\bar{g}_1, \bar{g}_2, \bar{g}_3)$, avec l'indicatrice indiquant le scénario choisi et le *offset* $\bar{Q}_3$. Ainsi, nous obtenons le nouvel estimateur fluctué $\bar{Q}_3^1$ qui sera utilisé comme variable dépendante dans le modèle de régression en fonction des variables de l'exposition et des variables confondantes aux périodes précédentes.

Pour TMLE avec *SuperLearner*, les librairies que nous avons utilisées pour effectuer les prédictions sont les modèles linéaires généralisés (GLM) et les modèles additifs généralisés (GAM). Il suffit simplement de remplacer les fonctions de régression que nous avons utilisées dans la méthode de TMLE par la fonction *SuperLearner* du package **SuperLearner** en indiquant les librairies sélectionnées. Le modèle de fluctuation reste un modèle de régression linéaire et ne subit pas cette transformation.

Toutes ces approches ont été mises en œuvre sous R-Software à partir de fonctions que nous avons créées. En effet, au moment où ces analyses ont été réalisées, aucun module R publiquement disponible ne semblait permettre de les réaliser. Les erreurs-types ont été estimées à l'aide de la méthode de bootstrap avec 1 000 répétitions, en recalculant les estimations et en prenant l'erreur-type des estimations. Les intervalles de confiance ont été calculés à l'aide de 2,5ème et 97,5ème quantiles des estimations obtenues par le bootstrap.

Le tableau suivant résume les différentes estimations des coefficients associés à l'ordonnée à l'origine (*intercept*) et aux niveaux d'exposition (*intermittent, cessation, onset et chronic*) pour les méthodes ci-dessus, leurs intervalles de confiance et le temps de calcul en secondes (*s*).

TABLE 3.1: Effet causal estimé de l'exposition répétée au DER sur la pression artérielle systolique ambulatoire (en mm Hg)

		IC à 95%		
Méthodes	Estimations	Borne inférieure	Borne supérieure	Temps de calcul (s)
MSM-IPTW				1,36
Intercept	124,413	123,686	125,140	
intermittent	1,490	-0,328	3,310	
cessation	1,112	-0,480	2,700	
onset	0,194	-1,494	1,880	
chronic	1,636	-0,317	3,590	
MSM-calcul-g				203,45
Intercept	124,476	123,801	125,129	
intermittent	1,025	0,286	1,839	
cessation	1,038	-0,170	2,184	
onset	1,012	-0,176	2,376	
chronic	2,050	0,573	3,679	
MSM-TMLE				249,77
Intercept	124,339	123,642	125,115	
intermittent	2,419	0,762	4,293	
cessation	1,286	-0,303	2,791	
onset	1,331	-0,139	3,049	
chronic	1,363	-0,623	3,093	
MSM-TMLE-SL				1227,53
Intercept	124,310	123,643	124,962	
intermittent	2,444	0,863	4,134	
cessation	1,235	-0,174	2,624	
onset	1,489	-0,043	3,042	
chronic	1,458	-0,367	3,301	
approche classique repétée				1,20
Intercept	119,635	115,760	123,509	
intermittent	1,595	-0,070	3,262	
cessation	1,342	-0,104	2,790	
onset	1,328	-0,298	2,955	
chronic	1,204	-0,595	3,004	
approche classique				0,70
Intercept	120,737	116,992	124,481	
intermittent	1,553	-0,118	3,226	
cessation	1,309	-0,137	2,755	
onset	1,602	-0,024	3,229	
chronic	1,461	-0,336	3,258	

Les différentes expositions répétées au DER sont associées à une augmentation de la PA en comparaison au groupe de référence des jamais exposés, bien que les différences n'atteignent que rarement la significativité statistique.

D'ailleurs, ces résultats valident ce qui a été montré dans les études précédentes [Trudel et al. (2018)]

et suggèrent donc que la mise en œuvre des interventions et des politiques visant à réduire durablement les facteurs de stress psychosociaux au travail peuvent contribuer à améliorer la pression artérielle et donc avoir un effet bénéfique sur la santé cardiovasculaire. Bien entendu, les estimations associées à l'exposition à long terme sont positives, significatives et non nulles.

Alors que TMLE (avec ou sans SL) estime que l'exposition intermittente au DER est associée à une augmentation d'environ 2.4 mm Hg de la PA, le calcul-g estime cette association à 1.025 mm Hg et les autres méthodes à environ 1.5 mm Hg. L'association entre l'exposition "onset" et la PA varie également considérablement d'une méthode à l'autre, de 0.194 mm Hg pour IPTW à 1.602 mm Hg pour l'approche classique.

La largeur des intervalles de confiance à 95% varie également considérablement d'une méthode à l'autre. Par exemple, pour l'exposition intermittente, le calcul-g a l'intervalle de confiance le plus court avec une largeur de 1.55, tandis que pour les autres méthodes, l'étendue de l'intervalle de confiance est de l'ordre de 3.3. Pour l'exposition "cessation" et "onset", le calcul-g et TMLE avec SL ont les intervalles de confiance les plus courts avec des largeurs respectivement de 2.35 et de 2.55 pour le calcul-g et de 2.79 et 2.09 pour TMLE avec SL. L'exposition chronique a rapporté des intervalles de confiance de largeur de 3.09 pour le calcul-g et de 3.3 pour l'approche classique répétée.

En présence de confusion dépendante du temps, les trois méthodes MSM sont théoriquement sans biais sous certaines hypothèses de modélisation statistiques. L'IPTW suppose que le modèle pour le traitement est correct, le calcul-g suppose que celui pour l'issue est correct et le TMLE suppose qu'au moins l'un des deux est correct. L'utilisation de l'apprentissage-automatique via le SL pour le TMLE réduit encore davantage les hypothèses statistiques nécessaires.

En théorie, les méthodes classiques ne fonctionnent pas pour une exposition cumulée, mais les résultats montrent qu'il n'y a pas une différence majeure avec les approches de MSM dans cette étude. De plus, la première méthode de l'approche classique répétée consiste à ajuster pour les variables qui sont à la fois confondantes et intermédiaires, elle fournit ainsi des estimations plus petites, alors que la seconde consiste à ajuster uniquement pour les variables initiales.

Nos résultats montrent que la méthode de TMLE avec SL est la plus exigeante du point de vue computationnel. En effet, il a fallu plus de 20.45 minutes avec 1000 échantillons bootstrap pour produire une estimation de notre paramètre d'intérêt avec cette méthode, alors que IPTW n'a pris que 1.36 secondes.

La comparaison que nous avons présentée a révélé certaines différences importantes entre les différentes méthodes. En raison de ses propriétés théoriques, le TMLE est la méthode qui est le plus susceptible d'avoir produit des estimations valides. Toutefois, il est impossible de déterminer si les différences observées entre les méthodes dans cet exemple reflète une différence réelle dans la qualité des estimations ou si elles sont attribuables à des fluctuations aléatoires. Les études de simulations basées sur des données synthétiques permettent d'évaluer la performance de méthodes statistiques dans

un contexte où les vraies valeurs sont connues, mais ont l'inconvénient de ne pas bien refléter la complexité des véritables bases de données. Dans le prochain chapitre, nous proposons des algorithmes de simulation "Plasmode", combinant données réelles et synthétiques, adaptés aux données longitudinales. Ce type de simulation nous permettra d'évaluer et la performance des différentes approches d'ajustement dans un contexte réaliste.

Chapitre 4

Étude de simulation

Les simulations représentent un outil couramment utilisé pour examiner et comparer les caractéristiques de fonctionnement de différentes méthodes statistiques. Bien évidemment, la majorité des données utilisées dans les études épidémiologiques contiennent de nombreuses variables collectées que nous voulons simuler. Pour simuler un tel ensemble de données tout en prenant en compte les associations entre les différentes variables, Vaughan et al. (2009) a suggéré de créer des simulations Plasmode.

Dans ce mémoire, nous décrivons un cadre statistique et informatique pour la création d'ensembles de données de simulation répliquées d'une façon générale en se basant sur la méthode de Vaughan. L'objectif de ce travail est de permettre l'évaluation des approches d'ajustement de confusion dans les données simulées qui préservent les caractéristiques complexes et le contenu informationnel des données, mais qui ont également un véritable effet de traitement connu.

4.1 Le Plasmode

Les simulations Plasmode peuvent aider à surmonter les défis inhérents à la génération d'un ensemble de données simulées à partir de données réelles. Cattell and Jaspars (1967) ont utilisé très tôt ce terme lorsqu'ils ont défini un Plasmode comme un ensemble de valeurs numériques correspondant à un modèle théorique mathématique. Le fait que l'ajustement du modèle soit connu est soit parce que les données simulées sont produites mathématiquement pour s'adapter aux données, soit parce que nous avons une situation réelle dont nous savons avec certitude qu'elle doit produire des données de ce type. Plus précisément, Mehta et al. (2004) décrivent le Plasmode comme un ensemble de données réelles dont la véritable structure est connue.

Cette méthode repose sur le ré-échantillonnage à partir des covariables observées et des données d'exposition sans modifier entièrement les jeux de données générés afin de préserver les associations empiriques entre ces variables.

Les algorithmes Plasmode qui ont été développés jusqu'à présent ne permettent pas d'évaluer les

méthodes indiquées dans le chapitre 2. Dans la plupart des articles, par exemple Schneeweiss et al. (2019), le Plasmode est plutôt appliqué sur des études avec une exposition ponctuelle et non pas sur des études longitudinales avec une exposition et des covariables variant dans le temps.

4.1.1 Le Plasmode pour une exposition ponctuelle

Pour simplifier la présentation, nous considérons le cas d'une variable d'exposition A binaire $(0/1)$ et d'un ensemble de covariables *(baseline)* potentiellement confondantes L.

L'exposition et l'issue sont générées sur la base des coefficients estimés à partir des vraies données [Pang et al. (2016)]. L'exposition ponctuelle A est générée conditionnellement aux covariables L. En effet, les valeurs des paramètres de l'ordonnée à l'origine β_{0_A} et ceux des covariables β_{L_A} sont définis comme les coefficients estimés correspondants à l'ajustement d'un modèle logistique binaire avec la variable réponse A et les variables indépendantes L sur les données réelles. Par conséquent, la variable d'exposition A est générée suivant une distribution Bernoulli, telle que :

$$P(A = 1|L) = \frac{exp(\beta_{0_A} + \beta'_{L_A}\mathbf{L})}{1 + exp(\beta_{0_A} + \beta'_{L_A}\mathbf{L})}.$$

De même, dans l'ajustement du modèle de l'issue, l'ordonnée à l'origine β_{0_Y}, le coefficient bêta associé à A, β_{A_Y}, et le vecteur de coefficients bêta associé à L, β_{L_Y}, sont équivalents aux estimations des coefficients à partir des données réelles. Finalement, l'issue est générée à partir de l'exposition A et des covariables L. Par exemple, pour une issue binaire :

$$P(Y = 1|A,L) = \frac{exp(\beta_{0_Y} + \beta_{A_Y}A + \beta'_{L_Y}L)}{1 + exp(\beta_{0_Y} + \beta_{A_Y}A + \beta'_{L_Y}L)}.$$

L'ensemble du processus de génération de variables est répété J fois. Nous construisons ainsi J jeux de données simulés de taille $m \leq n$, où n est la taille de la cohorte complète.

Pour une exposition cumulée dans le temps, l'algorithme de Plasmode devient plus complexe surtout au niveau computationnel. Nous avons donc développé de nouveaux algorithmes qui permettent de générer des données longitudinales, à partir d'une base de données réelle, tout en prenant en compte les associations entre ces variables afin de comparer les performances des différentes méthodes de MSM.

L'ensemble des algorithmes que nous allons décrire ont été codés comme des fonctions R suffisamment générales pour pouvoir être utilisées dans d'autres contextes que celui que nous étudions dans ce mémoire. Nous effectuons dans la section 4.1.2 et 4.1.4 deux études de simulation Plasmode, l'une en utilisant une approche paramétrique standard et l'autre en utilisant une méthode non paramétrique, notamment les forêts aléatoires.

4.1.2 Le Plasmode paramétrique pour une exposition cumulative

La structure de base de l'algorithme Plasmode comprend les étapes suivantes :

* Disposer d'une base de données originale (cohorte) de taille n sur laquelle les simulations seront basées. Cet ensemble de données fournit toutes les informations que nous utiliserons pour construire les ensembles de données simulées, telles que l'exposition (A_t=1 : si le patient a subi un traitement au temps t, A_t=0 : si le patient n'a pas reçu de traitement au temps t), l'ensemble des covariables L_t à tout temps t, l'issue mesurée à la fin du suivi Y.

* Estimer les associations entre l'exposition, les variables confondantes et l'issue. Nous recommandons de spécifier les covariables qui seront susceptibles d'être associées à l'issue, à l'exposition ou aux facteurs de confusion précédents, telles que l'âge, le sexe...

* Fixer les paramètres, dont la taille de l'échantillon à produire m, le coefficient multiplicateur et le nombre de simulations à effectuer J.

* Effectuer un échantillon aléatoire avec remise de taille $m \leq n$ des données originales.

    1. Simuler l'exposition A_1 :

        a) Construire un modèle de régression de l'exposition A_1 en fonction des covariables L_1 à partir de la banque de données complète.

        b) Extraire le vecteur des coefficients associé à ce modèle $\hat{\beta}_{A_1}$.

        c) Multiplier les valeurs de $\hat{\beta}_{A_1}$ par un coefficient multiplicateur choisi par l'utilisateur. Le nouveau coefficient obtenu par cette étape $\hat{\beta}^*_{A_1}$ sert à fluctuer les données soit dans le sens positif ou négatif, pour affaiblir or renforcer positivement ou négativement les associations entre les variables explicatives et la variable dépendante.

        d) Générer une nouvelle variable dichotomique ${A_1}^*$ à partir du nouveau vecteur de coefficients $\hat{\beta}^*_{A_1}$ selon une loi binomiale avec $P(A^*_1 = 1|L_1) = expit(L_1\hat{\beta}^*_{A1})$ sur les données échantillonnées.

        e) Remplacer A_1 par ${A_1}^*$ dans la nouvelle base de données simulée.

    2. Simuler les covariables L_t et l'exposition A_t, $\forall t = 2, \ldots, k-1$:

pour $t = 2,\ldots,k-1$ **faire**

a) Construire un modèle de régression de chaque élément de l'ensemble des covariables L_t selon sa distribution (logistique, linéaire avec erreurs normales, logistique multinomiale), en fonction de l'exposition $\bar{A}_{t-1}$ et de l'ensemble des covariables antérieures $\bar{L}_{t-1}$ à partir de la banque de données complète.

b) Extraire le vecteur des coefficients associé à ce modèle, $\hat{\beta}_{L_t}$.

c) Multiplier les valeurs de $\hat{\beta}_{L_t}$ par le coefficient multiplicateur prédéfini par l'utilisateur et sauvegarder le nouveau vecteur de coefficients obtenu $\hat{\beta}^*_{L_t}$.

d) Générer une nouvelle variable $L_t{}^*$ à partir du nouveau vecteur de coefficients et des données simulées, selon la distribution de l'élément de L_t et remplacer l'ancienne variable de L_t par la nouvelle variable $L_t{}^*$ dans les données simulées.

e) Construire un modèle de régression de l'exposition A_t en fonction de l'ensemble des covariables $\bar{L}_t$ et de l'exposition antérieure $\bar{A}_{t-1}$ sur les données originales.

f) Extraire le vecteur des coefficients associé à ce modèle $\hat{\beta}_{A_t}$.

g) De même, multiplier les valeurs de $\hat{\beta}_{A_t}$ par le coefficient multiplicateur et obtenir le nouveau vecteur de coefficients $\hat{\beta}^*_{A_t}$.

h) Générer une nouvelle variable dichotomique A^*_t à partir du nouveau vecteur de coefficients selon une loi binomiale sur les données échantillonnées avec :

$$P(A^*_t = 1|\bar{L}_t,\bar{A}_{t-1}) = expit(\bar{L}_t\hat{\beta}^*_{A_{tL}} + \bar{A}_{t-1}\hat{\beta}^*_{A_{tA}})\cdot$$

Le vecteur des coefficients $\hat{\beta}_{A_t}$ se décompose en deux parties : les coefficients estimés $\hat{\beta}^*_{A_{tL}}$ associés au vecteur de covariables $\bar{L}_t$ et les coefficients $\hat{\beta}^*_{A_{tA}}$ associés au vecteur de l'exposition $\bar{A}_{t-1}$.

i) Remplacer A_t par A^*_t.

3. Simuler l'issue Y :

a) Construire un modèle de régression de l'issue Y en fonction de l'ensemble de l'exposition et des covariables respectivement $\bar{A}_t$ et $\bar{L}_t$ $\forall t = 1,\ldots,k-1$ sur les données originales.

b) Remplacer les valeurs des coefficients $\hat{\beta}_y$ par le nouveau vecteur de coefficients $\hat{\beta}^*_y\cdot$

c) Générer une nouvelle variable Y^* à partir du nouveau vecteur de coefficients, et remplacer Y dans la base de données simulée par $Y^*\cdot$

4.1.3 Les étapes de Plasmode paramétrique contrefactuel

L'algorithme précédent permet de simuler des données en se basant sur les relations observées dans les données réelles. Toutefois, le véritable effet de l'exposition cumulée sur l'issue finale ne peut pas

être déterminé directement à partir de cet algorithme. Il s'agit d'un problème commun dans les études de simulation en inférence causale [Morris et al. (2019)].

Une solution simple couramment utilisée consiste à estimer l'effet causal réel à partir de simulations de Monte Carlo. Le principe général est de générer un nouvel ensemble de données en utilisant exactement les mêmes équations, mais en fixant le niveau d'exposition à un niveau donné. Ceci permet au final de simuler les issues contrefactuelles sous ce niveau d'exposition donné.

La moyenne des issues contrefactuelles simulées estime l'espérance contrefactuelle $E(Y^{\bar{a}})$. Lorsque toutes les espérances contrefactuelles sont simulées de cette façon, les véritables valeurs des paramètres du MSM peuvent finalement être estimées en ajustant une régression de l'issue en fonction de l'exposition répétée sur les données contrefactuelles obtenues.

Nous avons ajouté dans la fonction paramétrique de la section 4.1.2 une option qui permet de fixer les valeurs d'exposition. Autrement dit, au lieu que les expositions soient générées d'une façon aléatoire, elles seraient fixées à des valeurs décidées par l'utilisateur lorsque cette option est utilisée. Par exemple, pour une exposition $A_3 = (1,0,1)$, on simulera l'issue et les covariables telles que tous les individus ont exactement les mêmes valeurs de l'exposition sélectionnée. Dans ce cas, on n'aura pas besoin de nombre de simulations ou de la taille de la base de données simulées. On simulera une seule banque de la taille originale.

L'avantage de telles simulations, c'est qu'elles nous permettent de calculer les vraies valeurs des paramètres et que nous pouvons donc tester les performances des méthodes que nous utiliserons après pour calculer le biais, la déviation standard (Std), la racine de l'erreur quadratique moyenne (RMSE) et la couverture de l'intervalle de confiance.

4.1.4 Le Plasmode non paramétrique pour une exposition cumulative

Pour générer les données de Plasmode d'une façon non paramétrique, nous commençons tout d'abord par l'ajustement des modèles de forêt aléatoire du traitement et de chaque covariable sur les données originales pour générer les nouvelles variables des simulations Plasmode en utilisant le package **randomForest** sous R, avec les paramètres par défaut. De même, un modèle de forêt aléatoire de l'issue en fonction des variables de l'exposition et des covariables à tout temps de suivi est ensuite ajusté à partir des données complètes. L'algorithme suivant résume les étapes du Plasmode non paramétrique :

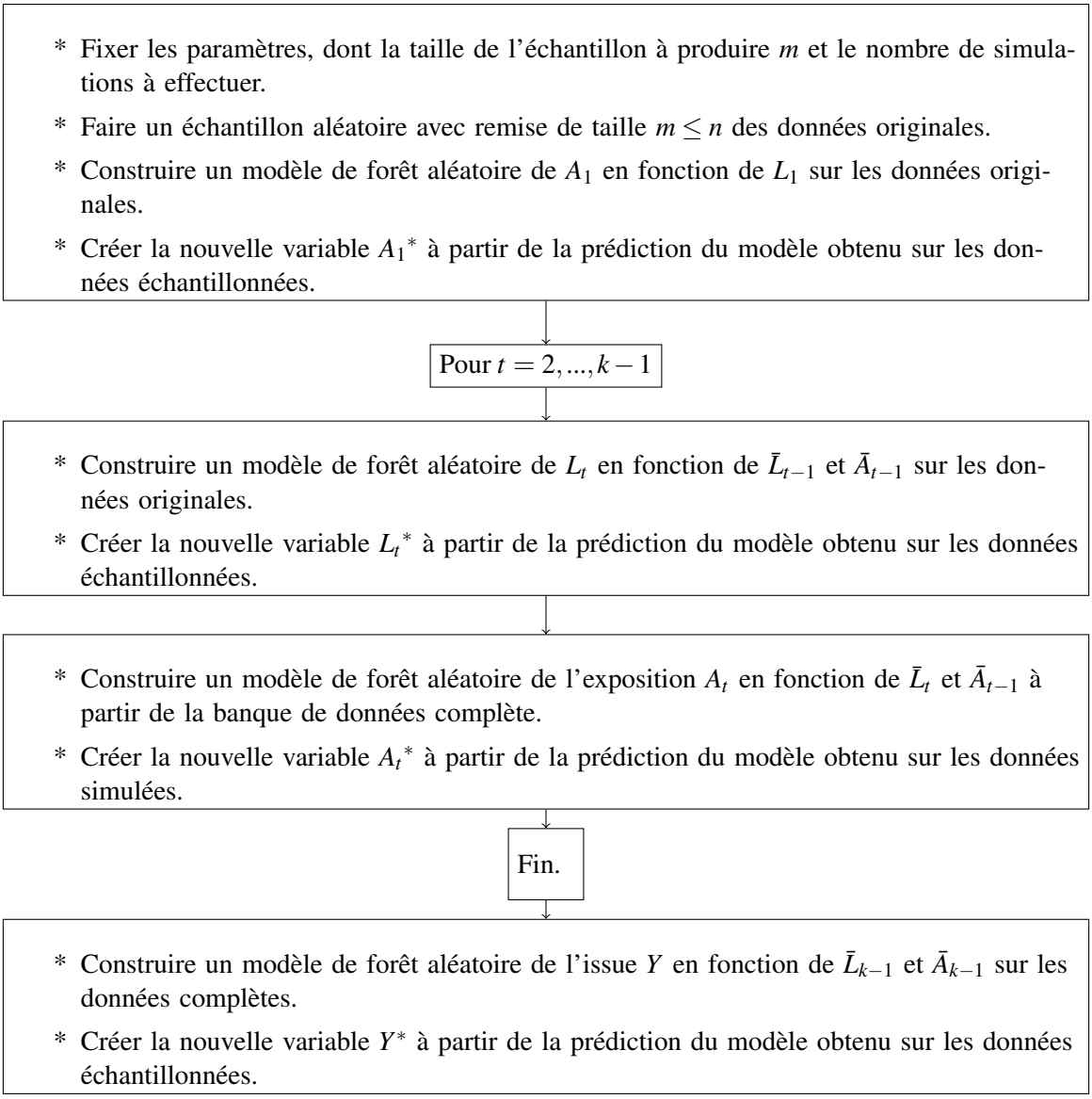


FIGURE 4.1 – Algorithme de Plasmode non paramétrique

4.1.5 Les étapes de Plasmode contrefactuel non paramétrique

Il s'agit du même principe que pour le Plasmode contrefactuel paramétrique. En effet, l'utilisateur choisit au préalable les scénarios sous lesquels il veut faire ses simulations. Les prédictions sont ainsi effectuées sur les individus qui ont exactement les mêmes valeurs de l'exposition pour chaque scénario.

Les vrais coefficients sont calculés à partir de cette fonction de Plasmode contrefactuelle, en spécifiant que l'option prédéfinie de l'exposition est égale à la matrice $(2^{k-1} \times [k-1])$ de tous les scénarios possibles, c'est-à-dire de 2^{k-1} régimes possibles.

En bref, nous simulons les données selon tous les régimes contrefactuels, puis nous roulons la régression de l'issue observée en fonction de l'exposition répétée sur toutes les données contrefactuelles. Ceci nous donne les vraies valeurs des paramètres associées à chaque exposition répétée.

4.1.6 Le Plasmode appliqué sur la cohorte

Approche paramétrique :

En utilisant les simulations Plasmode et grâce aux données sur la cohorte, nous avons comparé les méthodes d'ajustement de confusion suivantes : IPTW, calcul-g, TMLE, TMLE avec *SuperLearner* et les approches classiques de confusion avec ajustement pour les covariables initiales et pour les covariables variant dans le temps, afin d'évaluer l'effet des stresseurs psychosociaux au travail sur la PA.

Tout d'abord, nous avons extrait des données originales un échantillon aléatoire avec remplacement de taille $m = 500$. Ensuite, pour chaque observation, une nouvelle exposition simulée, des nouvelles covariables et une nouvelle issue simulée ont été générées à partir des modèles de régression, en utilisant un coefficient multiplicateur égal à 1.5, comme indiqué dans la section 4.1.2.

Sur cette base, nous avons simulé $J = 500$ ensembles de données indépendants. Par la suite, nous avons obtenu les coefficients estimés et les écarts-type associés à chaque méthode pour chaque exposition répétée ainsi que les intervalles de confiance à 95%. Pour l'IPTW, l'intervalle de confiance a été obtenu avec un estimateur robuste de la variance, alors que pour le calcul-g, le TMLE et TMLE-SL, l'intervalle de confiance correspondait à $\pm$ 1,96 fois l'écart-type de 50 réplications de bootstrap non paramétrique. Cette stratégie de bootstrap a été utilisée, plutôt que celle basée sur les percentiles, afin de réduire les temps de l'exécution des simulations.

Pour calculer les vraies valeurs des coefficients, nous avons utilisé la fonction de Plasmode paramétrique contrefactuel de la section 4.1.3 en spécifiant tous les scénarios possibles ($2^3 = 8$ au total). Après, nous avons créé une nouvelle base de données contenant les issues contrefactuelles simulées et les différentes expositions. Finalement, nous avons effectué un modèle de régression de l'issue (PA) en fonction de l'exposition répétée sur le jeu de données créé. Les coefficients obtenus de cette régression sont équivalents aux vraies valeurs des coefficients.

Sur ce, nous avons comparé les méthodes d'ajustement en utilisant trois mesures : le biais, le RMSE et la couverture de l'intervalle de confiance :

* Le biais relatif estimé est défini comme étant la différence entre la moyenne des coefficients et les vrais valeurs des coefficients sur les J ensembles, tel que :

$$\text{biais relatif} = \frac{\text{moyenne des coefficients estimés} - \text{vrais coefficients}}{\text{vrais coefficients}} \times 100\cdot$$

  Cette formule donne le pourcentage par lequel les coefficients estimés diffèrent de la valeur

réelle des coefficients dans toutes les réplications. Des valeurs négatives de ce biais indiquent que les coefficients sont sous-estimés.

* Le RMSE est écrit sous la forme de la racine de la somme de la variance de l'estimateur et du biais au carré de l'estimateur, tel que :

$$RMSE = \sqrt{std^2 + biais^2}.$$

* La couverture de l'intervalle de confiance à 95% estimée est calculée comme la proportion des intervalles de confiance ayant la borne inférieure plus petite que la vraie valeur du paramètre et la borne supérieure plus grande que la vraie valeur du paramètre.

Pour certaines méthodes, il y a eu des problèmes de convergence dans quelques cas. En effet TMLE n'a pas convergé pour 25 unités, alors que TMLE avec SL n'a pas convergé pour 20 observations (individus). Les calculs des résultats pour le TMLE et TMLE-SL ont été effectués en excluant leurs réplications problématiques respectives. Nous trouvons au tableau 4.1 les résultats pour les simulations de Plasmode paramétrique.

TABLE 4.1: Estimation de l'effet de l'exposition répétée au DER sur la pression artérielle systolique ambulatoire (en mm Hg) avec un échantillon de 500 individus par l'approche paramétrique.

	Méthodes	Intercept	intermittent	cessation	onset	chronic
Biais Relatif	*IPTW*	1.31	13.44	19.28	-5.29	1.04
	calcul-g	1.27	10.45	28.99	-5.91	8.01
	TMLE	1.31	6.47	10.54	-1.10	-0.42
	TMLE-SL	1.31	5.99	11.43	-1.66	-0.63
	classique répétée	-5.54	13.78	-0.12	-5.24	-11.07
	classique	-5.58	36.56	9.06	27.82	0.75
Std	*IPTW*	0.75	2.90	2.13	2.17	1.36
	calcul-g	0.68	0.63	1.27	1.16	1.25
	TMLE	0.70	1.77	1.69	1.48	1.23
	TMLE-SL	0.70	1.78	1.68	1.48	1.23
	classique répétée	3.79	1.82	1.54	1.51	1.24
	classique	3.56	1.82	1.55	1.52	1.24
RMSE	*IPTW*	1.82	2.90	2.14	2.17	1.36
	calcul-g	1.74	0.65	1.34	1.17	1.28
	TMLE	1.80	1.78	1.70	1.48	1.23
	TMLE-SL	1.79	1.79	1.69	1.48	1.23
	classique répétée	7.98	1.83	1.54	1.52	1.29
	classique	7.91	1.90	1.55	1.58	1.24
Couverture IC	*IPTW*	0.43	0.93	0.95	0.95	0.98
	calcul-g	0.34	0.94	0.93	0.96	0.94
	TMLE	0.33	0.93	0.94	0.94	0.95
	TMLE-SL	0.33	0.93	0.95	0.93	0.94
	classique répétée	0.53	0.96	0.96	0.95	0.94
	classique	0.48	0.95	0.96	0.93	0.95

Les résultats obtenus dans ce tableau et dans la figure 4.2 montrent qu'en général le plus grand biais est présent dans les approches classiques qui n'utilisent aucun ajustement pour la confusion dépendante du temps, surtout pour la catégorie intermittente. Aussi, le biais de calcul-g reste assez élevé par rapport aux autres méthodes de MSM. Généralement, l'approche de TMLE avec/sans SL présente les meilleurs résultats par rapport au biais pour certaines catégories. En utilisant *SuperLearner*, les performances de TMLE ont été légèrement améliorées en termes de RMSE pour la catégorie *cessation*, mais pas pour les autres mesures de performances.

En termes de RMSE, le plus petit est celui de calcul-g et de TMLE avec/sans SL pour presque tous les paramètres, avec des valeurs respectivement de 1.74 et 1.79, alors que celui de IPTW et des approches classiques était plus élevé avec une valeur de 1.82. Bien que le biais de la méthode IPTW soit faible (1.31), l'erreur-standard est élevée (0.75), ce qui entraîne une sur-couverture de l'intervalle de confiance, en particulier pour la catégorie chronique. En fait, c'est parce que nous avons utilisé un estimateur conservateur de la variance (l'estimateur robuste).

La majorité des méthodes ont donné des intervalles de confiance à 95% avec une couverture assez proche du niveau attendu (95%). Pourtant, le taux de couverture pour l'ordonnée à l'origine est faible pour toutes méthodes. C'est d'autant plus surprenant que le biais est faible. En effet, l'erreur-type est sous-estimée, c'est-à-dire que la moyenne des variances estimées est inférieure à la variance des coefficients estimés (résultats non présentés). Il en résulte une petite couverture de l'IC d'environ 40%, et donc des intervalles de confiance qui sont en moyenne trop courts.

Approche non paramétrique :

Après avoir généré des données à l'aide de la méthode paramétrique, nous avons simulé 500 ensembles de données indépendants, en utilisant l'approche non paramétrique de Plasmode pour comparer entre elles les méthodes d'ajustement de confusion en utilisant les mêmes mesures de performance qu'avec le Plasmode paramétrique.

Les vraies valeurs des coefficients sont calculés suivant les mêmes étapes de l'approche paramétrique, sauf qu'au lieu d'utiliser la fonction de Plasmode paramétrique contrefactuelle, nous avons utilisé la fonction de Plasmode non paramétrique contrefactuel de la section 4.1.5.

TABLE 4.2: Estimation de l'effet de l'exposition répétée au DER sur la pression artérielle systolique ambulatoire (en mm Hg) avec un échantillon de 500 par l'approche non paramétrique.

	Méthodes	Intercept	intermittent	cessation	onset	chronic
Biais Relatif	*IPTW*	0.41	-27.51	23.80	-34.15	25.08
	calcul-g	0.37	-36.76	-24.54	-30.51	-1.01
	TMLE	0.37	-14.10	-59.34	-1.95	-41.10
	TMLE-SL	0.35	-12.40	-54.31	5.93	-36.23
	classique répétée	-3.77	-59.44	-56.41	7.88	-52.05
	classique	-2.72	-64.04	-51.14	51.38	-43.04
Std	*IPTW*	0.31	1.66	2.38	2.02	3.51
	calcul-g	0.29	0.71	1.09	0.95	1.42
	TMLE	0.30	1.11	1.17	0.92	2.07
	TMLE-SL	0.30	1.15	1.19	0.90	2.01
	classique répétée	1.55	0.96	1.27	0.94	2.08
	classique	1.50	0.94	1.21	0.99	2.14
RMSE	*IPTW*	0.59	1.70	2.40	2.07	3.54
	calcul-g	0.54	0.88	1.14	1.02	1.42
	TMLE	0.55	1.13	1.40	0.92	2.20
	TMLE-SL	0.53	1.17	1.38	0.90	2.11
	classique répétée	4.93	1.28	1.46	0.94	2.29
	classique	3.69	1.31	1.38	1.17	2.28
Couverture IC	*IPTW*	0.69	0.89	0.83	0.88	0.69
	calcul-g	0.65	0.85	0.92	0.93	0.92
	TMLE	0.62	0.92	0.87	0.94	0.88
	TMLE-SL	0.65	0.89	0.88	0.93	0.91
	classique répétée	0.15	0.82	0.79	0.92	0.76
	classique	0.40	0.83	0.82	0.87	0.76

D'après le tableau 4.2 et la figure 4.2, les résultats obtenus avec la simulation non paramétrique de Plasmode étaient positifs, dans la mesure où une réduction du biais a été observée pour l'approche de TMLE avec *SuperLearner* par rapport aux autres méthodes. En effet, le biais de TMLE avec *SuperLearner* était de 0.35, alors que celui associé à l'IPTW était de 0.41. En ce qui concerne les méthodes classiques, il y avait un biais supérieur près de 3. La couverture des intervalles de confiance était inférieure au niveau attendu (95%) surtout pour les approches classiques, tandis que l'approche de TMLE avec SL a donné les meilleures performances pour certaines catégories de l'exposition, et ce pour la couverture, le biais, l'écart-type et le RMSE. Ensuite, le calcul-g a également donné des résultats comme même performants avec une couverture d'intervalle de confiance appropriée. Cependant, l'approche de IPTW a donné de mauvais résultats dans le cadre des modèles structurels marginaux.

Simulations paramétriques vs non paramétriques :

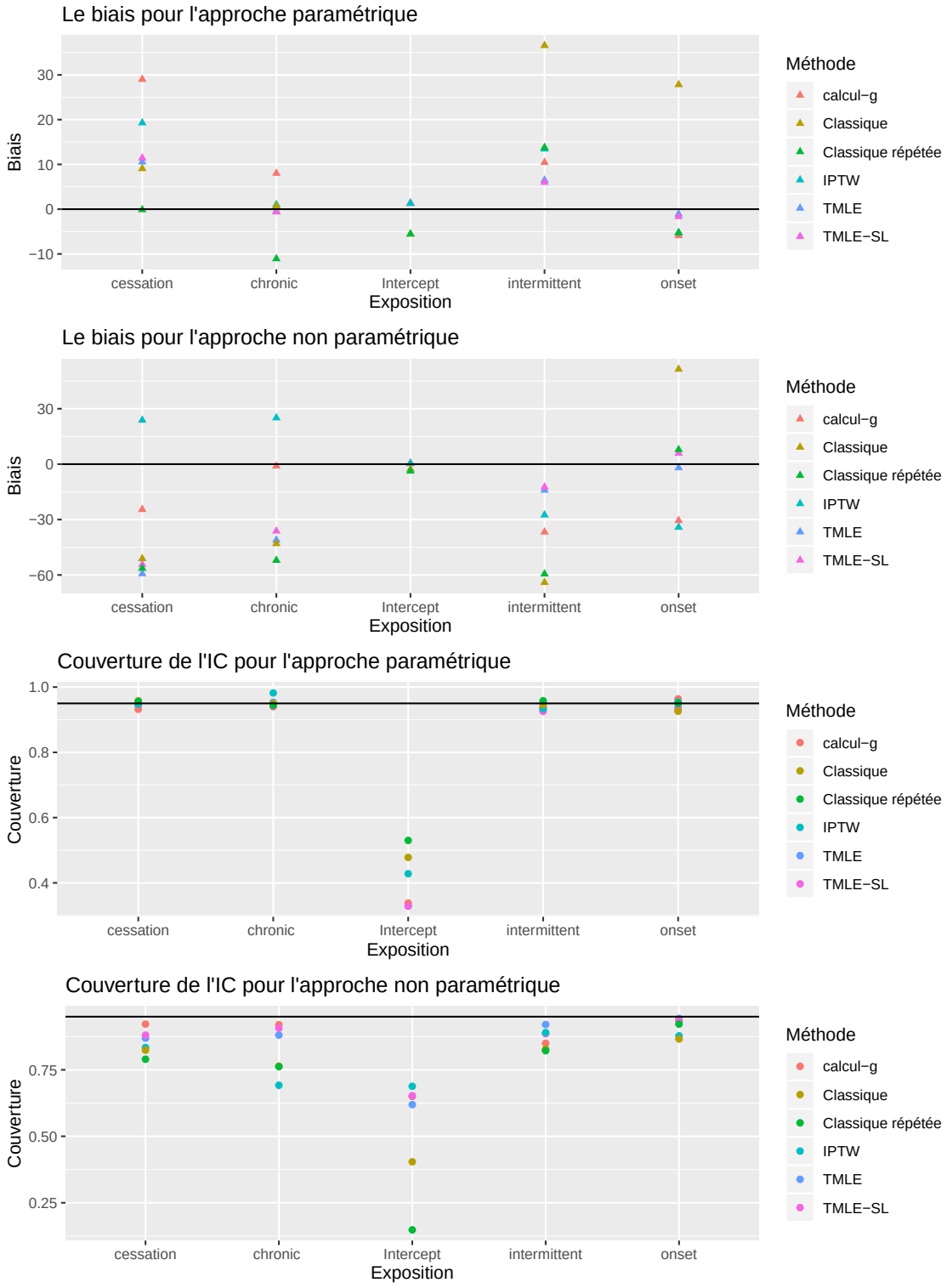
Dans le contexte des simulations paramétriques, les méthodes sont souvent testées en analysant des données simulées selon un modèle supposé ou une distribution connue. Cette stratégie de test peut aboutir à une vision trop optimiste des performances des méthodes utilisées [Benidt and Nettleton

(2015)].

Bien évidemment, sur la base des résultats obtenus, nous avons constaté que les mesures de performances des méthodes paramétriques utilisées, y compris les approches de MSM et les approches classiques, sont généralement meilleures pour les simulations paramétriques que pour celles non paramétriques. Toutefois, des études de simulations basées sur une approche paramétrique peuvent donner une vue trompeuse de l'efficacité des méthodes utilisées. Donc, nous pensons que l'utilisation d'une approche non paramétrique, sans faire d'hypothèse de distribution paramétrique, donne une image plus précise des performances de ces méthodes.

Sachant que nous ne nous attendions pas à des résultats parfaits, les performances des simulations non paramétriques sont tout de même satisfaisants. En fait, la méthode doublement robuste de TMLE avec/sans *SuperLearner* se comporte de façon similaire entre les deux méthodes de simulation et il n'y a pas de grande différence entre les mesures des performances des deux approches de simulations.

FIGURE 4.2 – Comparaisons de biais et de couverture de l'intervalle de confiance entre les méthodes



Conclusion

Dans ce mémoire, nous nous sommes intéressés à la problématique de l'analyse comparative des performances des méthodes de correction de biais de confusion dans un contexte d'une exposition répétée dans le temps. Les approches de MSM énoncées dans la présente étude n'ont pratiquement jamais été utilisées dans le domaine des stresseurs psychosociaux au travail et la pertinence de les utiliser n'a jamais été évaluée dans ce contexte.

Cette étude nous a permis donc de mieux comprendre les performances des méthodes d'ajustement. En effet, parmi toutes les méthodes comparées, celle du TMLE avec *SuperLearner* apparaît comme étant la plus robuste selon les critères du biais, de l'erreur quadratique moyenne et de la couverture de l'intervalle de confiance. De plus, le calcul-g a donné de bons résultats proches de TMLE avec *SuperLearner*. La performance de IPTW est, quant à elle, faible par rapport aux autres méthodes d'ajustement. En général, la majorité des approches de MSM ont pu réduire le biais par rapport aux approches classiques de régression. Toutefois, il est important de reconnaître que les MSM sont des approches qui sont plus difficiles à implanter que les méthodes classiques et requièrent des ressources computationnelles plus importantes, en particulier pour TMLE avec *SuperLearner*.

Certes, la méthode de TMLE avec *SuperLearner* semble être la plus performante dans le contexte d'une étude longitudinale avec des expositions cumulées aux stresseurs psychosociaux au travail, cependant, il y a un effort supplémentaire à payer surtout au niveau computationnel. Nous avons constaté à partir des résultats que la méthode de TMLE avec SL prend presque une demi-heure pour rouler sur une seule base de données, alors que les approches classiques prennent quelques secondes pour donner les résultats. De surcroît, les approches classiques ont donné une couverture de l'intervalle de confiance acceptable, mais des biais qui sont très élevés.

Puisque les modèles de MSM gagnent en popularité dans le contexte des expositions psychosociales au travail, nous croyons que les résultats que nous avons obtenus sont extrêmement importants pour les applications futures des MSMs. En effet, dans le cas d'une exposition variant dans le temps, les modèles structurels marginaux sont le plus souvent utilisés. Pour garantir une estimation sans biais de l'effet causal, ces derniers doivent satisfaire certaines conditions, notamment la bonne spécification du modèle reliant l'issue à l'exposition.

En fait, dans la mesure où l'on ne connait pas le vrai modèle causal qui relie l'issue à l'historique

d'exposition, il est très difficile de valider cette hypothèse. Bien entendu, ces méthodes qui sont théoriquement supérieures n'auront pas un si grand avantage dans ce contexte particulier probablement en raison de la mauvaise spécification du modèle de la pression artérielle en fonction de l'historique de l'exposition qui peut produire des estimations biaisées.

À la vue de nos résultats, nous recommandons aux chercheurs du domaine des stresseurs psychosociaux au travail d'utiliser les approches de MSM, en particulier la méthode de TMLE avec *SuperLearner*. Il est vrai que c'est une méthode pertinente, mais elle nécessite que les analystes disposent de ressources nécessaires pour l'implanter.

De plus, notre étude vient apporter un nouveau souffle à la littérature notamment du fait que nous avons introduit la simulation Plasmode appliquée sur des données longitudinales. En effet, les études antérieures avaient, généralement, été réalisées dans le contexte de l'exposition binaire. Le cas de l'exposition cumulée n'a jamais été exploré dans la littérature.

Finalement, notre contribution peut être une piste de réflexions pour d'autres recherches. En fait, l'amélioration des fonctions de Plasmode longitudinal que nous avons créées pour comparer la performance des différentes méthodes de MSM peut faire l'objet de recherches futures. Une option possible consiste à ajouter un argument qui spécifie le sens de fluctuation des données, soit au sens positif ou négatif, dépendamment des variables. D'autres travaux pourraient impliquer l'ajout d'une option qui spécifie un coefficient multiplicateur différent associé à chaque variable pour renforcer ou affaiblir ces liens avec l'exposition, l'issue ou les covariables.

Bibliographie

LJ Appel. Lifestyle modification as a means to prevent and treat high blood pressure. *Journal of American Society of Nephrology*, 14 :99–102, 2003.

S Benidt and D Nettleton. Simseq : a nonparametric approach to simulation of rna-sequence datasets. *Statistics in Medicine*, 13 :2131–2140, 2015.

J Bouyer, D Hémon, S Cordier, and F Derriennic. *Épidémiologie*. Tec & Doc Lavoisier, 2003.

L Breiman. Stacked regressions. *Machine learning*, 24 :49–64, 1996.

C Brisson, J Moisan, M Vézina, A Vinet, and GR Dagenais. Maladies cardiovasculaires et environnement de travail, protocole (cardiovascular diseases and the work environment, protocol). *Canadian Institutes of Health Research*, 1994.

C Brisson, N Laflamme, and J Moisan. Environnement psychosocial, maladie coronarienne et tension artérielle, protocole (psychosocial environment, coronary heart disease and blood pressure, protocol). *Canadian Institutes of Health Research*, 2001.

NR Campbell, E Burgess, G Taylor, E Wilson, J Cléroux, JG Fodor, L Leiter, and JD Spence. Lifestyle changes to prevent and control hypertension : do they work ? *Canadian Medical Association journal*, 160 :1341–1343, 1999.

RB Cattell and J Jaspars. A general plasmode (no. 30-10-5-2) for factor analytic exercises and research. *Multivariate Behavioral Research*, 67 :1–212, 1967.

AV Chobanian, GL Bakris, HR Black, WC Cushman, LA Green, JL Jr Izzo, DW Jones, BJ Materson, S Oparil, JT Jr Wright, and EJ Roccella. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure : the jnc 7 report. *Journal of the American Medical Association*, 289(19) :2560–2571, 2003.

N Claes and N Jacobs. The precardio-study protocol - a randomized clinical trial of a multidisciplinary electronic cardiovascular prevention programme. *BMC Cardiovascular Disorders*, 4(7) :27, 2007.

NR Cook, IM Lee, JM Gaziano, D Gordon, PM Ridker, JE Manson, CH Hennekens, and JE Buring. Low-dose aspirin in the primary prevention of cancer : the women's health study : a randomized controlled trial. *The Journal of the American Medical Association*, 294(1) :47–55, 2005.

DR Cox and N Wermuth. Causality : a statistical view. *International Statistical Review*, 72 :285–305, 2004.

M Gilbert-Ouimet, X Trudel, C Brisson, A Milot, and M Vézina. Adverse effects of psychosocial work factors on blood pressure : systematic review of studies on demand-control-support and effort-reward imbalance models. *Scandinavian Journal Of Work Environment and Health*, 40(2) :109–132, 2014.

NS Godtfredsen, E Prescott, and M Osler. Effect of smoking reduction on lung lancer risk. *The Journal of the American Medical Association*, 294(12) :1505–1510, 2005.

S Gruber and MJ Van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1) :26, 2010.

MA Hernán. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4) :265–271, 2004.

MA Hernán, B Brumback, and JM Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 11(5) :561–570, 2000.

MA Hernán, B Brumback, and JM Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96 :440–448, 2001.

DJ Horvitz, DG et Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 :663–685, 1952.

EH Kennedy. Semiparametric theory and empirical processes in causal inference. *ICSA Book Series in Statistics*, 8 :141–167, 2016.

A Kouvonen, M Kivimäki, M Virtanen, T Heponiemi, M Elovainio, J Pentti, A Linna, and J Vahtera. Effort-reward imbalance at work and the co-occurrence of lifestyle risk factors : cross-sectional survey in a sample of 36,127 public sector employees. *BMC Public Health*, 6(1) :1–24, 2006.

CM Lawes, HS Vander, MR Law, P Elliott, S MacMahon, and A Rodgers. Blood pressure and the burden of coronary heart disease, in coronary heart disease epidemiology : from aetiology to public health. *Oxford University Press*, 2005.

B Lecoutre. Expérimentation, inférence statistique et analyse causale. *Revue de l'Association pour la Recherche Cognitive*, 2004.

JH Markovitz, KA Matthews, M Whooley, CE Lewis, and KJ Greenlund. Increases in job strain are associated with incident hypertension in the cardia study. *Annals Of Behavioral Medicine*, 28(1) : 4–9, 2004.

T Mehta, M Tanik, and DB Allison. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nature Genetics*, 36 :943–947, 2004.

TP Morris, IR White, and MJ Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11) :2074–2102, 2019.

M Pang, T Schuster, KB Filion, ME Schnitzer, M Eberg, and RW Platt. Effect estimation in point-exposure studies with binary outcomes and high-dimensional covariate data- a comparison of targeted maximum likelihood estimation and inverse probability of treatment weighting. *The International Journal of Biostatistics*, 2016.

ML Petersen, KE Porter, S Gruber, Y Wang, and MJ Van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods In Medical Research*, 21 :31–54, 2012.

EC Polley, E LeDell, C Kennedy, S Lendle, and MJ Van der Laan. Package "superlearner". *2.0-4 ed*, 2011.

JM Robins. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *The International Journal of Computer Mathematics*, 14 :923–945, 1987.

JM Robins, MA Hernán, and B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5) :550–560, 2000.

DB Rubin. Bayesian inference for causal effects : the role of randomization. *The Annals of Statistics*, 6(1) :34–58, 1978.

A Sall, K Aubé, X Trudel, C Brisson, and D Talbot. A test for the correct specification of marginal structural models. *Statistics in Medicine*, 00 :1–6, 2019.

S Schneeweiss, R Wyss, MJ Van der Laan, SD Lendle, C Ju, and JM Franklin. Methods for improving confounding control in comparative effectiveness research using electronic healthcare databases. *Patient-Centered Outcomes Research Institute*, 2019.

ME Schnitzer, MJ Van der Laan, EE Moodie, and RW Platt. Effect of breastfeeding on gastrointestinal infection in infants : a targeted maximum likelihood approach for clustered longitudinal data. *Annals of Applied Statistics*, 8(2) :703–725, 2014.

J Siegrist. Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology*, 1(1) :1–27, 1996.

J Siegrist and A Rödel. Work stress and health risk behavior. *Scandinavian journal of work, environment & health*, 32(6) :473–481, 2006.

SA Swanson, MA Hernán, M Miller, JM Robins, and TS Richardson. Partial identification of the average treatment effect using instrumental variables : Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522) :933–947, 2018.

X Trudel, C Brisson, A Milot, B Masse, and M Vézina. Effort-reward imbalance at work and 5-year changes in blood pressure : the mediating effect of changes in body mass index among 1400 white-collar workers. *International archives of occupational and environmental health*, 89(8) : 1229–1238, 2016a.

X Trudel, C Brisson, A Milot, B Masse, and M Vézina. Adverse psychosocial work factors, blood pressure and hypertension incidence : repeated exposure in a 5-year prospective cohort study. *Journal of Epidemiology and Community Health*, 70(4) :402–408, 2016b.

X Trudel, C Brisson, M Gilbert-Ouimet, and A Milot. Psychosocial stressors at work and ambulatory blood pressure. *Current Cardiology Reports*, 20(12) :127, 2018.

MJ Van der Laan and D Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2 :1–40, 2006.

MJ Van der Laan, EC Polley, and AE Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6 :1–25, 2007.

CM Vargas, DD Ingram, and RF Gillum. Incidence of hypertension and educational attainment the NHANES I epidemiologic followup study. *American Journal of Epidemiology*, 152(3) :272–278, 2000.

LK Vaughan, J Divers, MA Padilla, DT Redden, HK Tiwari, D Pomp, and DB Allison. The use of plasmodes as a supplement to simulations : a simple example evaluating individual admixture estimation methodologies. *Computational Statistics Camp ; Data Analysis*, 53 :1753–1766, 2009.

PK Whelton, J He, LJ Appel, JA Cutler, S Havas, TA Kotchen, EJ Roccella, R Stout, C Vallbona, MC Winston, and J Karimbakas. Primary prevention of hypertension : clinical and public health advisory from the national high blood pressure education program. *The Journal of the American Medical Association*, 288 :1882–1888, 2002.

(WHO) World Health Organization. Cardiovascular diseases. *Fact sheet*, 2011.