UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Moyennage de modèles pour l'estimation d'effets causaux avec la méthode de pondération par les probabilités inversées

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MALORIE CHABOT-BLANCHET

MARS 2011

REMERCIEMENTS

L'aboutissement de ce mémoire a été possible avec l'aide de plusieurs personnes qui m'ont apporté un important soutien tout au long de mes études.

Je tiens à remercier sincèrement ma directrice de recherche, Geneviève Lefebvre, pour tout le temps qu'elle a consacré à la réalisation de ce projet et aux nombreuses relectures de ce mémoire. Je la remercie aussi pour son enthousiasme qui m'a toujours motivée à continuer et à me dépasser, ainsi que pour toute la confiance qu'elle m'a accordée.

Merci aux professeurs du département de mathématiques, et plus particulièrement à Fabrice Larribe et Sorana Froda, pour leur encadrement, leur aide et leurs conseils tout au long du baccalauréat et de la maîtrise.

Je tiens aussi à remercier le personnel de soutien du département de mathématiques. Un merci particulier à Manon Gauthier et Gisèle Legault pour leur aide.

Merci aussi à mes collègues et amis de l'université, notamment Hugues, Jérôme et Marie, pour ces années d'études agréablement passées en votre compagnie.

Merci à Vincent pour sa patience et sa bienveillance à mon égard, ainsi que pour son soutien au quotidien.

Mes remerciements vont également à ma famille et mes amis, pour leur écoute et leurs encouragements durant les dernières années.

Enfin, je remercie le Conseil de recherches en sciences naturelles et en génie du Canada, le Fonds québécois de la recherche sur la nature et les technologies et l'Université du Québec à Montréal et pour leur soutien financier sous forme de bourses d'études.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Pour estimer un effet causal dans les études d'observation en épidémiologie, les méthodes de pondération par les probabilités inversées et les méthodes doublement robustes sont couramment utilisées. Il n'est toutefois pas facile de spécifier correctement le modèle de traitement et les estimateurs associés sont particulièrement sensibles à un choix de modèle incorrect. Le but principal de ce projet est de déterminer si le fait de prendre une moyenne sur plusieurs modèles pourrait améliorer la performance des estimateurs par pondération par les probabilités inversées, en comparaison à une estimation basée sur un seul modèle. Pour ce faire, nous utilisons les critères d'ajustement AIC et BIC pour associer un poids (probabilité) à chacun des modèles. Nous nous intéressons plus particulièrement à deux façons d'utiliser ces poids 1) soit la pondération externe qui considère une moyenne des estimations obtenues sous chacun des modèles de l'ensemble des modèles considérés, et 2) la pondération interne qui effectue une moyenne des scores de propension obtenus sous chacun des modèles pour ensuite obtenir l'estimation correspondante. Nous comparons les résultats obtenus sous chacun des modèles individuellement, puis sous les différentes façons proposées de considérer un ensemble de modèles. Nous regardons la performance des techniques lorsque le vrai modèle fait ou ne fait pas partie des modèles considérés. Nous obtenons que la pondération apporte un compromis intéressant pour pallier l'incertitude reliée à la sélection du modèle de traitement. Nous observons que l'estimateur basé sur la pondération interne semble avoir une variance plus petite que l'estimateur basé sur la pondération externe et que l'estimateur par pondération par les probabilités inversées employé sur les modèles individuellement, surtout lorsqu'ils sont appliqués sur des échantillons de petite taille.

Mots clés : estimation causale, sélection de modèle, moyennage de modèles, étude d'observation, pondération par probabilités inversées.

INTRODUCTION

L'épidémiologie est un domaine d'application de la statistique qui est au coeur des études effectuées en santé, que ce soit lors de la mise à l'essai de nouveaux médicaments et traitements préventifs, ou encore lorsque nous évaluons l'impact de facteurs de risque pouvant influencer l'apparition de maladies. Un des grands défis statistiques que nous y retrouvons est d'estimer l'effet causal d'un certain facteur lorsque les données proviennent d'une étude d'observation plutôt que d'une étude expérimentale. Bien que les études expérimentales soient considérées comme le contexte idéal pour l'estimation d'effets de type causal, en réalité, les études d'observation sont beaucoup plus fréquentes. Une des principales difficultés rencontrées dans la pratique est la présence des facteurs confondants qui compliquent l'estimation de l'effet causal. L'effet de la variable d'intérêt peut être difficile à isoler, surtout à l'intérieur d'études d'observation. Différentes techniques sont utilisées pour contrôler les variables confondantes et obtenir une estimation sans biais. Cependant, pour ce faire, le statisticien doit faire la supposition que toutes les variables confondantes sont mesurées et qu'il a choisi le bon modèle pour représenter les relations entre les variables. Pour cette étape aussi, différentes méthodes existent pour effectuer le choix du modèle.

La technique d'estimation d'un effet causal à laquelle nous nous intéressons pour ce projet est l'estimation par la pondération par les probabilités inversées. Elle requiert la spécification d'un modèle de prédiction pour le traitement et l'estimateur est très sensible au choix de ce modèle. Cependant, il n'existe pas de façon de confirmer que le modèle choisi est le bon et souvent l'incertitude au niveau de la sélection du modèle n'est pas prise en compte. Le but de ce projet est de pallier cette incertitude en proposant une approche d'analyse qui incorpore un ensemble de modèles possibles. Nous étudions une technique de moyennage qui consiste à pondérer des quantités estimées en fonction des

probabilités des modèles utilisés pour obtenir ces quantités. Nous déclinons la technique de moyennage de deux façons que nous appellerons *pondération interne* et *pondération externe*.

Ainsi, nous voulons comparer les résultats obtenus à partir des estimateurs pondérés *internes* et *externes* avec ceux provenant de chacun des modèles de l'ensemble proposé, ainsi qu'avec les résultats provenant d'autres estimateurs qui prennent en considération l'ensemble de modèles. Nous comparons les différents estimateurs sur la base du biais, la variance et l'erreur quadratique moyenne. De plus, nous nous intéressons à la performance des méthodes lorsque le vrai modèle fait partie, ou pas, des modèles considérés dans l'ensemble déterminé. Afin d'obtenir une compréhension élargie du comportement des estimateurs, nous présentons les résultats obtenus pour deux exemples simulés. Un premier plus simple afin d'étudier plus facilement comment les poids accordés aux modèles et les poids associés aux individus d'un échantillon interviennent dans les estimations obtenues avec les différents estimateurs. Un second exemple imite un cas qui se rapproche de la réalité en représentant des variables qui peuvent être observées dans la pratique. Dans ce deuxième exemple, nous rencontrons un problème différent de sélection de modèle : la sélection de modèle se fait principalement sur la manière de décrire les variables plutôt que sur la forme de leurs effets sur le traitement et la variable résultat.

Le premier chapitre introduit les concepts de base que nous rencontrons en épidémiologie. Nous y expliquons également quelles sont les différences entre les études expérimentales et observationnelles, de même que les défis engendrés par ces dernières. Dans le deuxième chapitre, nous présentons différentes méthodes couramment utilisées pour réduire le biais dû aux variables confondantes lors de la construction d'une étude ou lors de l'analyse des données dans le but d'estimer un effet causal. Les estimateurs qui nous intéressent plus particulièrement pour ce travail sont présentés dans le troisième chapitre. Enfin, dans le quatrième chapitre, deux exemples servent à illustrer les performances des estimateurs et une section sur les poids aide à la compréhension des résultats observés.

CHAPITRE I

CONCEPTS DE BASE

En épidémiologie, le but d'une étude est souvent d'estimer l'effet d'une exposition à un certain facteur (par exemple, un traitement ou un contaminant) sur la santé d'un individu ou d'une population d'individus. Nous nous intéressons aux variables qui causent ou qui préviennent une maladie ou un décès. Nous définissons une cause comme étant un événement, une condition ou une caractéristique nécessaire à l'occurrence de la maladie ou décès au moment et dans les conditions dans lesquels elle/il est survenu(e) (Kenneth et Greenland, 1998).

1.1 Effet causal et mesure d'association

Considérons le contexte médical suivant dans lequel la variable réponse Y représente la présence de la maladie si $Y = 1$ et son absence si $Y = 0$. D'autres variables sont considérées, soient X où $X = 1$ représente la prise ou l'exposition à un traitement ou un contaminant et $X = 0$ l'absence d'exposition, et les variables C pour représenter les autres caractéristiques mesurées chez les individus d'intérêt. Posons $Y_{x=1}$ la variable réponse qui aurait été observée chez un individu sous l'exposition ($x = 1$) et $Y_{x=0}$ la variable réponse qui aurait été observée en absence d'exposition ($x = 0$). Ces variables sont des résultats potentiels. Nous disons que l'exposition a un effet causal si $Y_{x=0} \neq Y_{x=1}$. Cependant, nous ne pouvons habituellement pas observer plus d'un résultat parce que cela nécessiterait de pouvoir reproduire des conditions totalement semblables pour un individu sous l'exposition et sans l'exposition. C'est pourquoi on ne peut généralement

pas évaluer l'effet causal chez un individu. Cette approche, où un des cas d'exposition est réalisé, mais l'autre non, est appelée l'approche *contre-factuelle* (Hernan, 2004).

Dans une population, nous disons que l'exposition a un effet causal si $P[Y_{x=1} = 1] \neq P[Y_{x=0} = 1]$, où $P[Y_x = 1]$ est la proportion d'individus qui auraient développé le résultat $Y = 1$ si l'ensemble des individus avaient reçu l'exposition x. En d'autres termes, l'exposition a un effet causal si la proportion d'individus ayant développé la maladie dans le cas où tous les individus auraient été exposés au traitement est différente de la proportion d'individus ayant développé la maladie dans le cas où tous les individus n'auraient pas été exposés au traitement.

Nous sommes donc intéressés à connaître le lien causal qui existe entre la variable d'exposition X et le résultat Y. L'association observée entre la variable d'intérêt et le résultat permet d'estimer cette causalité dans certaines situations. Pour ce faire, les chercheurs comparent un groupe d'individus qui ont été exposés au traitement à d'autres individus qui n'y ont pas été. Quelques mesures d'association sont :

$$Pr[Y = 1|X = 1] - Pr[Y = 1|X = 0],$$

$$Pr[Y = 1|X = 1]/Pr[Y = 1|X = 0],$$

$$(Pr[Y = 1|X = 1]/Pr[Y = 0|X = 1])/(Pr[Y = 1|X = 0]/Pr[Y = 0|X = 0]),$$

respectivement la différence de risque, le rapport de risque (aussi appelé risque relatif) et le rapport de cotes. Quand X et Y ne sont pas associés, on dit que X ne prédit pas Y ou encore que X et Y sont indépendants et on le note par : $X \coprod Y$.

1.2 Variables confondantes

Tel que mentionné en Section 1.1, le but est d'isoler l'effet de la variable d'intérêt X sur la réponse Y. Cependant, une maladie a souvent plus d'une cause, certaines que nous connaissons et d'autres que nous ignorons. Ces autres causes, si elles ne sont pas distribuées de manière homogène à travers les groupes observés peuvent brouiller les estimations du lien causal. Ces variables, qui sont à la fois associées à l'exposition et

qui sont des causes ou marqueurs de causes de la maladie sont appelées des *variables confondantes* (Rothman, 2002). Plus spécifiquement, une variable confondante doit être une cause à la fois de la variable résultat et du traitement. Elle doit donc apparaître avant l'événement résultat et ne doit pas être située dans le chemin causal menant du traitement à la variable résultat. Cette relation est souvent représentée à l'aide d'un graphe orienté acyclique (voir ci-bas, Figure 1.1), où l'orientation d'une flèche indique la cause, à la queue de la flèche, et le résultat, à la tête de la flèche (Greenland, Pearl et Robins, 1999). Les variables confondantes peuvent induire un biais important dans l'estimation de l'effet causal de l'exposition au traitement/contaminant étudié.

Voici une façon de comprendre la problématique des variables confondantes à l'aide d'une approche contre-factuelle, tirée de l'article « Confounding and Collapsibility in Causal Inference » (Greenland, Robins et Pearl, 1999). Supposons que le but de notre expérience est d'évaluer l'effet d'un traitement sur un paramètre μ dans la population A. Nous aimerions pouvoir comparer l'effet du traitement $x = 1$ à l'absence de traitement (ou à un placebo) $x = 0$ sur le paramètre μ dans la population A. Théoriquement, l'effet exact peut se mesurer par $\mu_{A1} - \mu_{A0}$ ou μ_{A1}/μ_{A0} si $\mu_{A0} \neq 0$. Nous ne pouvons cependant qu'observer la population sous la condition $x = 1$. Nous avons donc recours à un deuxième groupe, B, que nous espérons très semblable à A auquel nous assignons la condition $x = 0$ que A n'a pas reçu. Nous nous en servons alors pour comparer μ_{A1} à μ_{B0}. Les deux groupes sont semblables si $\mu_{B0} = \mu_{A0}$ et nous disons qu'ils sont *échangeables*.

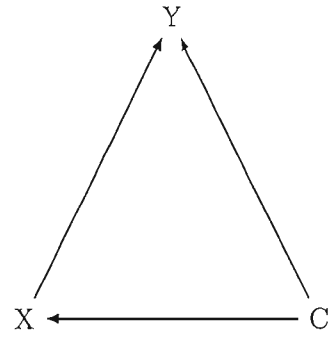


Figure 1.1 Graphe orienté acyclique.

Si, contre nos attentes, le groupe B diffère de A de telle sorte que $\mu_{B0} \neq \mu_{A0}$, nous disons alors qu'il y a présence de facteurs confondants. La mesure d'effet qui utilise μ_{B0} à la place de μ_{A0} est donc confondue. La différence $\mu_{A1} - \mu_{B0}$ représente alors un mélange de l'effet réel du traitement et de la différence entre les deux groupes A et B sous le traitement $x = 0$ (le biais) :

$$\mu_{A1} - \mu_{B0} = (\mu_{A1} - \mu_{A0}) + (\mu_{A0} - \mu_{B0}).$$

Avec un exemple numérique, voyons quel genre d'erreur peut engendrer l'omission d'une variable confondante. Considérons le cas où nous observons l'effet d'un traitement X sur la variable Y ($Y = 1$ indique la présence d'une maladie ; $Y = 0$ l'absence de maladie), et ce, en présence d'une variable confondante C dichotomique. Nous observons 1000 individus qui prennent le traitement ($X = 1$) et 1000 individus qui ne le prennent pas ($X = 0$). Le Tableau 1.1 résume ce qui est observé.

Si la variable confondante C est omise, nous calculons alors le rapport de risque en ne tenant compte que du traitement et nous obtenons la valeur suivante :

$$\begin{aligned} Pr[Y = 1|X = 1]/Pr[Y = 1|X = 0] \\ = \frac{440/1000}{620/1000} \\ \approx 0{,}7097\,. \end{aligned}$$

Notre conclusion est donc que la prise du traitement diminue le risque de maladie.

Tableau 1.1 Données observées pour un traitement X, une variable résultat Y et une variable confondante C

	X=0		X=1		
	Y=0	Y=1	Y=0	Y=1	
C=0	60	540	80	120	800
C=1	320	80	480	320	1200
	380	620	560	440	2000

Voyons maintenant l'effet de l'inclusion de la variable confondante C dans le calcul du rapport de risque. Supposons que la variable C représente un indice de l'importance qu'un individu accorde à sa santé ($C = 1$ indique que l'individu est très soucieux ; $C = 0$ que l'individu est peu soucieux). Dans ce contexte, nous pouvons imaginer qu'un individu plus soucieux de sa santé est généralement en meilleure santé et a moins de chance de développer la maladie. De plus, un individu soucieux de sa santé est également plus enclin à prendre un traitement préventif X qu'un individu peu soucieux de sa santé. En connaissant cette information, nous pouvons l'inclure dans notre analyse et alors obtenir un rapport de risque qui nous amène à une conclusion différente :

$$\frac{Pr[Y=1|X=1,C=0]\cdot Pr[C=0] + Pr[Y=1|X=1,C=1]\cdot Pr[C=1]}{Pr[Y=1|X=0,C=0]\cdot Pr[C=0] + Pr[Y=1|X=0,C=1]\cdot Pr[C=1]}$$

$$= \frac{\frac{120}{200}\cdot\frac{800}{2000} + \frac{320}{800}\cdot\frac{1200}{2000}}{\frac{540}{600}\cdot\frac{800}{2000} + \frac{80}{400}\cdot\frac{1200}{2000}}$$

$$= 1\,.$$

Dans ce cas, nous ne pouvons plus dire que le traitement a un effet protecteur puisque le risque de maladie est le même chez les individus qui ont été traités et ceux qui n'ont pas été traités. L'effet observé précédemment n'était pas dû au traitement lui-même, mais plutôt au fait que les individus non traités sont moins soucieux de leur santé.

1.3 Études expérimentales vs études d'observation

La présence potentielle de variables confondantes fait partie intégrante de toutes études épidémiologiques. Nous devons donc examiner en détail les moyens qui existent pour contrôler ces facteurs. Le contrôle des variables confondantes s'effectue soit à l'étape de la construction de l'étude, soit à l'étape de l'analyse des données recueillies. L'approche idéale en épidémiologie pour éliminer (ou du moins, réduire) le biais occasionné par ces variables est de réaliser des études expérimentales, c'est-à-dire des études où l'attribution des individus aux différents groupes exposé et non exposé est aléatoire.

Avec un nombre de sujets suffisamment grand, nous pouvons alors présumer que les variables confondantes sont balancées à travers les groupes. L'avantage de ce type d'étude est que non seulement les variables confondantes connues et mesurées sont balancées, mais également celles que nous ne mesurons ou ne connaissons pas. Cependant, rares et souvent plus coûteuses sont les occasions que les chercheurs ont de réaliser de telles études. Nous pouvons penser par exemple qu'il est tout à fait non éthique de provoquer l'exposition d'une personne à ce que nous croyons être un contaminant pour voir l'effet de ce dernier. Une solution de remplacement aux études expérimentales est les études d'observation. Dans ce type d'étude, nous observons des sujets qui s'exposent ou non à la variable d'intérêt, mais ce, de leur propre gré. Il est alors fort probable que les gens exposés soient différents des gens non exposés par leurs autres caractéristiques qui, si elles sont des causes de la maladie, sont alors des variables confondantes. Pensons par exemple à une étude dans laquelle nous désirons connaître l'effet de la consommation de café sur une maladie. Si un petit nombre d'heures de sommeil augmente la probabilité de souffrir de cette maladie, nous pouvons nous attendre à ce que la variable « nombre d'heures de sommeil » soit une variable confondante, car il y a des chances que les gens qui dorment moins consomment davantage de café. Il est possible d'atténuer le biais engendré par les variables confondantes dans les études d'observation. Les méthodes les plus couramment utilisées sont décrites dans le Chapitre 2. Mentionnons toutefois que dans le contexte des études d'observation, le contrôle des facteurs confondants ne peut se faire que sur ceux qui sont mesurés. Ceci implique qu'il est tout à fait possible que l'effet de la variable d'exposition soit encore confondu même après l'application de ces techniques. Ceci est la raison pour laquelle les études expérimentales sont souvent considérées comme étant supérieures aux études d'observation. En effet, dans les études expérimentales, nous pouvons faire la supposition que les facteurs qui ne sont pas mesurés sont distribués de manière similaire à travers les différents groupes d'exposition car ces derniers sont créés de manière aléatoire. L'effet des facteurs confondants n'entraîne donc pas une différence entre les individus assignés au traitement et ceux qui n'y sont pas.

CHAPITRE II

TECHNIQUES CLASSIQUES UTILISÉES EN ÉTUDES D'OBSERVATION

Tel que mentionné précédemment, nous examinons dans ce chapitre les méthodes disponibles pour réduire le biais dû aux variables confondantes dans le cadre d'études d'observation. À cette fin, nous voyons qu'il est possible d'intervenir lors de la construction de l'étude ou bien lors de l'analyse des données.

2.1 Lors de la construction de l'étude

Nous pouvons tout d'abord réduire le biais dû aux variables confondantes par la façon dont nous choisissons les groupes observés. Le but quand nous construisons nos populations est de les rendre échangeables. Deux techniques souvent utilisées sont la restriction et l'appariement.

2.1.1 Restriction

Premièrement, nous pouvons restreindre l'étude à un certain groupe d'individus. Par exemple, si nous présumons que le lieu de résidence est une variable confondante, nous pouvons réaliser l'étude avec des gens qui vivent tous dans un même quartier. Ainsi, comme il n'y a pas de variation, la variable « lieu de résidence » répartie uniformément à travers l'échantillon ne peut plus confondre l'effet ciblé. Les conclusions faites suite à cette restriction sont par contre elles aussi restreintes quant à la population à laquelle

elles se rapportent. De plus, il se peut que pour faire l'expérience avec un nombre d'individus suffisamment grand, le fait de se restreindre à une certaine tranche de la population rende la tâche de trouver les sujets plus longue ou difficile.

2.1.2 Appariement

Une façon de contourner les inconvénients qui viennent avec la restriction est de plutôt apparier les individus qui ont des caractéristiques semblables de façon à balancer les groupes. Ainsi, si nous nous référons au même exemple que plus haut, plutôt que d'observer un seul quartier, il suffit d'avoir les mêmes proportions de sujets qui viennent d'un quartier dans les deux groupes, exposé et non exposé. Encore ici, nous pouvons avoir de la difficulté à trouver assez de gens si nous désirons apparier les sujets sur plusieurs variables. De plus, s'il y a une perte de suivi pour certains individus, les groupes peuvent être alors débalancés.

2.2 Lors de l'analyse de données

Après que les données aient été récoltées, nous pouvons encore réduire l'effet de confusion dû aux variables confondantes mesurées. Voyons dans cette sous-section quelques-unes des techniques les plus fréquemment utilisées.

2.2.1 Stratification

La *stratification* nécessite de regrouper les sujets selon leur valeur observée pour la variable confondante C. Par exemple, nous pouvons regrouper les sujets qui viennent d'un même quartier. Nous pouvons aussi transformer une variable continue comme l'âge en une variable catégorique avec différentes classes d'âge et assigner chacun des individus à l'une de celles-ci. Nous évaluons ensuite l'effet causal à l'intérieur de chacune des classes, puis le risque global s'obtient à partir d'une moyenne pondérée des risques calculés sur l'ensemble des classes. Pour une variable résultat binaire, nous pouvons

donc obtenir le rapport de risque causal par standardisation de la façon suivante :

$$\frac{Pr[Y_{x=1}=1]}{Pr[Y_{x=0}=1]} = \frac{\sum_c Pr[Y_{x=1}=1|C=c]Pr[C=c]}{\sum_c Pr[Y_{x=0}=1|C=c]Pr[C=c]}. \tag{2.1}$$

Pour estimer le rapport de risque causal, nous estimons les probabilités des résultats contre-factuels par les probabilités observées dans l'étude d'observation, car nous présumons qu'après standardisation les groupes exposé et non exposé sont échangeables à l'intérieur de chacune des classes (i.e. ils sont échangeables conditionnellement à C) :

$$\frac{\sum_c \hat{Pr}[Y=1|C=c,X=1]\hat{Pr}[C=c]}{\sum_c \hat{Pr}[Y=1|C=c,X=0]\hat{Pr}[C=c]}.$$

Dans le cas où Y est linéaire plutôt que dichotomique, nous calculons la moyenne de la variable Y, $\hat{\mu}$, à l'intérieur de chacune des classes pour les individus exposés et non exposés et la différence de risque s'estime alors comme suit :

$$\sum_c \hat{\mu}_{C=c,X=1}\hat{Pr}[C=c] - \sum_c \hat{\mu}_{C=c,X=0}\hat{Pr}[C=c].$$

Ceci est valide dans la mesure où nous supposons également qu'il n'y a pas de variables confondantes non mesurées. Nous réalisons que le nombre de classes augmente rapidement lorsque la stratification est effectuée sur plusieurs variables. En fait, l'augmentation est exponentielle, car si nous voulons stratifier p variables en 2 classes chacune, nous avons alors 2^p classes. Il est important de toujours avoir au moins un individu exposé et non exposé pour chacune des classes. Cette technique est donc applicable tant que nous n'avons pas trop de variables qui nécessitent d'être contrôlées ou tant que nous n'avons pas besoin d'un raffinement trop fin si nous catégorisons une variable continue. En effet, nous pourrions nous retrouver avec trop peu de données par groupe ou pire encore, avec aucune donnée. Lorsqu'il y a plusieurs variables confondantes à considérer dans l'analyse, la stratification sur chacune d'elles (ou l'appariement) s'avère souvent trop complexe. Il est courant d'avoir alors recours à la méthode de régression et les méthodes basées sur ce que nous appelons le *score de propension.*

2.2.2 Régression

Cette technique est probablement la plus couramment utilisée pour estimer l'effet de chacune des variables et plus particulièrement l'effet de l'exposition. Nous nous servons souvent des régressions linéaires ou logistiques pour modéliser la relation entre la variable résultat et l'exposition et les covariables mesurées dans l'étude. En supposant que nous ayons une variable résultat Y, une exposition X et deux covariables C_1 et C_2, voyons ce que nous indique la régression.

Si Y est continue, comme ce serait le cas si elle représentait l'âge au décès par exemple, nous pouvons utiliser la régression linéaire. Si nous ne considérons pas d'interaction entre les variables, nous pouvons alors estimer la relation $E[Y|X,C_1,C_2] = \alpha + \beta_x X + \beta_1 C_1 + \beta_2 C_2$ par $\hat{E}[Y|X = x,C_1 = c_1,C_2 = c_2] = \hat{\alpha} + \hat{\beta}_x x + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2$. Les coefficients, β, représentent l'effet associé à une augmentation d'une unité de la variable à laquelle il est rattaché lorsque les autres variables sont fixées. Par exemple, pour β_x :

$$\begin{aligned} &E[Y|X = x+1,C_1 = c_1,C_2 = c_2] - E[Y|X = x,C_1 = c_1,C_2 = c_2] \\ &= \alpha + \beta_x(x+1) + \beta_1 c_1 + \beta_2 c_2 - \alpha - \beta_x x - \beta_1 c_1 - \beta_2 c_2 \\ &= \beta_x \cdot 1 \end{aligned}$$

Dans le cas où l'exposition est représentée par une variable dichotomique (0 pour non exposé ou 1 pour exposé), β_x exprime alors directement l'effet de l'exposition, car c'est une unité qui différencie les deux groupes.

Un avantage de la régression linéaire est que nous pouvons inclure un grand nombre de variables dans le modèle considéré. La principale difficulté qui se pose lorsque nous voulons faire une régression est de décider de la forme du modèle. Les paramètres estimés sont convergents seulement lorsque nous faisons la supposition que le modèle utilisé est le bon, ce qui n'est jamais une certitude en pratique. Premièrement, nous devons faire un choix quant aux variables qui sont incluses dans le modèle. Il faut idéalement que toutes les variables qui confondent l'effet de l'exposition sur la variable résultat soient mesurées et considérées dans le modèle pour qu'il soit possible de les contrôler

et ainsi ajuster l'effet du traitement. En effet, le paramètre β_x ne doit être interprété causalement que si toutes les variables confondantes sont considérées dans le modèle. De plus, la relation entre les variables explicatives et la variable dépendante pourrait être linéaire, quadratique ou autre. La sélection du modèle est donc de première importance pour l'estimation des paramètres.

Un problème se pose si Y n'est pas une variable continue, mais plutôt une variable binaire (0 ou 1) puisque la régression linéaire pourrait faire correspondre à $Pr[Y = 1|X,C_1,C_2]$ des valeurs plus petites que 0 ou plus grandes que 1. Une solution à ce problème est d'utiliser la régression logistique. Dans ce cas, la relation entre Y, X, C_1 et C_2 s'exprime alors comme suit :

$$Pr[Y = 1|X, C_1, C_2] = \frac{\exp(\alpha + \beta_x X + \beta_1 C_1 + \beta_2 C_2)}{1 + \exp(\alpha + \beta_x X + \beta_1 C_1 + \beta_2 C_2)} . \tag{2.2}$$

Puisque l'exponentielle est strictement positive, le rapport (équation 2.2) est toujours compris entre 0 et 1.

La fonction logit nous permet d'établir une relation entre les variables explicatives et la variable résultat sous cette forme :

$$\ln \frac{Pr[Y = 1|X, C_1, C_2]}{Pr[Y = 0|X, C_1, C_2]} = \ln \frac{Pr[Y = 1|X, C_1, C_2]}{1 - Pr[Y = 1|X, C_1, C_2]} = \alpha + \beta_x X + \beta_1 C_1 + \beta_2 C_2 .$$

Voyons comment nous pouvons interpréter β_x :

$$\begin{aligned}
& \beta_x \cdot 1 \\
= \; & (\alpha + \beta_x \cdot 1 + \beta_1 c_1 + \beta_2 c_2) - (\alpha + \beta_x \cdot 0 + \beta_1 c_1 + \beta_2 c_2) \\
= \; & \ln \frac{Pr[Y = 1|X = 1, C]}{1 - Pr[Y = 1|X = 1, C]} - \ln \frac{Pr[Y = 1|X = 0, C]}{1 - Pr[Y = 1|X = 0, C]} \\
= \; & \ln \frac{Pr[Y = 1|X = 1, C]}{Pr[Y = 0|X = 1, C]} - \ln \frac{Pr[Y = 1|X = 0, C]}{Pr[Y = 0|X = 0, C]} ,
\end{aligned}$$

où C=$(C_1 = c_1, C_2 = c_2)$.

Le coefficient β_x représente donc la différence en logit provoquée par l'exposition.

2.2.3 Score de propension

Nous avons recours aux méthodes faisant intervenir le score de propension entre autres lorsqu'il nous semble plus facile de modéliser correctement le modèle ayant comme variable réponse le traitement X et comme variables explicatives les variables confondantes C que le modèle de Y en fonction de X et de C. Le score de propension se définit comme étant la probabilité d'être exposé conditionnellement aux autres caractéristiques mesurées : $Pr[X = 1|C]$. Cette probabilité est généralement estimée avec une régression logistique dans laquelle la variable réponse est l'exposition et les variables explicatives sont les variables confondantes et possiblement leurs transformations. Par la suite, nous pouvons nous servir du score de propension comme variable dans une régression, ou comme variable à partir de laquelle nous stratifions ou nous apparions. Le score de propension est un résumé à une seule dimension de l'information apportée par toutes les variables utilisées pour le construire. Dorénavant dans ce texte, $e(C)$ désignera le score de propension et $e(C,\hat{\alpha})$, où $\hat{\alpha}$ est le vecteur des coefficients obtenus par la régression logistique, son estimation. Rosenbaum et Rubin (1983) ont montré que pour une valeur du score de propension fixée, l'association observée à l'aide de ces méthodes d'analyse entre le traitement et la variable résultat est une estimation non biaisée de l'effet causal moyen du traitement pour cette même valeur de score de propension. Ce résultat est valide pour autant que le modèle pour le score de propension soit correctement spécifié. L'hypothèse nécessaire pour que l'estimation soit non biaisée est appelée *strongly ignorable treatment assignment* ; elle suppose que l'assignation au traitement et les résultats contre-factuels sont indépendants conditionnellement au vecteur de variables utilisé pour construire le score de propension, $Y_{0,1} \coprod X|C$ (Rosenbaum et Rubin, 1983).

Dans un contexte de stratification, nous pouvons, par exemple, séparer les individus en K strates définies par les quantiles des scores de propension. À l'intérieur de chacune des strates, les différentes variables ne sont pas nécessairement homogènes dans les groupes exposé et non exposé, mais nous nous attendons à ce que leurs distributions soient semblables. En d'autres termes, si nous considérons l'âge et le sexe comme va-

riables qui déterminent la probabilité d'être exposé, à l'intérieur de chacune des strates, nous n'avons pas nécessairement que des jeunes ou que des femmes, mais une même proportion de jeunes femmes dans les groupes exposé et non exposé. Nous pouvons ensuite estimer le rapport de risque causal par standardisation (équation 2.1), c'est-à-dire en comparant la variable résultat du groupe exposé à celle du groupe non exposé en pondérant l'effet obtenu à l'intérieur de chaque strate par la probabilité d'être dans une strate donnée. Il est suggéré d'utiliser $K = 5$ strates pour stratifier les observations, car c'est souvent suffisant pour enlever 90% du biais (Rosenbaum et Rubin, 1983; Cochran, 1968).

Nous pouvons également utiliser le score de propension pour apparier les individus. Plutôt que de les apparier à partir de toutes leurs caractéristiques, nous nous servons de la valeur de leur score de propension. Bien sûr, il est peu probable que deux individus aient exactement le même score. Une façon de procéder est la suivante. Pour un individu exposé ayant un score de propension s, nous repérons les individus non exposés ayant un score dans une fenêtre $\pm$ 0.1 de s. Nous sélectionnons alors au hasard l'un d'entre eux pour l'appariement tout en le retirant des individus non exposés encore disponibles pour l'appariement. Malheureusement, il est possible qu'il soit difficile d'apparier tous les cas puisque, surtout dans les probabilités extrêmes, il se peut qu'un des groupes (exposé ou non exposé) soit sous-représenté.

Comme mentionné précédemment, il est aussi possible d'utiliser le score de propension comme une variable explicative à l'intérieur de la régression. Sous la supposition que l'assignation au traitement et les résultats contre-factuels sont indépendants conditionnellement au vecteur de variables utilisé pour construire le score de propension, nous avons :

$$E[Y_x|X = x, e(C) = e(c)] = E[Y_x|e(C) = e(c)]\,.$$

Si nous supposons également que la relation entre le résultat et le score de propension est linéaire, c'est-à-dire

$$E[Y_x|e(C) = e(c)] = \alpha_x + \gamma_x e(c)\,,$$

alors la statistique

$$(\hat{\alpha}_{x=1} - \hat{\alpha}_{x=0}) + (\hat{\gamma}_{x=1} - \hat{\gamma}_{x=0})\overline{e(c)}\,,$$

où $\overline{e(c)} = \frac{1}{n}\sum e(c)$, est un estimateur sans biais de l'effet causal moyen du traitement (Rosenbaum et Rubin, 1983). Dans la régression, cette réduction du nombre de variables explicatives due à l'utilisation du score de propension, plutôt que du vecteur des variables confondantes, permet de conserver plus de degrés de liberté pour l'erreur résiduelle. Cela est donc particulièrement utile si nous désirons inclure un grand nombre de variables par rapport au nombre d'individus disponibles. Pour estimer l'effet causal du traitement, nous pouvons donc inclure le score de propension et quelques variables jugées plus importantes dans le modèle de régression, évitant ainsi une surparamétrisation (D'Agostino, 1998).

2.2.4 Pondération par les probabilités inversées

Le but de cette méthode est de simuler ce qui serait arrivé si toute la population avait été exposée au traitement et si elle avait également été non exposée. Nous créons donc une *pseudo*-population deux fois plus grande que la population originale où chaque sujet s'y retrouve à la fois comme étant exposé et non exposé. Ainsi, les facteurs confondants se retrouvent balancés entre les groupes exposé et non exposé. Cette pseudo-population est créée en pondérant chaque sujet par l'inverse de la probabilité conditionnelle de recevoir l'exposition qu'il a reçue $Pr^{-1}(X = x|C = c)$, c'est-à-dire par l'inverse de son score de propension $e^{-1}(C) = e^{-1}(c)$. Par exemple, si parmi 12 personnes qui vivent dans le quartier A, 2 reçoivent l'exposition, nous avons alors $Pr[X = 1|C = A] = 2/12$. Ainsi, nous accordons à chacun de ces sujets un poids de $12/2 = 6$ et la pseudo-population contient alors 12 personnes vivant dans le quartier A qui ont reçu l'exposition. De même, les 10 personnes qui n'ont pas été exposées dans la population originale se voient attribuées un poids de $12/10 = 1.2$, ce qui crée également 12 personnes non exposées vivant dans le quartier A dans la pseudo-population. Notons qu'ainsi davantage de poids est donné aux individus dont l'exposition observée x est plus rare conditionnellement à $C = c$. La différence de risque est estimée à partir de cette pseudo-population. Par

exemple, pour Y dichotomique, nous remplaçons $Pr[Y_{x=1} = 1]$ par la proportion de gens qui ont reçu l'exposition $x = 1$ et qui ont comme variable résultat $Y = 1$ parmi toute la pseudo-population. Similairement, nous remplaçons $Pr[Y_{x=0} = 1]$ par la proportion de gens qui ont reçu l'exposition $x = 0$ et qui ont aussi $Y = 1$ comme variable résultat parmi la pseudo-population entière. De façon plus générale, pour Y dichotomique ou non, nous estimons la différence de risque de la façon suivante :

$$\frac{1}{n}\sum_{i=1}^{n}\frac{X_iY_i}{e(C_i,\hat{\alpha})} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-X_i)Y_i}{1-e(C_i,\hat{\alpha})}.$$

2.2.5 Méthodes doublement robustes

Les méthodes vues ci-dessus requièrent la spécification correcte du modèle de prédiction du traitement ou du modèle de régression de la variable résultat Y. Toutefois, il n'y a pas de façon d'être absolument certain que le modèle utilisé est le bon et donc l'estimation que nous en tirons peut s'avérer biaisée. D'autres méthodes, que nous disons *doublement robustes*, existent et possèdent une propriété très intéressante. Les méthodes doublement robustes combinent l'estimation par régression et l'estimation par pondération par les probabilités inversées. Avec cette approche, nous estimons l'effet causal moyen par :

$$\frac{1}{n}\sum_{i=1}^{n}\frac{X_iY_i - (X_i - e(C_i,\hat{\alpha}))m_1(C_i,\hat{\beta}_1)}{e(C_i,\hat{\alpha})} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-X_i)Y_i + (X_i - e(C_i,\hat{\alpha}))m_0(C_i,\hat{\beta}_0)}{1-e(C_i,\hat{\alpha})},$$

où $m_x(C_i,\beta_x) = E[Y|X = x,C] = E[Y_x|C]$, $x = 0,1$ (Lunceford et Davidian, 2004). La première partie à l'intérieur des sommes est similaire à ce que nous trouvons dans l'estimateur par les probabilités inversées, mais à cela est ajouté un autre terme qui tient compte des modèles de régression. L'avantage principal de cet estimateur est qu'il demeure non biaisé dans les trois cas suivants :

(a) Les trois modèles sont correctement spécifiés ;

(b) Le modèle pour le score de propension est correctement spécifié, mais un ou les deux modèles $m_x(C,\beta_x)$ ne le sont pas ;

(c) Les deux modèles $m_x(C,\beta_x)$ sont correctement spécifiés, mais le modèle de propension ne l'est pas,

ce qui offre une certaine protection contre l'erreur. Pour voir pourquoi il en est ainsi, trouvons l'espérance de l'estimateur de μ_1 :

$$
\begin{aligned}
&E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{X_iY_i-(X_i-e(C_i,\hat{\alpha}))m_1(C_i,\hat{\beta}_1)}{e(C_i,\hat{\alpha})}\right]\\
&=\frac{1}{n}\sum_{i=1}^{n}E\left[\frac{X_iY_i-(X_i-e(C_i,\hat{\alpha}))m_1(C_i,\hat{\beta}_1)}{e(C_i,\hat{\alpha})}\right]\\
&=\frac{1}{n}\sum_{i=1}^{n}E\left[\frac{X_i(X_iY_{1i}+(1-X_i)Y_{0i})-(X_i-e(C_i,\hat{\alpha}))m_1(C_i,\hat{\beta}_1)}{e(C_i,\hat{\alpha})}\right](i)\\
&=\frac{1}{n}\sum_{i=1}^{n}E\left[\frac{X_i^2Y_{1i}+X_i(1-X_i)Y_{0i}-(X_i-e(C_i,\hat{\alpha}))m_1(C_i,\hat{\beta}_1)}{e(C_i,\hat{\alpha})}\right]\\
&=\frac{1}{n}\sum_{i=1}^{n}E\left[\frac{X_iY_{1i}-(X_i-e(C_i,\hat{\alpha}))m_1(C_i,\hat{\beta}_1)}{e(C_i,\hat{\alpha})}\right](ii)\\
&=\frac{1}{n}\sum_{i=1}^{n}E\left[\frac{X_iY_{1i}-(X_i-e(C_i,\hat{\alpha}))(m_1(C_i,\hat{\beta}_1)-Y_{1i})-X_iY_{1i}+e(C_i,\hat{\alpha})Y_{1i}}{e(C_i,\hat{\alpha})}\right]\\
&=\frac{1}{n}\sum_{i=1}^{n}E\left[Y_{1i}-\frac{(X_i-e(C_i,\hat{\alpha}))(m_1(C_i,\hat{\beta}_1)-Y_{1i})}{e(C_i,\hat{\alpha})}\right]\\
&=\frac{1}{n}\sum_{i=1}^{n}E[Y_{1i}]-\frac{1}{n}\sum_{i=1}^{n}E\left[\frac{(X_i-e(C_i,\hat{\alpha}))(m_1(C_i,\hat{\beta}_1)-Y_{1i})}{e(C_i,\hat{\alpha})}\right]\\
&=\mu_1-\frac{1}{n}\sum_{i=1}^{n}E\left[\frac{(X_i-e(C_i,\hat{\alpha}))(m_1(C_i,\hat{\beta}_1)-Y_{1i})}{e(C_i,\hat{\alpha})}\right];
\end{aligned}
$$

(i) $Y_i = X_iY_{1i} + (1 - X_i)Y_{0i}$, car selon que l'individu est exposé ou non, Y_i observé sera égal au résultat contre-factuel Y_{1i} ou Y_{0i} ;

(ii) car X étant égal soit à 0, soit à 1, $X_i^2 = X$ et $X_i(1 - X_i) = 0$.

(Davidian, 2007)

Nous trouvons donc que l'estimateur doublement robuste de μ_1 est non biaisé si

$$E\left[\frac{(X_i-e(C_i,\hat{\alpha}))(m_1(C_i,\hat{\beta}_1)-Y_{1i})}{e(C_i,\hat{\alpha})}\right]=0.$$

Cela est précisément le cas si le modèle pour le score de propension ou le modèle de régression pour Y_1 sont correctement spécifiés. De manière similaire, nous pouvons montrer que l'espérance de la seconde partie de l'estimateur est μ_0 si le modèle de propension ou le modèle de régression sont correctement spécifiés. Suivant ces résultats, nous avons

bien que cet estimateur doublement robuste peut nous donner une bonne estimation de l'effet causal. Cependant, dans le cas où ni le modèle pour le score de propension et ni les modèles pour $Y_{0,1}$ ne sont corrects, l'estimateur doublement robuste ne performe pas nécessairement mieux que la régression ou la pondération par probabilités inversées, séparément (Kang et Schafer, 2007). Aussi, il est bien connu que, dans le cas où la modélisation du modèle de régression pour Y est correcte, la variance des estimateurs résultant de l'utilisation de la méthode doublement robuste est plus large qu'avec l'utilisation d'une simple méthode de régression (Bang et Robins, 2005).

CHAPITRE III

MÉTHODOLOGIE

Parmi les techniques d'estimation vues au chapitre précédent, celle qui nous intéresse plus particulièrement dans le cadre de ce mémoire est l'estimation par pondération par les probabilités inversées (section 2.2.4). Comme il a déjà été mentionné, cette méthode requiert que nous spécifiions un modèle de prédiction pour le traitement, modèle que nous utilisons pour estimer le score de propension de chacun des individus. Il n'est toutefois pas facile de spécifier correctement le modèle de traitement et les estimateurs sont particulièrement sensibles au choix de ce modèle. Afin de spécifier le modèle, nous espérons avoir une assez bonne connaissance des variables qui sont reliées à la variable résultat ainsi qu'au processus d'exposition au traitement. En pratique, la connaissance des relations entre les variables est souvent imparfaite et les méthodes de sélection de variables par étapes sont souvent utilisées en complément pour spécifier le modèle de traitement (Weitzen et al., 2004). Le critère d'information d'Akaike (AIC) et le critère d'information bayésien (BIC) sont eux aussi couramment utilisés pour faire de la sélection de modèle. Ces critères sont basés sur un calcul incorporant la vraisemblance maximale et le nombre de paramètres du modèle. Ce sont des mesures qui indiquent la qualité de l'ajustement d'un modèle aux données et nous nous en servons pour comparer les modèles. Le AIC et le BIC se calculent ainsi :

$$\begin{aligned} AIC &= 2k - 2\ln(L)\,; \\ BIC &= k\ln(n) - 2\ln(L)\,, \end{aligned}$$

où L est la vraisemblance maximale, k le nombre de paramètres et n la taille de l'échantillon. Le modèle présentant le plus petit AIC ou BIC est généralement préféré aux autres. Notons que puisque le nombre de paramètres multiplie le logarithme du nombre d'individus dans le calcul du BIC, ce dernier est plus pénalisant que le AIC envers les modèles qui ont plusieurs paramètres (aussitôt que $n \geq 8$).

Souvent, suite à la sélection d'un modèle, les analyses statistiques sont faites et des conclusions tirées sans tenir compte de l'incertitude liée à la sélection de ce modèle. Si, contre nos attentes, le modèle choisi n'est pas le bon, l'estimateur n'est plus sans biais et les conclusions qui en découlent sont invalides. Nous nous intéressons donc plutôt ici à une façon de faire qui incorpore dans les analyses un ensemble de modèles potentiels. L'approche que nous privilégions plus particulièrement utilise une moyenne pondérée des scores de propension calculés sous chaque modèle considéré pour réaliser une seule estimation de l'effet causal à l'aide de la méthode d'estimation par pondération par les probabilités inversées.

3.1 Estimations présentées

Pour illustrer plus clairement en quoi consistent les estimations que nous présentons, commençons tout d'abord par définir certaines notations qui sont employées :

- $i = 1, \ldots, n$: les individus d'un jeu de données ;
- $j = 1, \ldots, m$: les modèles faisant partie de l'ensemble des modèles considérés ;
- $k = 1, \ldots, ech$: les jeux de données qui sont générés ;
- $X_{i,k}$: variable dichotomique représentant la présence de l'exposition étudiée ($X = 0$: absence de l'exposition ; $X = 1$: présence de l'exposition) pour l'individu i dans le jeu de données k ;
- $Y_{i,k}$: la variable résultat de l'individu i dans le jeu de données k ;
- $e(C_{i,k}, \hat{\alpha}_{j,k})$: le score de propension estimé pour l'individu i avec le modèle j dans le jeu de données k.

3.1.1 Estimation sous chaque modèle

Pour chacun des modèles qui font partie de l'ensemble des modèles qui sont considérés, nous calculons l'estimation de l'effet causal par pondération par les probabilités inversées qui en découle. Pour un jeu de données k, nous avons donc autant d'estimations que de modèles suggérés, soient $\hat{\theta}_{1,k}$, ..., $\hat{\theta}_{m,k}$. Nous calculons chacune d'elles à partir de l'équation suivante :

$$\hat{\theta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_{i,k}Y_{i,k}}{e(C_{i,k},\hat{\alpha}_{j,k})} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-X_{i,k})Y_{i,k}}{1-e(C_{i,k},\hat{\alpha}_{j,k})}. \tag{3.1}$$

Sous chaque modèle, les valeurs de $X_{i,k}$ et $Y_{i,k}$ demeurent les mêmes pour un individu puisque le choix du modèle intervient seulement dans le calcul du score de propension. Puisque c'est ce score qui sert à la pondération pour chaque individu, un modèle de traitement incorrect peut engendrer de trop grands ou trop petits poids pour certains individus et ainsi biaiser l'estimation. Les individus qui ont une forte propension à recevoir l'exposition contraire à celle qu'ils ont véritablement reçue sont ceux qui ont une plus grande influence. Par exemple, un individu qui a été exposé ($X_{i,k} = 1$), mais dont le score de propension estimé par un certain modèle j_1 est $e(C_{i,k},\hat{\alpha}_{j_1,k}) = 0{,}05$, a un poids de $1/0{,}05 = 20$ dans le calcul de l'estimation. Sous un modèle différent, j_2, ce même individu peut présenter un score de propension estimé de $e(C_{i,k},\hat{\alpha}_{j_2,k}) = 0{,}02$, ce qui n'est pas si loin de l'estimation obtenue du premier modèle. Cependant, avec 0,02 plutôt que 0,05 comme score de propension, son poids augmente de plus que le double ($1/0{,}02 = 50$), ce qui peut grandement influencer le résultat de l'estimation. C'est pourquoi nous disons que l'estimateur est très sensible au choix du modèle.

3.1.2 Estimation sous le modèle ayant le plus petit AIC ou BIC

La première forme d'estimation impliquant plus d'un modèle que nous présentons est l'estimation obtenue du modèle ayant la plus petite valeur de AIC et celle obtenue du modèle ayant le plus petit BIC. Comme il a été mentionné précédemment, ces mesures servent souvent à déterminer de quel modèle il semble le plus probable que les données

proviennent. Nous reprenons alors les m modèles faisant partie de l'ensemble des modèles considérés et calculons pour chacun d'eux les valeurs de AIC et BIC qui y sont associées ($AIC_{j,k}$ et $BIC_{j,k}$). Nous gardons finalement l'estimation provenant du modèle le plus probable selon le AIC, puis selon le BIC, c'est-à-dire les modèles ayant les plus petites valeurs. Ces deux estimations, pour le jeu de données k, sont notées $\hat{\theta}_{AIC,k}$ et $\hat{\theta}_{BIC,k}$ respectivement.

3.1.3 Estimation par la pondération externe

Nous considérons ensuite dans ce travail deux façons de pondérer nos estimations en fonction des modèles. La première, que nous appelons « pondération externe », consiste en un moyennage des m estimations obtenues des m différents modèles considérés. Nous déclinons cette méthode d'estimation en deux différents résultats selon que nous utilisons le critère AIC ou BIC pour la pondération. Pour un jeu de données k, les notations utilisées pour représenter ces deux estimations sont $\hat{\theta}_{ext,AIC_k}$ et $\hat{\theta}_{ext,BIC_k}$ respectivement. Nous utilisons les valeurs de AIC et BIC pour évaluer la plausibilité d'un certain modèle parmi l'ensemble de tous les modèles considérés. Nous assignons à chacun des modèles une probabilité qui dépend de la valeur du critère qui y est associée et de la valeur de ce critère à travers l'ensemble de tous les modèles. C'est cette probabilité que nous employons ensuite comme un poids. Pour calculer cette probabilité nous devons tout d'abord définir $\Delta_{AIC_{j,k}} = AIC_{j,k} - AIC_{min,k}$, où $AIC_{min,k}$ est le plus petit AIC parmi tous les modèles proposés, et ce, pour le jeu de données k. Le poids que nous associons à chaque modèle, pour le jeu de données k, se calcule à partir de l'équation suivante (Burnham et Anderson, 2004) :

$$p_{AIC_{j,k}} = \frac{\exp(-\Delta_{AIC_{j,k}}/2)}{\sum_{j=1}^{m} \exp(-\Delta_{AIC_{j,k}}/2)}. \tag{3.2}$$

De façon tout à fait similaire, nous déterminons les poids $p_{BIC_{j,k}}$ en fonction du critère BIC :

$$p_{BIC_{j,k}} = \frac{\exp(-\Delta_{BIC_{j,k}}/2)}{\sum_{j=1}^{m} \exp(-\Delta_{BIC_{j,k}}/2)}, \tag{3.3}$$

où $\Delta_{BIC_{j,k}} = BIC_{j,k} - BIC_{min,k}$.

Remarquons qu'une plus petite valeur du critère, comparativement aux autres, donne un plus grand poids et que la somme des poids, sous chacun des critères, pour un jeu de données particulier, égale un.

L'estimation par pondération externe reprend les m estimations décrites plus haut (équation 3.1), dans la sous-section 3.1.1, et le poids associé à chacun des modèles (équations 3.2 et 3.3), les combine et en fait deux nouvelles estimations :

$$\hat{\theta}_{ext,AIC_k} = \sum_{j=1}^{m} p_{AIC_{j,k}} \cdot \hat{\theta}_{j,k}\,, \tag{3.4}$$

si nous utilisons le AIC comme mesure d'ajustement, ou

$$\hat{\theta}_{ext,BIC_k} = \sum_{j=1}^{m} p_{BIC_{j,k}} \cdot \hat{\theta}_{j,k}\,, \tag{3.5}$$

si nous utilisons plutôt le BIC.

3.1.4 Estimation par la pondération interne

L'estimation par « pondération interne » se base aussi sur un moyennage à travers les différents modèles. Cependant, plutôt que de pondérer les estimations des m modèles, nous pondérons les scores de propension utilisés à l'intérieur de l'estimateur par pondération par les probabilités inversées (équation 3.1). Ici, nous avons donc une seule estimation de l'effet causal basée sur les m modèles simultanément. Nous obtenons également avec cette méthode deux résultats selon que nous utilisons le AIC ou le BIC et nous les désignons par les notations $\hat{\theta}_{int,AIC_k}$ et $\hat{\theta}_{int,BIC_k}$, pour un jeu de données k.

L'estimation par pondération interne s'effectue comme suit. Sous chacun des modèles sélectionnés, nous estimons le score de propension de chacun des individus. Nous avons donc, dans le jeu de données k, $e(C_{i,k},\hat{\alpha}_1)$, ..., $e(C_{i,k},\hat{\alpha}_m)$, pour $i = 1,\ldots,n$. Nous calculons également, de la même façon que précédemment, le poids accordé à chaque modèle, soit à partir du critère AIC, soit à partir du critère BIC (équations 3.2 et 3.3).

Pour le moment, définissons l'estimateur en fonction du AIC. Pour chaque individu, nous procédons à un moyennage de ses m scores de propension en les pondérant chacun par la probabilité associée au modèle duquel ils proviennent. Notons ce nouveau score de propension pour un l'individu i $e_{AIC}(C_{i,k},\hat{\alpha}_k)$ et nous le déterminons de cette façon :

$$e_{AIC}(C_{i,k},\hat{\alpha}_k) = \sum_{j=1}^{m} p_{AIC_{j,k}} \cdot e(C_{i,k},\hat{\alpha}_{j,k}) \,. \tag{3.6}$$

Ensuite, nous utilisons ce nouveau score de propension dans le calcul de l'estimation par pondération par les probabilités inversées afin d'obtenir ce que nous appelons l'estimation par pondération interne :

$$\hat{\theta}_{int,AIC_k} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_{i,k}Y_{i,k}}{e_{AIC}(C_{i,k},\hat{\alpha}_k)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-X_{i,k})Y_{i,k}}{1-e_{AIC}(C_{i,k},\hat{\alpha}_k)} \,. \tag{3.7}$$

De manière analogue, nous pouvons décider de travailler à partir du BIC et utiliser plutôt

$$\hat{\theta}_{int,BIC_k} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_{i,k}Y_{i,k}}{e_{BIC}(C_{i,k},\hat{\alpha}_k)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-X_{i,k})Y_{i,k}}{1-e_{BIC}(C_{i,k},\hat{\alpha}_k)} \,, \tag{3.8}$$

où

$$e_{BIC}(C_{i,k},\hat{\alpha}_k) = \sum_{j=1}^{m} p_{BIC_{j,k}} \cdot e(C_{i,k},\hat{\alpha}_{j,k}) \,. \tag{3.9}$$

3.1.5 Estimateurs qui considèrent un ensemble de modèles potentiels

Pour faire référence aux six derniers estimateurs, ceux qui prennent en compte un ensemble de modèles potentiels, nous utilisons dorénavant l'abréviation ECMP (Estimation qui Considère l'ensemble des Modèles Potentiels). Pour la construction des estimateurs ECMP, nous utilisons les critères AIC et BIC. Ces derniers évaluent la qualité de l'ajustement des variables explicatives à la variable d'exposition X. Le choix des modèles se fait donc en ne tenant aucunement compte de la relation de ces variables avec la variable résultat Y. Cela n'est pas la situation idéale dans notre cas. Par exemple, des modèles plus grands, incluant des variables qui ont un effet sur Y, mais pas sur X, seront pénalisés, ce qui n'est pas souhaitable. Il a en effet déjà été suggéré que l'ajout des variables qui ont un effet sur Y, même si elles n'en ont pas sur X, peut diminuer la

variance de l'estimateur (Brookhart et al., 2006). L'utilisation des critères AIC et BIC est néanmoins courante dans la pratique.

Tel que mentionné précédemment, nous souhaitons voir si le fait de prendre une moyenne sur plusieurs modèles (pondération externe ou interne) peut améliorer l'estimation des effets de type causal. De plus, nous croyons que le fait de pondérer les scores de propension (pondération interne) peut aider à stabiliser les poids vers des valeurs moins extrêmes et ainsi diminuer la variance de l'estimateur par rapport aux autres estimateurs qui ne considèrent qu'un seul modèle pour le calcul du score de propension. Pour ce faire, nous comparons la méthode de pondération interne à celle de la pondération externe, à l'estimation obtenue des modèles présentant les plus petits AIC et BIC et aux estimations obtenues sous chaque modèle séparément. Les mesures que nous utilisons pour comparer les estimateurs sont le biais, la variance et l'erreur quadratique moyenne à travers k jeux de données. Nous examinons la technique lorsque le vrai modèle de traitement fait et ne fait pas partie de l'ensemble des modèles considérés.

CHAPITRE IV

SIMULATIONS

4.1 Premier exemple

Pour évaluer l'efficacité des estimateurs proposés dans le chapitre précédent, nous devons faire la supposition que nous connaissons et mesurons toutes les variables confondantes conceptuelles qui interviennent pour brouiller le lien de causalité entre l'exposition X et la variable résultat Y. Ce que nous ne savons pas, par contre, est la ou les formes sous lesquelles ces variables influencent X et Y. La sélection du modèle de traitement se fait donc au niveau de la forme des variables plutôt que de leur présence ou leur absence. En pratique, nous faisons cette même supposition que nous connaissons et mesurons toutes les variables confondantes pour pouvoir présumer que les estimations soient valides, mais nous n'avons pas de moyen de vérifier que cette supposition est vraie.

Pour les besoins de ce mémoire, nous générons nous-mêmes toutes nos données à l'aide du logiciel R (version 2.11.1). Nous présentons un premier exemple assez simple dans lequel interviennent trois variables confondantes, soit une variable C_1 générée d'une loi de Bernouilli avec probabilité 0,5, et deux variables, C_2 et C_3, générées d'une loi normale de moyenne 0 et de variance 1 :

$$C_1 \sim \mathcal{B}(0{,}5)\,; C_2, C_3 \sim \mathcal{N}(0,1)\,.$$

Ces trois variables jouent un rôle pour déterminer à la fois la probabilité pour un individu de recevoir ou non l'exposition X et la valeur de sa variable résultat Y. La relation entre

X, C_1, C_2 et C_3 s'exprime comme suit :

$$Pr[X = 1|C_1, C_2, C_3] = \frac{\exp(g(C_1, C_2, C_3))}{1 + (\exp g(C_1, C_2, C_3))}\,; \tag{4.1}$$

où $\quad g(C_1, C_2, C_3) = -0{,}5C_1 + 0{,}4C_2 + 0{,}25C_2^2 - 0{,}15C_2^3 + 0{,}4C_3 + 0{,}3C_1C_2 - 0{,}3C_1C_3\,.$

La probabilité (4.1), calculée pour chacun des individus à partir de ses propres valeurs pour les variables C_1, C_2 et C_3, est utilisée comme paramètre d'une loi de Bernouilli afin de générer aléatoirement l'exposition au traitement X. Après avoir simulé les variables C_1, C_2, C_3 et X, la variable résultat Y est générée à partir de la relation suivante :

$$\begin{aligned} Y &= 5 - 3C_1 - 2C_2 + 1{,}5C_2^2 - C_2^3 + 2C_3 \\ &+ 1{,}5C_1C_2 - 1{,}5C_1C_3 + 2V_1 - 2V_2 + 2V_3 + 2X + \epsilon\,, \end{aligned}$$

où V_1 et V_3 sont des variables générées d'une loi de Bernouilli de probabilité 0,5, V_2 une variable générée d'une loi normale $\mathcal{N}(0,1)$ et ϵ une variable générée d'une loi normale $\mathcal{N}(0,9)$:

$$V_1, V_3 \sim \mathcal{B}(0{,}5); V_2 \sim \mathcal{N}(0,1); \epsilon \sim \mathcal{N}(0,9)\,.$$

La variable ϵ est simplement un bruit aléatoire dans la détermination de la valeur de Y. Nous évaluons qu'ainsi environ 80% de la variabilité de la variable Y est déterminée par les variables C_1, C_2, C_3, V_1, V_2, V_3 et X, et 20% est due au hasard (ϵ). Notons que V_1, V_2 et V_3 ont un effet sur Y seulement et ne sont donc pas des variables confondantes. Le graphe acyclique représenté dans la Figure 4.1 illustre les relations de causalité existantes dans l'exemple présenté.

Nous simulons ainsi des échantillons de différentes tailles afin d'étudier le comportement des estimateurs. Nous présentons les résultats pour des échantillons de 300, 500 et 1500 individus. Pour chacune de ces tailles d'échantillon, nous simulons 5000 jeux de données. Pour le jeu de données k, $k = 1, \ldots, 5000$, sont calculées l'estimation sous chacun des modèles, ainsi que les estimations sous le modèle ayant le plus petit AIC ou BIC ($\hat{\theta}_{AIC,k}$ et $\hat{\theta}_{BIC,k}$), les estimations par la pondération externe ($\hat{\theta}_{ext,AIC_k}$ et $\hat{\theta}_{ext,BIC_k}$) et les estimations par la pondération interne ($\hat{\theta}_{int,AIC_k}$ et $\hat{\theta}_{int,BIC_k}$) (équations 3.1, 3.4, 3.5, 3.7

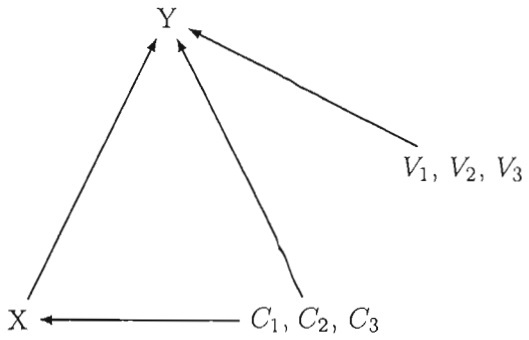


Figure 4.1 Graphe orienté acyclique représentant les liens de causalité entre Y, X, C_1, C_2, C_3, V_1, V_2 et V_3.

et 3.8). Pour comprendre la composition des estimateurs pondérés externes et internes, nous incluons dans les résultats les probabilités moyennes a posteriori (équations 3.2 et 3.3) de chacun des modèles. Afin de comparer chacune des estimations, nous présentons leur biais, variance et erreur quadratique moyenne (EQM).

Comme nous l'avons déjà mentionné, l'estimateur pondéré par les probabilités inversées est très sensible au choix du modèle de prédiction de X. Après quelques simulations, nous constatons rapidement que des poids trop extrêmes, même chez seulement quelques individus, peuvent engendrer une estimation grandement erronée et gonfler exagérément la variance de l'estimateur à travers les 5000 répétitions. Les individus qui engendrent ces poids extrêmes représentent des cas « rares » et puisque nous travaillons avec des échantillons dont la taille n'est pas suffisamment grande pour estomper leur impact dans l'estimation de l'effet causal, la variabilité engendrée par leur présence est trop grande par rapport à l'information apportée. Nous prenons donc la décision d'imposer une valeur maximale au poids que peut prendre un individu dans le calcul des estimations. Nous fixons cette valeur à 100. En pratique, cela signifie que nous imposons une valeur de 0,01 au score de propension de tout individu qui est exposé, mais dont le score de propension estimé est inférieur à 0,01 et une valeur de 0,99 aux individus non exposés dont le score de propension estimé dépasse 0,99. Cette pratique, appelée troncation des poids, est souvent employée en pratique afin d'obtenir un compromis entre le biais et la

Tableau 4.1 Différents modèles considérés dans les estimations

	C_1	C_2	C_2^2	C_2^3	C_3	C_3^2	C_3^3	V_1, V_2, V_3	C_1C_2	C_1C_3
M1 ⋆vrai⋆ ; (M8)	♦	♦	♦	♦	♦			(♦)	♦	♦
M2 ; (M9)	♦	♦			♦			(♦)		
M3 ; (M10)	♦	♦	♦		♦	♦		(♦)		
M4 ; (M11)	♦	♦	♦	♦	♦	♦	♦	(♦)		
M5 ; (M12)	♦	♦			♦			(♦)	♦	♦
M6 ; (M13)	♦	♦	♦		♦	♦		(♦)	♦	♦
M7 ; (M14)	♦	♦	♦	♦	♦	♦	♦	(♦)	♦	♦

Les variables entre ( ), V_1, V_2 et V_3, sont ajoutées aux modèles entre ( ), M8 à M14.

variance (Cole et Hernan, 2008) des estimateurs par la pondération par les probabilités inversées.

Nous voulons évaluer l'efficacité des estimateurs lorsque le vrai modèle de prédiction du traitement, c'est-à-dire M1, fait partie de l'ensemble des modèles utilisés pour les estimations de l'effet causal et quand il n'en fait pas partie. Pour ce faire, nous considérons trois situations. Dans la première, le vrai modèle est parmi l'ensemble des modèles considérés ; dans la deuxième, le vrai modèle ne fait pas partie de l'ensemble des modèles, mais d'autres modèles sensiblement proches du vrai modèle s'y retrouvent, et dans le dernier cas, ni le vrai modèle, ni des modèles qui emboîtent le vrai modèle ne sont considérés. Nous retrouvons dans le Tableau 4.1 les quatorze modèles utilisés dans l'une ou l'autre de ces situations. Un losange indique que la variable fait partie du modèle et les variables entre parenthèses, V_1, V_2 et V_3, sont ajoutées aux modèles entre parenthèses (M8, ..., M14).

Préalablement aux calculs des estimations qui nous intéressent, nous estimons le lien causal sans tenir compte des variables autres que l'exposition. Pour ce faire, nous prenons simplement la différence de la variable Y entre les groupes d'individus exposés et non exposés. À travers les 5000 jeux de données, avec une taille d'échantillon de 300 individus, nous obtenons qu'en moyenne l'effet de la variable X est $\hat{\mu}_{X=1} - \hat{\mu}_{X=0} =$

3,378, avec une variance de 0,655. Avec une taille d'échantillon de 500, nous obtenons $\hat{\mu}_{X=1} - \hat{\mu}_{X=0} = 3{,}384$ avec une variance de 0,407, et en augmentant la taille à 1500, nous obtenons $\hat{\mu}_{X=1} - \hat{\mu}_{X=0} = 3{,}396$ avec une variance de 0,126. Nous savons que la vraie différence de risque est de 2. La différence entre la vraie valeur et la valeur estimée par différence de moyennes est due à l'effet des variables confondantes que nous devons considérer dans nos analyses.

4.1.1 Première situation

Pour cette première situation, nous utilisons tous les modèles M1, ... , M14 (Tableau 4.1). Le vrai modèle est donc parmi l'ensemble des modèles considérés. Le Tableau 4.2 présente les résultats obtenus quand nous considérons des échantillons de 300 individus. Les modèles qui ont une probabilité moyenne supérieure à 8% d'être le bon modèle selon le AIC ou le BIC sont présentés. Pour ce même ensemble de modèles, les Tableaux 4.3 et 4.4 présentent les résultats obtenus pour des échantillons de 500 et 1500 individus. Les résultats complets sont disponibles pour consultation dans les Tableaux A.1, A.2 et A.3 de l'appendice A pour $n = 300$, $n = 500$ et $n = 1500$ respectivement.

Nous pouvons voir que les biais obtenus pour les estimations sous les modèles M1, M7, M8 et M14 sont sensiblement les mêmes. Cela s'explique par le fait que M1 est emboîté dans M7, M8 et M14, et que les variables ou la forme des variables qui se trouvent en supplément dans ces trois derniers modèles ($C_3^2, C_3^3, V_1, V_2, V_3$) ne sont pas des variables confondantes. Les moyennes des estimations sous les autres modèles s'éloignent toutes davantage de la vraie valeur 2, car pour chacun de ces modèles l'une ou l'autre des formes des variables confondantes est manquante. Les modèles M2 et M9 étant les plus parcimonieux des modèles sont en effet ceux qui produisent les plus grands biais de l'estimation. Nous remarquons également que les modèles M1, M7, M8 et M14, malgré leur similarité quant à la valeur du biais, produisent des variances différentes. La variance supérieure sous M7 par rapport à M1 est probablement causée par les variables C_3^2 et C_3^3 qui ne sont pas utiles dans cette situation. Cependant, dans le cas du modèle M8,

Tableau 4.2 Résultats obtenus en considérant tous les modèles (M1 à M14) avec $n = 300$ et $ech = 5000$

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M7}$	$\hat{\theta}_{M8}$
Biais	0,211	1,477	1,112	0,423	0,214	0,206
Variance	0,789	0,495	1,118	0,683	0,828	0,759
EQM	0,833	2,678	2,354	0,862	0,874	0,801
$\overline{p_{AIC}} \times 10^2$	31,491	2,824	6,528	14,053	12,922	9,514
$\overline{p_{BIC}} \times 10^2$	16,556	53,455	17,023	5,485	0,292	0,062

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,290	1,007	0,347	0,974	0,357	0,911
Variance	0,799	0,732	0,673	0,537	0,521	0,363
EQM	0,883	1,747	0,794	1,486	0,648	1,193

malgré le fait que V_1, V_2 et V_3 ne soient pas des variables confondantes, la variance est réduite par l'ajout de ces variables. De manière générale, les estimations moyennes de l'effet causal sont plus près de la vraie valeur et les variances sont réduites lorsque la taille de l'échantillon augmente.

Pour évaluer les estimateurs ECMP, nous devons examiner les résultats séparément selon les trois tailles échantillonnales $n = 300, 500, 1500$. Pour $n = 300$, nous remarquons que $\hat{\theta}_{AIC}$ est celui des six estimateurs ECMP qui est le plus près en moyenne de la vraie valeur. Les deux estimateurs qui utilisent une pondération basée sur le critère AIC ($\hat{\theta}_{ext,AIC}$ et $\hat{\theta}_{int,AIC}$) ont également un biais plus petit que les estimateurs pondérés par le BIC. Effectivement, nous voyons qu'en moyenne, le critère AIC choisit le modèle M1 et ceux qui lui ressemblent comme étant les plus probables. Quant au critère BIC, il désigne plus souvent les modèles M2 et M3 comme étant les plus probables. Comme nous l'avons mentionné dans le précédent chapitre, le critère BIC a tendance à pénaliser davantage la complexité des modèles. Les modèles M2 et M3 sont parmi les plus petits modèles, mais

Tableau 4.3 Résultats obtenus en considérant tous les modèles (M1 à M14) avec $n = 500$ et $ech = 5000$

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M7}$	$\hat{\theta}_{M8}$
Biais	0,143	1,467	1,187	0,367	0,144	0,150
Variance	0,579	0,277	1,021	0,507	0,592	0,546
EQM	0,599	2,428	2,430	0,642	0,612	0,569
$\overline{p_{AIC}} \times 10^2$	41,931	0,357	2,419	11,784	16,924	12,407
$\overline{p_{BIC}} \times 10^2$	40,135	28,630	16,806	8,779	0,411	0,091

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,170	0,622	0,206	0,629	0,220	0,605
Variance	0,572	0,626	0,508	0,455	0,419	0,326
EQM	0,601	1,014	0,550	0,850	0,467	0,692

Tableau 4.4 Résultats obtenus en considérant tous les modèles (M1 à M14) avec $n = 1500$ et $ech = 5000$

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M7}$	$\hat{\theta}_{M8}$
Biais	0,082	1,468	1,267	0,319	0,084	0,082
Variance	0,265	0,087	0,442	0,230	0,267	0,251
EQM	0,272	2,243	2,046	0,333	0,275	0,258
$\overline{p_{AIC}} \times 10^2$	53,318	$< 0,001$	0,001	1,373	21,994	16,187
$\overline{p_{BIC}} \times 10^2$	96,172	0,024	0,373	2,850	0,435	0,057

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,082	0,085	0,084	0,089	0,086	0,091
Variance	0,261	0,266	0,256	0,265	0,254	0,261
EQM	0,268	0,273	0,263	0,273	0,261	0,270

également ceux qui produisent un plus grand biais. La pondération externe inhérente aux estimateurs $\hat{\theta}_{ext,AIC}$ et $\hat{\theta}_{ext,BIC}$ a pour conséquence de réduire la variance de ces estimateurs en comparaison à $\hat{\theta}_{AIC}$ et $\hat{\theta}_{BIC}$, et ainsi réduit également l'EQM. Pour ce qui est des estimateurs par pondération interne, $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$, nous voyons que leur variance est encore plus petite que celle de $\hat{\theta}_{ext,AIC}$ et $\hat{\theta}_{ext,BIC}$, respectivement, de même que leur EQM. Pour les trois types d'estimateur, l'EQM est nettement plus petite pour les déclinaisons AIC que BIC parce que le biais engendré par les modèles sélectionnés par le BIC est plus important que la réduction de la variance.

Avec une taille échantillonnale de $n = 500$ individus, les deux critères, AIC et BIC, désignent désormais le modèle M1 comme étant le plus probable avec une probabilité moyenne d'environ 40%. Le reste des probabilités est encore principalement attribué aux modèles M7 et M8 par le AIC et aux modèles M2 et M3 par le BIC. Le fait que le BIC accorde maintenant plus d'importance au premier modèle, tout comme le fait le AIC, fait en sorte que les estimations basées sur le AIC et le BIC sont un peu moins éloignées qu'elles le sont avec 300 individus. Cependant, les méthodes basées sur le AIC donnent encore de meilleurs résultats, tant au niveau du biais que de l'EQM. Toujours dans ce cas, les estimateurs $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$ présentent des EQM plus petites que $\hat{\theta}_{ext,AIC}$ et $\hat{\theta}_{ext,BIC}$, qui elles en présentent de plus petites que $\hat{\theta}_{AIC}$ et $\hat{\theta}_{BIC}$.

Lorsque nous considérons des échantillons de plus grande taille, par exemple ici avec $n = 1500$, nous observons que les comportements des critères AIC et BIC sont plutôt différents. Le AIC donne toujours une plus grande importance au vrai modèle M1, mais conserve des probabilités non négligeables pour les modèles M7 et M8, qui emboîtent M1. Le BIC, lui, accorde presque 100% du poids au modèle M1. De plus, comme le critère BIC sélectionne le modèle M1 comme étant le plus probable dans 98% des 5000 réplications, cela entraîne que les estimateurs $\hat{\theta}_{BIC}$, $\hat{\theta}_{ext,BIC}$ et $\hat{\theta}_{int,BIC}$ sont pratiquement équivalents. Nous remarquons que les résultats obtenus pour les trois estimations basées sur le critère AIC sont eux aussi très près les uns des autres, mais contrairement à ce que nous observons avec le BIC, trois modèles plutôt qu'un seul se partagent 90% du poids. Cela est dû au fait que, comme nous l'avons mentionné précédemment, les

estimations moyennes obtenues sous les modèles M1, M7 et M8 sont presque les mêmes et que, comme nous avons augmenté le nombre d'individus, les variances observées se rapprochent considérablement. Dans cet exemple, nous observons qu'à partir d'une taille échantillonnale suffisamment grande, les six estimateurs ECMP donnent des résultats pratiquement équivalents.

À travers les échantillons de différentes tailles, nous observons que les estimateurs $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$ semblent offrir une performance plus intéressante au niveau de l'EQM que $\hat{\theta}_{AIC}$, $\hat{\theta}_{BIC}$, $\hat{\theta}_{ext,AIC}$ et $\hat{\theta}_{ext,BIC}$. Nous sommes d'autant plus satisfaits de constater que $\hat{\theta}_{int,AIC}$ a même cet avantage relativement à l'estimateur basé sur le vrai modèle, $\hat{\theta}_{M1}$, grâce à une diminution de la variance de celui-ci.

4.1.2 Deuxième situation

En deuxième lieu, nous nous intéressons aux estimations obtenues lorsque le vrai modèle, M1, ne fait pas partie de l'ensemble des modèles considérés pour l'estimation, mais que d'autres modèles qui lui ressemblent en font partie. Par exemple, avec des modèles comme M7, M8 ou M14, qui comprennent toutes les formes des variables qui interviennent dans le vrai modèle, nous pensons que nous sommes en mesure d'obtenir de très bons résultats. Pour la deuxième situation, nous considérons donc tous les mêmes modèles que pour la première situation, à l'exception de M1. Dans les Tableaux 4.5, 4.6 et 4.7, nous retrouvons les résultats obtenus pour les modèles M2, M3, M4, M7, M8 et M14, qui obtiennent des probabilités a posteriori supérieures à 8% pour l'une ou l'autre des tailles d'échantillon. Dans l'appendice A, les Tableaux A.4, A.5 et A.6 contiennent les résultats complets pour cette situation. Pour une même taille d'échantillon, les résultats sous chacun des modèles considérés sont évidemment identiques à ceux que nous obtenons dans la première situation. Ce sont les résultats des derniers estimateurs (ECMP) qui nous intéressent ici.

Lorsque nous examinons les résultats obtenus avec des échantillons de 300 observations, nous remarquons d'abord que chacun des six estimateurs d'intérêt a un plus grand

Tableau 4.5 Résultats obtenus en considérant les modèles M2 à M14 avec $n = 300$ et $ech = 5000$

	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M7}$	$\hat{\theta}_{M8}$	$\hat{\theta}_{M14}$
Biais	1,477	1,112	0,423	0,214	0,206	0,210
Variance	0,495	1,118	0,683	0,828	0,759	0,780
EQM	2,678	2,354	0,862	0,874	0,801	0,824
$\overline{p_{AIC}} \times 10^2$	3,599	8,296	20,204	21,856	15,421	5,637
$\overline{p_{BIC}} \times 10^2$	58,729	19,882	9,798	1,288	0,270	0,003

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,355	1,164	0,406	1,130	0,409	1,041
Variance	0,789	0,656	0,635	0,508	0,474	0,339
EQM	0,915	2,010	0,800	1,784	0,642	1,421

biais que dans la première situation. Ceci est dû au fait que le modèle M1, celui qui présentait précédemment le plus petit biais, n'est plus dans l'ensemble des modèles considérés. Les probabilités qui lui étaient assignées ($\overline{p_{AIC_1}} = 0{,}315$; $\overline{p_{BIC_1}} = 0{,}166$) sont désormais réparties sur l'ensemble des modèles restants. Le critère AIC redistribue 21% aux modèles M4, M7 et M8, qui ressemblent beaucoup à M1. Le critère BIC accorde seulement environ 6% de plus à ces trois modèles, alors que 8% sont ajoutés aux modèles M2 et M3. Au niveau du biais, l'écart s'agrandit donc entre les estimateurs pondérés qui sont basés sur le critère AIC et ceux qui utilisent le critère BIC. Nous remarquons également que la variance est un peu plus petite dans cette situation pour l'ensemble des estimateurs d'intérêt. Cependant, l'augmentation du biais est plus importante que la diminution de la variance, ce qui engendre des EQM plus grandes ou égales à celles calculées dans la première situation. Les principales conclusions restent toutefois les mêmes, c'est-à-dire que l'EQM des estimateurs basés sur le AIC est plus petite que celle des estimateurs basés sur le BIC et que les estimateurs par pondération interne donnent aussi ici de plus petites EQM que les quatre autres estimateurs ECMP. Enfin, nous

Tableau 4.6 Résultats obtenus en considérant les modèles M2 à M14 avec $n = 500$ et $ech = 5000$

	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M7}$	$\hat{\theta}_{M8}$	$\hat{\theta}_{M14}$
Biais	1,467	1,187	0,367	0,144	0,150	0,151
Variance	0,277	1,021	0,507	0,592	0,546	0,555
EQM	2,428	2,430	0,642	0,612	0,569	0,578
$\overline{p_{AIC}} \times 10^2$	0,525	3,452	18,741	32,875	22,761	7,871
$\overline{p_{BIC}} \times 10^2$	36,892	23,762	21,857	4,514	0,899	0,005

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,204	0,907	0,244	0,902	0,260	0,838
Variance	0,532	0,527	0,479	0,370	0,374	0,246
EQM	0,574	1,350	0,539	1,184	0,442	0,949

obtenons encore que l'estimateur $\hat{\theta}_{int,AIC}$ affiche la plus petite EQM en comparaison à l'estimateur basé sur chacun des treize modèles, ainsi qu'aux cinq autres estimateurs basés sur les critères d'information AIC ou BIC.

Dans la première situation, avec les échantillons de taille 500, le AIC accorde 42% et le BIC 41% en probabilité au modèle M1. En enlevant ce modèle, la redistribution de ces probabilités va encore principalement aux modèles M4 (7%), M7 (16%) et M8 (10%) par le AIC et aux modèles M2 (8%), M3 (7%) et M4 (13%) par le BIC. Nous tirons de cet exemple les mêmes conclusions générales que précédemment avec les échantillons de taille $n = 300$. Notons également que l'écart entre l'EQM calculées pour les estimateurs basés sur le AIC et ceux basés sur le BIC est plus grand dans la deuxième situation que dans la première avec $n = 500$. En effet, nous avons maintenant que l'EQM pour un estimateur BIC est le double du même estimateur en version AIC.

Tout comme dans la première situation, nous remarquons qu'en augmentant la taille échantillonnale à 1500, les résultats découlant des estimateurs en version AIC et en

Tableau 4.7 Résultats obtenus en considérant les modèles M2 à M14 avec $n = 1500$ et $ech = 5000$

	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M7}$	$\hat{\theta}_{M8}$	$\hat{\theta}_{M14}$
Biais	1,468	1,267	0,319	0,084	0,082	0,083
Variance	0,087	0,442	0,230	0,267	0,251	0,252
EQM	2,243	2,046	0,333	0,275	0,258	0,259
$\overline{p_{AIC}} \times 10^2$	$< 0,001$	0,003	2,683	50,828	34,180	11,539
$\overline{p_{BIC}} \times 10^2$	0,223	3,030	37,481	50,112	5,783	0,009

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,083	0,187	0,087	0,196	0,089	0,206
Variance	0,259	0,269	0,253	0,248	0,250	0,221
EQM	0,266	0,304	0,261	0,287	0,258	0,264

version BIC se rapprochent beaucoup. Cependant, la différence est plus grande dans la deuxième situation pour les estimations provenant des modèles ayant le plus petit AIC ou BIC et des estimateurs par pondération externe. En regardant la redistribution des probabilités, nous comprenons pourquoi il en est ainsi. Dans la première situation, le AIC accorde 53% de probabilité à M1 et le BIC 96%. Rappelons-nous que dans la première situation, le AIC sélectionne principalement les modèles M7 et M8 pour compléter, avec M1, la pondération, alors que le BIC met presque tout le poids sur M1. Les modèles M7 et M8 présentent une moyenne et une variance assez proches de M1. C'est ce qui fait que les versions AIC et BIC sont semblables. Dans la deuxième situation, le retrait de M1 entraîne principalement l'augmentation des probabilités des modèles M7 et M8 par le AIC et des modèles M4 et M7 par le BIC. Pour ce qui est de l'estimation sous le meilleur modèle, plus de 90% du temps ce sont les modèles M7 et M8 qui sont choisis par le AIC, alors que le BIC choisit plutôt M7 et M4, 39% et 53% du temps respectivement. Le modèle M4, contrairement aux modèles M7 et M8, n'emboîte pas le vrai modèle M1. Le biais est donc plus grand sous M4. C'est ce qui fait

que les écarts d'EQM entre $\hat{\theta}_{AIC}$ et $\hat{\theta}_{BIC}$ et entre $\hat{\theta}_{ext,AIC}$ et $\hat{\theta}_{ext,BIC}$ sont plus grands. Pour les estimateurs par pondération interne, l'estimateur $\hat{\theta}_{int,BIC}$ possède une variance suffisamment inférieure à $\hat{\theta}_{int,AIC}$ pour que la différence entre les EQM soit moindre.

Rappelons que, alors que dans la première situation nous considérons tous les modèles incluant le vrai (M1, ..., M14), dans la deuxième situation nous omettons le modèle M1. Comme nous l'espérions, il semble que dans la deuxième situation, l'estimateur par pondération interne fournisse encore une fois de bons résultats. En effet, tout comme dans la première situation, l'estimateur $\hat{\theta}_{int,AIC}$ est celui qui obtient la plus petite EQM et il semble avantageux d'utiliser cette technique surtout avec les échantillons de 300 et de 500 observations.

4.1.3 Troisième situation

La dernière situation que nous voulons étudier est le cas où ni le vrai modèle, ni des modèles lui ressemblant ne font partie de l'ensemble des modèles considérés. Nous choisissons donc d'inclure dans l'ensemble seulement les modèles M2, M3 et M5. Pour ces trois modèles C_2^3 est manquante et aucun n'inclut à la fois la forme quadratique de C_2 et les interactions entre C_1, C_2 et C_3. Ici, nous désirons déterminer comment les estimateurs provenant du moyennage de modèles iront chercher l'information à travers ces trois modèles. Les résultats complets pour les échantillons de taille 300, 500 et 1500 se trouvent dans les Tableaux 4.8, 4.9 et 4.10.

Avec aussi peu que 300 observations, le AIC autant que le BIC accorde plus de la moitié du poids total à un seul modèle. Dans le cas du AIC, c'est le modèle M3 qui obtient un poids moyen de 0,55, alors que le BIC favorise encore le modèle contenant le moins de paramètres, M2, lui accordant un poids moyen de 0,66. Le modèle M3 donne une estimation moyenne moins biaisée que les deux autres modèles, mais sa variance est près du double de celle obtenue sous M2 et M5. Cela fait en sorte que l'écart entre les résultats obtenus des estimateurs basés sur le AIC et sur le BIC est moins grand que dans les deux précédentes situations. Par contre, l'écart observé entre les estimations

Tableau 4.8 Résultats obtenus en considérant les modèles M2, M3 et M5 avec $n = 300$ et $ech = 5000$

	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M5}$
Biais	1,477	1,112	1,384
Variance	0,495	1,118	0,672
EQM	2,678	2,354	2,588
$\overline{p_{AIC}} \times 10^2$	17,413	55,179	27,408
$\overline{p_{BIC}} \times 10^2$	65,909	25,380	8,711

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	1,150	1,317	1,173	1,309	1,050	1,198
Variance	1,016	0,751	0,894	0,666	0,631	0,460
EQM	2,339	2,486	2,269	2,378	1,733	1,896

pondérées à l'interne et celles pondérées à l'externe est très marqué. Nous notons une amélioration à la fois au niveau du biais et de la variance lorsque c'est la pondération interne qui est utilisée.

Comme dans les deux premières situations, en augmentant la taille d'échantillon à 500, les trois types d'estimateurs qui considèrent l'ensemble de modèles donnent des résultats qui se rapprochent. Tant le critère AIC que BIC retirent du poids au modèle M2. Le premier le redistribue presque totalement à M3, alors que le BIC, tout en accordant davantage à M3, augmente également la probabilité de M5. Le fait intéressant à remarquer est que les deux estimateurs par pondération interne ($\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$) donnent maintenant de meilleurs résultats au niveau du biais et de la variance que les quatre autres estimateurs d'intérêt en variation AIC ou BIC. Notre proposition de pondérer le score de propension (pondération interne) est donc ici clairement la meilleure des façons suggérées pour prendre en compte l'incertitude de modèle puisqu'elle fournit à la fois les plus petits biais et les plus petites variances.

Tableau 4.9 Résultats obtenus en considérant les modèles M2, M3 et M5 avec $n = 500$ et $ech = 5000$

	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M5}$
Biais	1,467	1,187	1,369
Variance	0,277	1,021	0,335
EQM	2,428	2,430	2,210
$\overline{p_{AIC}} \times 10^2$	7,398	66,040	26,562
$\overline{p_{BIC}} \times 10^2$	49,150	39,355	11,495

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	1,200	1,292	1,213	1,296	1,101	1,172
Variance	0,921	0,641	0,826	0,577	0,550	0,347
EQM	2,361	2,310	2,297	2,255	1,763	1,721

Tableau 4.10 Résultats obtenus en considérant les modèles M2, M3 et M5 avec $n = 1500$ et $ech = 5000$

	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M5}$
Biais	1,468	1,267	1,362
Variance	0,087	0,442	0,102
EQM	2,243	2,046	1,958
$\overline{p_{AIC}} \times 10^2$	0,075	84,086	15,838
$\overline{p_{BIC}} \times 10^2$	4,463	81,120	14,417

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	1,266	1,271	1,266	1,273	1,219	1,218
Variance	0,405	0,402	0,396	0,388	0,353	0,337
EQM	2,007	2,017	2,000	2,007	1,838	1,819

Contrairement aux deux premières situations, où en considérant des échantillons de taille 1500 les six estimateurs ECMP donnent des résultats relativement semblables pour le biais et la variance, nous voyons que pour la troisième situation les estimateurs par pondération interne semblent offrir les meilleurs résultats. Une réduction du biais et de la variance nous procure des EQM clairement plus petites que celles obtenues sous les autres estimateurs. Il semble donc réellement avantageux de préconiser l'approche par pondération interne dans ce cas.

Cette dernière situation nous amène à d'intéressantes conclusions au sujet des estimateurs étudiés. Il semble que lorsqu'aucun des modèles considérés n'inclut le vrai modèle, il soit préférable d'utiliser un moyennage à travers tous les modèles plutôt que d'en sélectionner un seul pour estimer l'effet causal. Le biais de l'estimation résultant d'un moyennage est souvent plus petit que le biais obtenu sous chacun des différents modèles et ceci est d'autant plus vrai lorsque c'est le critère AIC qui est utilisé avec les petits et moyens échantillons. Cependant, dans la troisième situation, nous remarquons que pour les trois types d'estimateurs ECMP avec des échantillons de taille $n = 500$ ainsi que pour l'estimateur par pondération interne quand $n = 1500$, le critère BIC fournit une EQM plus petite que le critère AIC. De plus, alors que dans les deux premières situations les résultats de nos estimateurs d'intérêt deviennent presque identiques pour des échantillons de 1500 observations, nous notons que dans la troisième situation, pour une telle taille d'échantillon, les estimateurs par pondération interne donnent des résultats plus intéressants que les autres estimateurs ECMP.

À travers les trois situations présentées, c'est-à-dire :

1. Le cas où nous considérons tous les modèles (M1 à M14) ;
2. Le cas où nous considérons tous les modèles sauf le vrai modèle (M2 à M14) ;
3. Le cas où nous ne considérons que les modèles incomplets M2, M3 et M5 ;

nous remarquons quelques tendances générales. Dans la plupart des cas, avec les deux types de pondération (interne et externe), nous obtenons que l'utilisation du critère AIC produit un biais plus petit et une variance plus grande que l'utilisation du critère BIC.

Ceci correspond à ce que nous pouvions prévoir, car nous savons que le AIC favorise souvent des modèles plus grands que le BIC, ce qui engendre un meilleur ajustement des données (plus petit biais), mais peut entraîner une plus grande variance. L'application de l'estimateur par pondération interne réduit clairement cette variance lorsque la taille échantillonnale n'est pas trop grande.

4.2 Sur les poids

4.2.1 Poids sous chacun des estimateurs

Dans cette section, nous expliquons pourquoi la variance est plus petite pour les estimateurs par pondération interne ($\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$) que pour les estimateurs par pondération externe ($\hat{\theta}_{ext,AIC}$ et $\hat{\theta}_{ext,BIC}$). Pour ce faire, nous regardons le poids attribué à un individu i dans le calcul de chacune des estimations. En effet, nous savons que la variance des estimateurs pondérés est directement liée aux poids donnés aux individus et que de grands poids entraînent généralement une grande variance. Pour faciliter la compréhension, supposons un contexte où nous considérons seulement deux modèles, A et B. Les poids (probabilités) associés à ces deux modèles sont dénotés par p_A et p_B et les scores de propension de l'individu i sont respectivement $e(C_i,\hat{\alpha}_A)$ et $e(C_i,\hat{\alpha}_B)$ sous le premier et le deuxième modèle. Notons que puisque la somme des poids associés aux modèles est égale à un, nous avons alors que $p_B = (1 - p_A)$. Supposons également que l'individu est exposé au traitement ($X = 1$). Nous avons alors que le poids qui lui est associé en considérant le modèle A est :

$$\frac{1}{e(C_i,\hat{\alpha}_A)}, \tag{4.2}$$

et nous notons ce poids $\omega_{i,A}$. Ce poids est multiplié par la variable Y_i et c'est ce qui complète la contribution de l'individu i à l'estimation. L'importance de la contribution de l'individu i à l'estimateur par pondération par les probabilités inversées sous le modèle A dépend donc de ces deux variables, Y_i et $e(C_i,\hat{\alpha}_A)$.

Lorsque nous appliquons le moyennage à l'aide de l'estimateur par pondération externe,

la variable Y_i demeure la même, mais le poids qui la multiplie est modifié. Nous avons alors le poids suivant :

$$\begin{aligned} \omega_{i,ext} &= p_A \cdot \frac{1}{e(C_i,\hat{\alpha}_A)} + p_B \cdot \frac{1}{e(C_i,\hat{\alpha}_B)} \\ &= \frac{p_A}{e(C_i,\hat{\alpha}_A)} + \frac{p_B}{e(C_i,\hat{\alpha}_B)} . \end{aligned} \tag{4.3}$$

Avec l'estimateur par pondération interne, le poids de ce même individu est plutôt :

$$\omega_{i,int} = \frac{1}{p_A \cdot e(C_i,\hat{\alpha}_A) + p_B \cdot e(C_i,\hat{\alpha}_B)} . \tag{4.4}$$

Proposition. *Nous montrons dans ce qui suit que le poids attribué par la pondération interne (4.4) est inférieur ou égal à celui attribué par la pondération externe (4.3) pour un individu i exposé au traitement X, c'est-à-dire :*

$$\frac{p_A}{e(C_i,\hat{\alpha}_A)} + \frac{p_B}{e(C_i,\hat{\alpha}_B)} \geq \frac{1}{p_A \cdot e(C_i,\hat{\alpha}_A) + p_B \cdot e(C_i,\hat{\alpha}_B)} . \tag{4.5}$$

Preuve.

Nous avons : $0 < p_A,\ p_B,\ e(C_i,\hat{\alpha}_A),\ e(C_i,\hat{\alpha}_B) < 1$ *et* $p_B = (1 - p_A)$.

$$\begin{aligned} (e(C_i,\hat{\alpha}_A) - e(C_i,\hat{\alpha}_B))^2 &\geq 0 ; \\ e(C_i,\hat{\alpha}_A)^2 - 2e(C_i,\hat{\alpha}_A) \cdot e(C_i,\hat{\alpha}_B) + e(C_i,\hat{\alpha}_B)^2 &\geq 0 ; \\ e(C_i,\hat{\alpha}_A)^2 + e(C_i,\hat{\alpha}_B)^2 &\geq 2e(C_i,\hat{\alpha}_A) \cdot e(C_i,\hat{\alpha}_B) ; \\ \frac{e(C_i,\hat{\alpha}_A)}{2e(C_i,\hat{\alpha}_B)} + \frac{e(C_i,\hat{\alpha}_B)}{2e(C_i,\hat{\alpha}_A)} &\geq 1 ; \end{aligned}$$

$$\begin{aligned} (1 - p_A) \cdot \left(\frac{e(C_i,\hat{\alpha}_A)}{2e(C_i,\hat{\alpha}_B)} + \frac{e(C_i,\hat{\alpha}_B)}{2e(C_i,\hat{\alpha}_A)} \right) &\geq (1 - p_A) ; \\ p_A + (1 - p_A) \cdot \left(\frac{e(C_i,\hat{\alpha}_A)}{2e(C_i,\hat{\alpha}_B)} + \frac{e(C_i,\hat{\alpha}_B)}{2e(C_i,\hat{\alpha}_A)} \right) - 1 &\geq 0 ; \\ 2p_A + (1 - p_A) \cdot \left(\frac{e(C_i,\hat{\alpha}_A)}{e(C_i,\hat{\alpha}_B)} + \frac{e(C_i,\hat{\alpha}_B)}{e(C_i,\hat{\alpha}_A)} \right) - 2 &\geq 0 ; \end{aligned}$$

Finalement,

$$\begin{aligned}
p_A\left(2p_A + (1-p_A)\cdot\left(\frac{e(C_i,\hat{\alpha}_A)}{e(C_i,\hat{\alpha}_B)} + \frac{e(C_i,\hat{\alpha}_B)}{e(C_i,\hat{\alpha}_A)}\right) - 2\right) &\geq 0\,;\\
p_A^2 + p_A\cdot(1-p_A)\cdot\left(\frac{e(C_i,\hat{\alpha}_A)}{e(C_i,\hat{\alpha}_B)} + \frac{e(C_i,\hat{\alpha}_B)}{e(C_i,\hat{\alpha}_A)}\right) - 2p_A + p_A^2 + 1 &\geq 1\,;\\
p_A^2 + p_A\cdot(1-p_A)\cdot\left(\frac{e(C_i,\hat{\alpha}_A)}{e(C_i,\hat{\alpha}_B)} + \frac{e(C_i,\hat{\alpha}_B)}{e(C_i,\hat{\alpha}_A)}\right) + (1-p_A)^2 &\geq 1\,;\\
p_A^2 + p_A\cdot p_B\cdot\left(\frac{e(C_i,\hat{\alpha}_A)}{e(C_i,\hat{\alpha}_B)} + \frac{e(C_i,\hat{\alpha}_B)}{e(C_i,\hat{\alpha}_A)}\right) + p_B^2 &\geq 1\,;\\
\frac{p_A^2\cdot e(C_i,\hat{\alpha}_A)}{e(C_i,\hat{\alpha}_A)} + \frac{p_A\cdot p_B\cdot e(C_i,\hat{\alpha}_A)}{e(C_i,\hat{\alpha}_B)} + \frac{p_A\cdot p_B\cdot e(C_i,\hat{\alpha}_B)}{e(C_i,\hat{\alpha}_A)} + \frac{p_B^2\cdot e(C_i,\hat{\alpha}_B)}{e(C_i,\hat{\alpha}_B)} &\geq 1\,;\\
(p_A\cdot e(C_i,\hat{\alpha}_A) + p_B\cdot e(C_i,\hat{\alpha}_B))\cdot\left(\frac{p_A}{e(C_i,\hat{\alpha}_A)} + \frac{p_B}{e(C_i,\hat{\alpha}_B)}\right) &\geq 1\,;
\end{aligned}$$

Ainsi,

$$\left(\frac{p_A}{e(C_i,\hat{\alpha}_A)} + \frac{p_B}{e(C_i,\hat{\alpha}_B)}\right) \geq \frac{1}{(p_A\cdot e(C_i,\hat{\alpha}_A) + p_B\cdot e(C_i,\hat{\alpha}_B))}\,.$$

Notons que les deux types de pondération donnent le même poids à l'individu si celui-ci présente le même score de propension sous le modèle A et le modèle B.

Pour la démonstration, nous avons supposé que l'individu i est exposé au traitement, mais la conclusion est la même s'il ne l'est pas. En effet, nous n'avons qu'à remplacer $e(C_i,\hat{\alpha}_A)$ et $e(C_i,\hat{\alpha}_B)$ par $(1-e(C_i,\hat{\alpha}_A))$ et $(1-e(C_i,\hat{\alpha}_B))$ pour le constater.

Corollaire. *Le poids attribué par la pondération interne (4.4) est inférieur ou égal à celui attribué par la pondération externe (4.3) pour un individu i non exposé au traitement X, c'est-à-dire :*

$$\left(\frac{p_A}{(1-e(C_i,\hat{\alpha}_A))} + \frac{p_B}{(1-e(C_i,\hat{\alpha}_B))}\right) \geq \frac{1}{1-(p_A\cdot e(C_i,\hat{\alpha}_A) + p_B\cdot e(C_i,\hat{\alpha}_B))} \tag{4.6}$$

Preuve. *Nous avons :*

$$\left(\frac{p_A}{e(C_i,\hat{\alpha}_A)} + \frac{p_B}{e(C_i,\hat{\alpha}_B)}\right) \geq \frac{1}{p_A\cdot e(C_i,\hat{\alpha}_A) + p_B\cdot e(C_i,\hat{\alpha}_B)}\,,$$

pour $0 < p_A,\ p_B,\ e(C_i,\hat{\alpha}_A),\ e(C_i,\hat{\alpha}_B) < 1.$

Ainsi,

$$\begin{aligned}
\left(\frac{p_A}{(1-e(C_i,\hat{\alpha}_A))}+\frac{p_B}{(1-e(C_i,\hat{\alpha}_B))}\right) &\geq \frac{1}{p_A\cdot(1-e(C_i,\hat{\alpha}_A))+p_B\cdot(1-e(C_i,\hat{\alpha}_B))};\\
\left(\frac{p_A}{(1-e(C_i,\hat{\alpha}_A))}+\frac{p_B}{(1-e(C_i,\hat{\alpha}_B))}\right) &\geq \frac{1}{p_A-p_A\cdot e(C_i,\hat{\alpha}_A)+p_B-p_B\cdot e(C_i,\hat{\alpha}_B)};\\
\left(\frac{p_A}{(1-e(C_i,\hat{\alpha}_A))}+\frac{p_B}{(1-e(C_i,\hat{\alpha}_B))}\right) &\geq \frac{1}{(p_A+p_B)-p_A\cdot e(C_i,\hat{\alpha}_A)-p_B\cdot e(C_i,\hat{\alpha}_B)};\\
\left(\frac{p_A}{(1-e(C_i,\hat{\alpha}_A))}+\frac{p_B}{(1-e(C_i,\hat{\alpha}_B))}\right) &\geq \frac{1}{1-(p_A\cdot e(C_i,\hat{\alpha}_A)+p_B\cdot e(C_i,\hat{\alpha}_B))}.
\end{aligned}$$

La Proposition 4.5 et son Corollaire 4.6 indiquent donc que la pondération interne réduit le poids d'un individu par rapport à la pondération externe si les scores de propension ne sont pas égaux. Nous présentons les Tableaux 4.11 et 4.12 afin de donner un aperçu de l'ampleur de la réduction du poids obtenue par l'utilisation de la pondération interne plutôt qu'externe. Nous utilisons encore un contexte à deux modèles A et B et nous présentons deux schémas de probabilités associées à ces modèles, soit le premier où le modèle A a une probabilité de 0,10 et le modèle B de 0,90 et le deuxième où les modèles A et B sont équiprobables. Nous supposons que l'individu i est exposé au traitement et calculons les poids associés à partir des équations 4.3 et 4.4.

Nous constatons qu'une importante diminution du poids se produit à l'aide de la pondération interne lorsqu'un individu présente un score de propension en contradiction avec son exposition réelle sous un modèle et que le second score utilisé pour le moyennage est plus près de l'exposition réelle (par exemple, lorsque $e(C_i,\hat{\alpha}_A) = 0{,}01$ et $e(C_i,\hat{\alpha}_B) = 0{,}5$ ou $e(C_i,\hat{\alpha}_B) = 0{,}99$, quand $X_i = 1$). Cette réduction du poids est plus grande lorsque des probabilités non négligeables sont accordées aux deux modèles, donc, dans le cas présenté ici, lorsque les modèles sont équiprobables. Nous remarquons une symétrie dans les ratios présentés (voir Tableau 4.11) : les ratios sont les mêmes si $e(C_i,\hat{\alpha}_A) = \varphi_1$ et $e(C_i,\hat{\alpha}_B) = \varphi_2$ ou $e(C_i,\hat{\alpha}_A) = \varphi_2$ et $e(C_i,\hat{\alpha}_B) = \varphi_1$ sous le même schéma de probabilité des modèles. Par contre, lorsque nous regardons la différence de

Tableau 4.11 Ratio du poids d'un individu obtenu par pondération interne vs pondération externe

(a) $p_A = 0{,}10\,;\ p_B = 0{,}90$

$\omega_{i,int}/\omega_{i,ext}$	$e(C_i,\hat{\alpha}_B) = 0{,}01$	$e(C_i,\hat{\alpha}_B) = 0{,}50$	$e(C_i,\hat{\alpha}_B) = 0{,}99$
$e(C_i,\hat{\alpha}_A) = 0{,}01$	1	0,188	0,103
$e(C_i,\hat{\alpha}_A) = 0{,}50$	0,188	1	0,958
$e(C_i,\hat{\alpha}_A) = 0{,}99$	0,103	0,958	1

(b) $p_A = 0{,}50\,;\ p_B = 0{,}50$

$\omega_{i,int}/\omega_{i,ext}$	$e(C_i,\hat{\alpha}_B) = 0{,}01$	$e(C_i,\hat{\alpha}_B) = 0{,}50$	$e(C_i,\hat{\alpha}_B) = 0{,}99$
$e(C_i,\hat{\alpha}_A) = 0{,}01$	1	0,077	0,040
$e(C_i,\hat{\alpha}_A) = 0{,}50$	0,077	1	0,892
$e(C_i,\hat{\alpha}_A) = 0{,}99$	0,040	0,892	1

Ratio du poids obtenu de la pondération interne sur le poids obtenu de la pondération externe ($\omega_{i,int}/\omega_{i,ext}$) pour un individu i exposé ($X_i = 1$). Un ratio est obtenu pour différents scores de propension en considérant deux modèles A et B dans deux schémas de probabilité de ces modèles.

Tableau 4.12 Différence du poids d'un individu obtenu par pondération externe et pondération interne

(a) $p_A = 0{,}10\,;\ p_B = 0{,}90$

$\omega_{i,ext} - \omega_{i,int}$	$e(C_i,\hat{\alpha}_B) = 0{,}01$	$e(C_i,\hat{\alpha}_B) = 0{,}50$	$e(C_i,\hat{\alpha}_B) = 0{,}99$
$e(C_i,\hat{\alpha}_A) = 0{,}01$	0	9,583	9,788
$e(C_i,\hat{\alpha}_A) = 0{,}50$	73,251	0	0,046
$e(C_i,\hat{\alpha}_A) = 0{,}99$	80,842	0,080	0

(b) $p_A = 0{,}50\,;\ p_B = 0{,}50$

$\omega_{i,ext} - \omega_{i,int}$	$e(C_i,\hat{\alpha}_B) = 0{,}01$	$e(C_i,\hat{\alpha}_B) = 0{,}50$	$e(C_i,\hat{\alpha}_B) = 0{,}99$
$e(C_i,\hat{\alpha}_A) = 0{,}01$	0	47,078	48,505
$e(C_i,\hat{\alpha}_A) = 0{,}50$	47,078	0	0,163
$e(C_i,\hat{\alpha}_A) = 0{,}99$	48,505	0,163	0

Différence du poids obtenu de la pondération externe et du poids obtenu de la pondération interne ($\omega_{i,ext} - \omega_{i,int}$) pour un individu i exposé ($X_i = 1$). Une différence est obtenue pour différents scores de propension en considérant deux modèles A et B dans deux schémas de probabilité de ces modèles.

poids, cette symétrie se retrouve seulement dans le deuxième schéma, lorsque les modèles sont équiprobables. Autrement, l'asymétrie peut être très importante. Par exemple, dans le premier schéma (a), la différence entre le poids obtenu avec la pondération externe et la pondération interne lorsque $e(C_i,\hat{\alpha}_A) = 0{,}01$ et $e(C_i,\hat{\alpha}_B) = 0{,}99$ est la suivante :

$$\begin{aligned} \omega_{i,ext} - \omega_{i,int} &= \left(\frac{0{,}10}{0{,}01} + \frac{0{,}90}{0{,}99}\right) - \frac{1}{0{,}10 \cdot 0{,}01 + 0{,}90 \cdot 0{,}99} \\ &\approx 10{,}909 - 1{,}121 \\ &\approx 9{,}788\,, \end{aligned}$$

mais, lorsque $e(C_i,\hat{\alpha}_A) = 0{,}99$ et $e(C_i,\hat{\alpha}_B) = 0{,}01$, elle est plutôt :

$$\begin{aligned} \omega_{i,ext} - \omega_{i,int} &= \left(\frac{0{,}10}{0{,}99} + \frac{0{,}90}{0{,}01}\right) - \frac{1}{0{,}10 \cdot 0{,}99 + 0{,}90 \cdot 0{,}01} \\ &\approx 90{,}101 - 9{,}259 \\ &\approx 80{,}842\,. \end{aligned}$$

Une telle réduction du poids peut avoir un énorme impact dans le calcul des estimations.

Voyons maintenant pourquoi cette diminution des poids fait en sorte que les estimateurs par pondération interne présentent une variance plus petite que les estimateurs par pondération externe. Pour examiner ce qui se produit, nous nous référons aux observations et estimations présentées dans la première situation (en considérant les modèles M1 à M14) avec les échantillons de taille $n = 300$. En regardant l'histogramme des estimations obtenues à partir de $\hat{\theta}_{ext,AIC}$ (Figure 4.2), nous constatons qu'une partie de la variance est causée par quelques échantillons qui produisent des estimations de l'effet causal très éloignées de la moyenne. En effet, nous comptons 23 échantillons sur 5000 dont l'estimation $\hat{\theta}_{ext,AIC}$ est inférieure à -2 (la moyenne est 2,347). Parmi ces échantillons, nous en sélectionnons un afin de voir de plus près pourquoi l'estimation est aussi petite et pour observer l'impact du choix de la pondération utilisée.

L'échantillon 2580, par exemple, nous fournit les estimations présentées dans le Tableau 4.13. Nous remarquons que l'estimation obtenue sous le modèle M4 est fortement erronée et qu'une grande probabilité est accordée à ce modèle. Regardons maintenant ce qui engendre une telle erreur. La Figure 4.3 montre le poids (Équation 4.2) des 300 individus

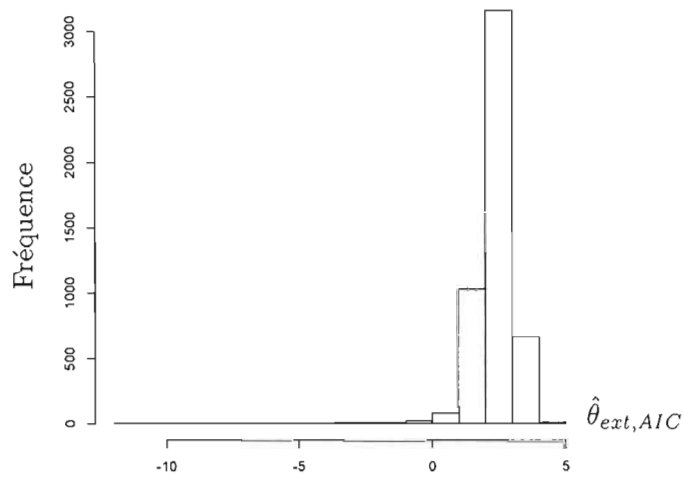


Figure 4.2 Histogramme des estimations $\hat{\theta}_{ext,AIC}$ observées à partir des 5000 réplications d'échantillon de taille $n = 300$ dans le contexte de la première situation (les modèles M1 à M14 sont inclus).

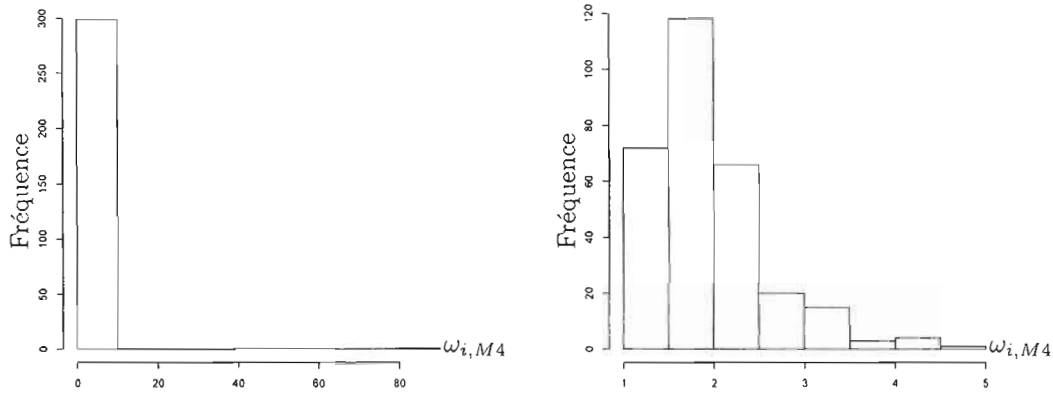
formant l'échantillon 2580 sous le modèle M4. Un individu se démarque nettement des autres car il a un poids supérieur à 80. En effet, le score de propension de cet individu sous le modèle M4 est $e(C_{93,2580},\hat{\alpha}_{M4,2580}) = 0.988$, mais son exposition est $X_{93,2580} = 0$. Tel que mentionné plus haut, la contribution d'un individu à l'estimation provient de son poids, mais également de la valeur de sa variable réponse Y. Il se trouve que cet individu a une valeur très écartée de la moyenne pour sa variable Y puisqu'elle est de $Y_{93,2580} = 34{,}7$. Il est possible de voir dans la Figure 4.4 où se situe cette valeur par rapport aux autres individus non exposés. Cet aussi grand poids combiné à une telle valeur de Y apporte une contribution qui est en grande partie la cause de l'erreur de l'estimation $\hat{\theta}_{M4}$. Puisque les probabilités attribuées par le AIC sont principalement accordées aux modèles M1, M4, M8 et M11, l'estimation obtenue avec l'estimateur par pondération externe $\hat{\theta}_{ext,AIC}$ présente aussi une erreur importante.

Tableau 4.13 Estimations obtenues en considérant tous les modèles (M1 à M14) pour l'échantillon 2580, de taille $n = 300$

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
Estimation	−1,766	3,074	1,132	−6,508	2,877	0,582	−8,303
$p_{AIC} \times 10^2$	18,120	0,279	5,270	37,495	0,076	0,925	6,017

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
Estimation	−0,256	3,505	2,191	−2,789	3,333	1,701	−4.350
$p_{AIC} \times 10^2$	9,759	0,140	2,302	16,301	0,041	0,430	2,845

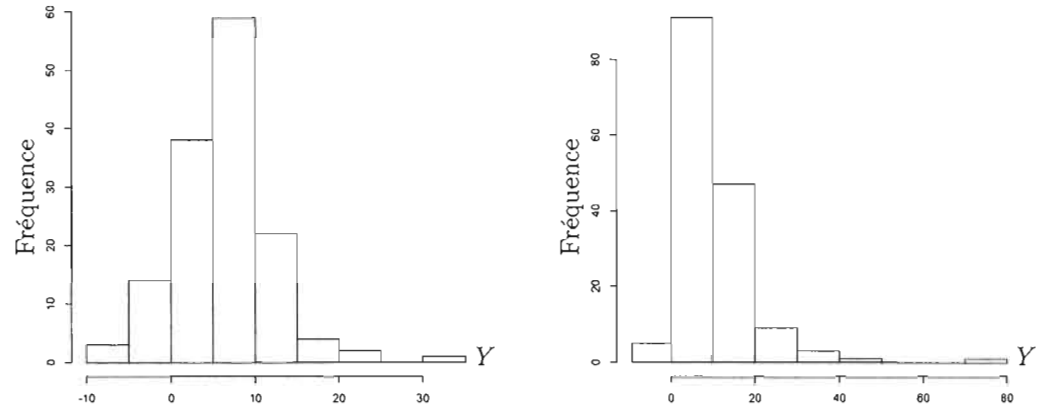
$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
−6.508	3.074	−3.723	1.893	−1.989	2.584



(a) tous les individus

(b) tous les individus, sauf celui qui a un poids supérieur à 80

Figure 4.3 Histogrammes des poids observés sous le modèle M4 dans l'échantillon 2580 de taille $n = 300$.



(a) $X = 0$ (b) $X = 1$

Figure 4.4 Histogrammes des valeurs observées de la variable réponse Y pour les individus exposés ($X = 1$) et non exposés ($X = 0$) dans l'échantillon 2580 de taille $n = 300$.

Nous utilisons maintenant ce même individu afin d'illustrer comment évolue son poids à travers chacun des modèles et surtout sous chacun des deux types de pondération (voir Tableau 4.14). C'est sous les modèles M4 et M7 que l'individu présente les plus grands poids. Sous les autres modèles qui ont une probabilité supérieure à 5%, il a des poids beaucoup plus petits. C'est dans des cas comme celui-là, c'est-à-dire quand un individu a un plus grand poids sous un modèle, mais que sous d'autres modèles importants son poids est plus petit, que le choix de la pondération interne a le plus d'impact. Lorsqu'un individu dans cette situation a également une variable Y qui s'éloigne de la moyenne des individus dans la même catégorie d'exposition que lui, la diminution du poids apportée par la pondération interne peut modifier considérablement l'estimation de l'effet causal. L'estimation obtenue suite à la pondération interne est moins extrême que sous la pondération externe ; elle se situe plus proche de la moyenne. La variance des estimations obtenues sous les 5000 réplications étant une estimation de la variance de l'estimateur, nous obtenons donc que la variance des estimateurs par pondération

Tableau 4.14 Poids de l'individu 93 de l'échantillon 2580 sous chacun des modèles, ainsi qu'avec la pondération externe et interne en version AIC en considérant tous les modèles (M1 à M14)

	ω_{M1}	ω_{M2}	ω_{M3}	ω_{M4}	ω_{M5}	ω_{M6}	ω_{M7}
	39,36	4,63	19,00	81,64	5,29	22,85	96,18
$p_{AIC} \times 10^2$	18,120	0,279	5,270	37,495	0,076	0,925	6,017

	ω_{M8}	ω_{M9}	ω_{M10}	ω_{M11}	ω_{M12}	ω_{M13}	ω_{M14}
	29,72	3,63	12,89	52,86	4,19	16,02	65,20
$p_{AIC} \times 10^2$	9,759	0,140	2,302	16,301	0,041	0,430	2,845

$\omega_{ext,AIC}$	$\omega_{int,AIC}$
58,50	43,32

interne est plus petite que la variance des estimateurs par pondération externe.

4.2.2 Une valeur maximale pour le poids

Comme il a été mentionné précédemment, nous avons imposé une valeur maximale au poids qui pouvait être attribué à un individu. Le poids d'un individu dépend à la fois de son attribution au traitement et de son score de propension. C'est donc sur le score de propension que nous imposons une borne afin que le poids d'un individu ne dépasse pas la valeur de 100. Nous portons ici attention aux impacts d'une telle décision en comparant les résultats obtenus avec et sans poids maximum. Le fait d'imposer un plafond au poids des individus induit systématiquement un biais de l'estimateur si au moins un individu voit son poids tronqué, et ce, même sous le vrai modèle de prédiction. Nous avons pris la décision d'appliquer la troncation des poids en espérant que le biais qui peut en découler soit plus petit que celui que nous pouvons rencontrer lorsqu'un individu a un poids extrême. Nous voulons maintenant considérer l'idée de fixer une valeur minimale/maximale (selon que l'individu est exposé ($X = 1$) ou non ($X = 0$)) au

score de propension utilisé dans l'estimateur par pondération interne, mais seulement une fois que le score ait été pondéré par la probabilité de chacun des modèles considérés. Ainsi, aucun plafond n'est imposé au score de propension sous chacun des m modèles, mais plutôt au score de propension moyenné (équations 3.6 et 3.9). Dans cette situation, un individu voit donc son score de propension être tronqué au plus une seule fois. Nous espérons alors qu'ils produisent un plus petit biais systématique que $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$ avec une troncation sous chacun des modèles, tout en gardant l'avantage d'une variance réduite. Nous utilisons le même exemple qu'à la sous-section 4.2.1, c'est-à-dire 5000 réplications d'échantillon de taille $n = 300$ dans le contexte de la première situation (M1 à M14 sont inclus), pour observer les différences obtenues avec les trois possibilités de troncation proposées : (1) poids maximal de 100 sous chacun des modèles / (2) aucun poids maximal / (3) poids maximal de 100 basé sur le score de propension moyenné.

Lorsque nous comparons les biais obtenus sous les trois méthodes (Tableau 4.15), nous n'observons pas une tendance générale pour tous les modèles. Toutefois, nous remarquons le fait que les estimateurs spécifiés sous le bon modèle et ceux qui l'emboîtent (M1, M7, M8 et M14) voient tous leur biais diminuer lorsque nous n'imposons pas de plafond aux poids. De même, lorsque nous n'imposons pas de plafond, les six estimateurs ECMP présentent un biais presque égal ou plus petit à celui observé avec un plafond. Rappelons-nous que dans cet exemple le critère AIC favorise surtout les modèles M1, M4, M7 et M8 et c'est pourquoi une diminution du biais est observable principalement pour les estimateurs ECMP qui sont basés sur le critère AIC. Quant au biais obtenu par $\hat{\theta}_{int,BIC}$ lorsqu'il n'y a aucune troncation (2), il est le même que le biais obtenu avec la troncation appliquée lors de la pondération interne (3) puisqu'aucun score de propension moyenné par les probabilités basées sur le critère BIC (équation 3.9) ne produit un poids supérieur à 100. Pour ce qui est de $\hat{\theta}_{int,AIC}$ lorsque la stratégie (3) est appliquée pour la troncation, nous obtenons un biais à mi-chemin entre ceux obtenus avec un plafond applicable sous chacun des modèles (1) et sans plafond (2).

Tableau 4.15 Biais sous chacune des possibilités de troncation à partir des 5000 réplications d'échantillon de taille $n = 300$ dans le contexte de la première situation (les modèles M1 à M14 sont inclus)

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
Biais (1)	0,211	1,477	1,112	0,423	1,384	0,887	0,214
Biais (2)	0,205	1,477	1,137	0,418	1,384	0,903	0,203

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
Biais (1)	0,206	1,473	1,111	0,421	1,378	0,882	0,210
Biais (2)	0,197	1,473	1,143	0,415	1,378	0,902	0,195

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais (1)	0,290	1,007	0,347	0,974	0,357	0,911
Biais (2)	0,280	1,004	0,337	0,973	0,350	0,910
Biais (3)	-	-	-	-	0,355	0,910

Sous chacun des modèles, l'utilisation d'un plafond pour le poids entraîne une réduction considérable de la variance de l'estimateur (voir Tableau 4.16) lorsqu'au moins un individu voit son poids tronqué. C'est la raison pour laquelle nous avons pris la décision d'imposer ce plafond. Comme nous l'avons mentionné dans la sous-section 4.2.1, les individus qui ont un poids important peuvent grandement influencer l'estimation de l'effet et produire un résultat éloigné de la moyenne (tel qu'obtenu, par exemple, avec l'échantillon 2580 que nous avons regardé plus particulièrement). Bien qu'un tel résultat puisse aider à réduire le biais de l'estimateur, il contribue aussi beaucoup à l'augmentation de la variance. La variance obtenue avec $\hat{\theta}_{int,AIC}$ lorsque la troncation est faite lors de la pondération interne (3) est assez près de celle obtenue avec $\hat{\theta}_{int,AIC}$ lorsque la troncation est appliquée sous chacun des modèles (1). La petite différence doit être due au fait que des individus qui se voient attribués un poids de 100 au moment d'effectuer la pondération interne ont probablement un poids légèrement plus petit lorsque

Tableau 4.16 Variance sous chacune des possibilités de troncation à partir des 5000 réplications d'échantillon de taille $n = 300$ dans le contexte de la première situation (les modèles M1 à M14 sont inclus)

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
Var (1)	0,789	0,495	1,118	0,683	0,672	0,967	0,828
Var (2)	0,985	0,495	1,961	0,866	0,672	1,556	1,405

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
Var (1)	0,759	0,409	1,082	0,624	0,589	0,893	0,780
Var (2)	1,073	0,409	2,512	0,846	0,589	1,630	1,618

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Var (1)	0,799	0,732	0,673	0,537	0,521	0,363
Var (2)	1,138	0,878	1,104	0,678	0,718	0,370
Var (3)	-	-	-	-	0,558	0,370

leurs scores de propension ont déjà été bornés sous chaque modèle. Pour ce qui est de l'estimateur $\hat{\theta}_{int,BIC}$ sous la stratégie de pondération (3), la variance obtenue est la même qu'avec $\hat{\theta}_{int,BIC}$ sans pondération (2) puisqu'aucun individu n'est affecté par l'application d'un plafond une fois son score de propension moyenné.

Lorsque nous prenons en considération à la fois le biais et la variance, à travers l'EQM, nous obtenons que sous chacun des modèles ainsi que pour tous les estimateurs ECMP, l'utilisation d'un plafond, quel qu'il soit, s'avère avantageuse. La réduction de la variance est suffisante pour contrebalancer l'augmentation du biais. Les estimateurs $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$ donnent des résultats semblables quand la troncation est appliquée lors de la pondération interne (3) ou sous chacun des modèles (1). Le choix du moment où nous effectuons la troncation des poids (sous chacun des modèles ou seulement lors de l'estimation par la pondération interne) n'est donc pas une décision qui a un grand impact sur les résultats observés.

Tableau 4.17 EQM sous chacune des possibilités de troncation à partir des 5000 réplications d'échantillon de taille $n = 300$ dans le contexte de la première situation (les modèles M1 à M14 sont inclus)

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
EQM (1)	0,833	2,678	2,354	0,862	2,588	1,753	0,874
EQM (2)	1,026	2,678	3,253	1,041	2,588	2,372	1,449

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
EQM (1)	0,801	2,579	2,316	0,801	2,486	1,670	0,824
EQM (2)	1,112	2,579	3,818	1,018	2,486	2,443	1,655

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
EQM (1)	0,883	1,746	0,794	1,486	0,648	1,192
EQM (2)	1,216	1,885	1,217	1,625	0,840	1,198
EQM (3)	-	-	-	-	0,684	1,198

4.3 Deuxième exemple

Nous présentons maintenant un exemple plus complexe afin de vérifier si les estimateurs moyennés offrent des performances aussi avantageuses que dans le premier exemple. Bien que ce second exemple soit aussi simulé par ordinateur, nous tentons de reproduire de façon relativement simple une situation réelle.

L'effet causal à estimer dans cet exemple est l'effet d'un traitement X sur la capacité pulmonaire totale[1] (CPT). Pour chaque individu, nous avons les informations suivantes :

- Son statut de fumeur (F) : actuellement fumeur (FA) / ancien fumeur (FD) / n'a jamais fumé (FJ) ;
- Le nombre de cigarettes fumées par jour (CJ) ;

[1] La capacité pulmonaire totale est le volume maximum d'air présent dans la poitrine lorsqu'elle est gonflée à fond. (Capital Souffle, 2010)

- Le nombre de mois où l'individu a été fumeur (MF) ;
- Le nombre de mois écoulés depuis qu'un ancien fumeur a arrêté de fumer (MA) ;
- L'âge (années) ;
- Le sexe : Femme / Homme ;
- La taille (centimètres) ;
- Le poids (kilogrammes) ;
- Le niveau d'activité physique pratiquée (AP) : nul (APN)/ moyen (APM) / élevé (APE) ;
- La pression systolique (millimètres de mercure) (PS) ;
- La pression diastolique (millimètres de mercure) (PD).

Nous savons que le tabagisme, l'âge, le sexe, le poids, la taille, l'activité physique et la pression sanguine sont des facteurs confondants. Cela est représenté par le diagramme de la Figure 4.5.

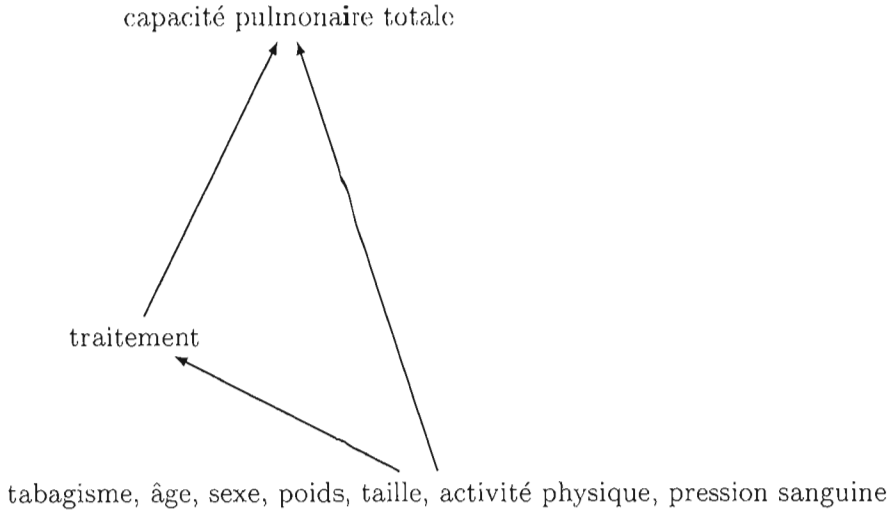


Figure 4.5 Graphe orienté acyclique représentant les liens de causalité entre la capacité pulmonaire totale, le traitement, le tabagisme, l'âge, le sexe, le poids, la taille, l'activité physique et la pression sanguine.

La particularité de ce problème vient du fait qu'une variable confondante peut être décrite de différentes façons. Par exemple, l'effet du tabagisme peut provenir simplement du fait qu'un individu soit actuellement un fumeur ou un non-fumeur, mais peut aussi changer selon le nombre de cigarettes fumées par jour, le temps écoulé depuis que l'individu a commencé à fumer ou a cessé de fumer et/ou le nombre de paquets-années[2] (PA) fumés. Nous pouvons également penser que le poids et la taille n'ont pas un effet linéaire chacun, mais que c'est plutôt leur interaction, sous la forme de l'indice de masse corporel[3] (IMC), qui a un effet. La pression sanguine peut aussi s'exprimer sous la forme de la pression sanguine moyenne (PSM) qui se calcule de la façon suivante : $PSM = \frac{PS+2PD}{3}$.

Pour cet exemple, nous avons imposé quelques limites aux valeurs que peuvent prendre certaines variables. Tous les individus présentés dans les échantillons ont entre 18 et 90 ans et nous considérons le nombre de mois où un individu a fumé, et possiblement arrêté de fumer, depuis ses 14 ans. Nous avons généré quelques variables qui sont dépendantes les unes des autres. La taille dépend du sexe. Le poids et l'IMC dépendent également du sexe, mais aussi de l'âge. Le niveau d'activité physique pratiquée dépend de l'âge et du statut de fumeur de l'individu. La pression sanguine dépend de l'âge, du sexe, du niveau d'activité physique et de l'IMC.

Appelons C l'ensemble des variables observées et calculées dans cet exemple, à l'exception de l'attribution au traitement X et de la variable résultat Y. La probabilité pour un individu d'être exposé au traitement X est déterminée par l'équation suivante :

$$P[X = 1|C] = \frac{\exp(g(C))}{1 + \exp(g(C))};$$

---

[2]Le paquet-année est une façon de mesurer la quantité de cigarettes fumées sur une longue période de temps. Il est calculé en multipliant le nombre de paquets de cigarettes fumées par jour par le nombre d'années pendant lesquelles l'individu a fumé.

[3]L'IMC se calcule ainsi : kg/m^2, où kg est le poids en kilogramme et m est la taille en mètre.

où

$$\begin{aligned} g(C) = & - 0{,}5 + 0{,}75I(F = FA) + 0{,}25I(F = FD) + 0{,}05PA - 0{,}005MA \\ & + 0{,}02\hat{a}ge - 0{,}25I(sexe = Femme) + 0{,}05IMC - 0{,}25I(IMC \geq 30) \\ & + 0{,}4I(AP = APM) + 0{,}6I(AP = APE) - 0{,}02PSM\,; \end{aligned}$$

et où $I(condition) = 1$ si la *condition* est vraie ; 0 sinon.

Tout comme dans le premier exemple, cette probabilité, calculée pour chacun des individus, est utilisée comme paramètre d'une loi Bernouilli pour générer aléatoirement l'attribution au traitement. La CPT est ensuite générée à partir de l'équation suivante :

$$\begin{aligned} CPT = & \; 1{,}7 - 0{,}2I(F = FA) - 0{,}0175PA + 0{,}0013MA - 0{,}01\hat{a}ge \\ & -0{,}3I(sexe = Femme) + 0{,}035taille - 0{,}0075poids \\ & -0{,}1I(IMC \geq 25 \;\&\; IMC < 30) - 0{,}2I(IMC \geq 30) \\ & +0{,}2I(AP = APM) + 0{,}45I(AP = APE) - 0{,}01PSM + 0{,}25X + \epsilon\,; \end{aligned}$$

où $\epsilon \sim \mathcal{N}(0,\, 0{,}25^2)$ est un bruit aléatoire. Ainsi, environ 80 à 85% de la variabilité de la variable résultat CPT est due aux variables mesurées chez les individus et le reste au hasard (ϵ). Le diagramme de la Figure 4.6 représente les liens causaux entre les variables confondantes, le traitement et la capacité pulmonaire totale, mais en précisant la forme sous laquelle les variables confondantes conceptuelles interviennent.

Puisque cet exemple contient plus de variables et est plus complexe que le premier exemple, nous simulons des échantillons de plus grandes tailles. Nous utilisons ici des échantillons de 500, 5000 et 10000 individus. Pour chacune de ces tailles, nous générons 5000 jeux de données. Nous souhaitons toujours voir comment performent les estimateurs sous différents ensembles de modèles utilisés pour pondérer les estimations. Les modèles utilisés sont représentés dans le Tableau 4.18. Le premier modèle, L1, représente le vrai modèle, c'est-à-dire le modèle qui est utilisé pour générer l'attribution au traitement. Notons que chacun des modèles L2 à L9 et L14 à L17 intervient sur une seule variable confondante à la fois. Par exemple, les modèles L2 à L9 diffèrent du vrai modèle

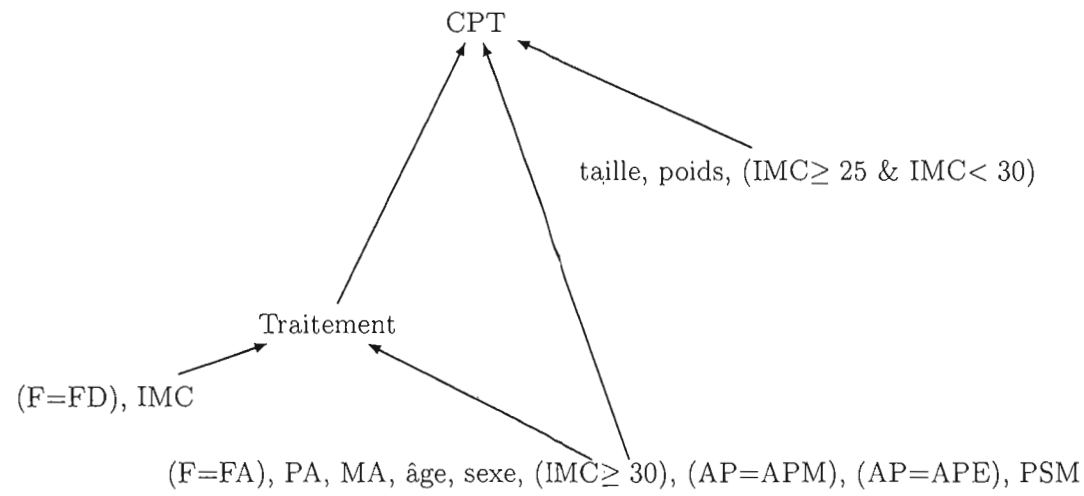


Figure 4.6 Diagramme représentant les liens causaux entre le traitement, la capacité pulmonaire totale et les différentes représentations des variables confondantes.

M1 seulement par les variables qui ont trait au tabagisme. Les modèles L10 à L13, L18 et L19 interviennent à la fois sur la taille et le poids, alors que L20, L21 et L22 agissent sur plus de trois variables confondantes simultanément. Tout comme dans le premier exemple, un poids plafond de 100 est utilisé sous chacun des modèles. Nous analysons les résultats obtenus sous les modèles individuellement, ainsi que ceux obtenus des estimateurs ECMP.

Dans le Tableau 4.19, nous pouvons observer que les individus des groupes exposé et non exposé ont des caractéristiques et habitudes de vie différentes. Nous y présentons les moyennes et proportions observées pour certaines variables chez les deux groupes d'exposition et les valeurs p obtenues du test de Student ou du chi-deux (χ^2). Nous remarquons que les variables qui ont trait au tabagisme ont une valeur p très significative, de même pour l'âge et le sexe. Nous pouvons aussi affirmer que les moyennes des autres variables confondantes sont différentes entre les deux groupes, à l'exception des variables représentant la taille, le poids et l'IMC chez la femme, qui ont des valeurs p non significatives. Lorsque nous estimons l'effet causal en calculant simplement la différence entre la capacité pulmonaire des gens exposés et non exposés, nous obtenons que le

Tableau 4.18 Différents modèles considérés dans les estimations

	F=FA	F=FD ou FA	CJ	MF	PA	MA	age	sexe=Femme	taille	poids	IMC	taille×poids	IMC≥20 & IMC<25	IMC≥25 & IMC<30	IMC≥30	AP=APM ou APE	AP=APE	PS	PD	PSM
L1	♦	♦			♦	♦	♦	♦			♦				♦	♦	♦			♦
L2	♦				♦	♦	♦	♦			♦				♦	♦	♦			♦
L3		♦			♦	♦	♦	♦			♦				♦	♦	♦			♦
L4	♦	♦					♦	♦			♦				♦	♦	♦			♦
L5	♦						♦	♦			♦				♦	♦	♦			♦
L6		♦					♦	♦			♦				♦	♦	♦			♦
L7	♦	♦			♦		♦	♦			♦				♦	♦	♦			♦
L8	♦	♦				♦	♦	♦			♦				♦	♦	♦			♦
L9	♦	♦	♦	♦	♦	♦	♦	♦			♦				♦	♦	♦			♦
L10	♦	♦			♦	♦	♦	♦	♦	♦					♦	♦	♦			♦
L11	♦	♦			♦	♦	♦	♦			♦					♦	♦			♦
L12	♦	♦			♦	♦	♦	♦	♦	♦			♦	♦		♦	♦			♦
L13	♦	♦			♦	♦	♦	♦						♦	♦	♦	♦			♦
L14	♦	♦			♦	♦	♦	♦			♦				♦	♦				♦
L15	♦	♦			♦	♦	♦	♦			♦				♦		♦			♦
L16	♦	♦			♦	♦	♦	♦			♦				♦	♦	♦		♦	
L17	♦	♦			♦	♦	♦	♦			♦				♦	♦	♦	♦		
L18	♦	♦			♦	♦	♦	♦			♦	♦			♦	♦	♦			♦
L19	♦	♦			♦	♦	♦	♦				♦	♦	♦		♦	♦			♦
L20	♦				♦	♦	♦	♦							♦	♦	♦			♦
L21	♦		♦	♦			♦	♦	♦	♦				♦	♦	♦		♦	♦	
L22	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦		♦	♦	♦	♦	♦	♦	

Tableau 4.19 Comparaisons des moyennes et proportions observées dans les groupes exposé et non exposé de l'échantillon 1930 de taille $n = 5000$

Variable	Traitement $X = 0$ ($n = 1980$)	$X = 1$ ($n = 3020$)	valeur p*
F=FA	0,20	0,36	**< 0,001**
F=DJ	0,53	0,48	**< 0,001**
CJ|(F=DJ ou F=FA)	15,11	15,99	**< 0,001**
MF|(F=DJ ou F=FA)	144,94	210,44	**< 0,001**
PA|(F=DJ ou F=FA)	8,68	14,10	**< 0,001**
MA|(F=DJ)	115,23	73,39	**< 0,001**
âge	43,43	45,64	**< 0,001**
sexe=Femme	0,53	0,48	**< 0,001**
taille|sexe=Femme	164,08	163,99	0,580
taille|sexe=Homme	176,03	175,93	0,597
poids|sexe=Femme	69,25	69,67	0,416
poids|sexe=Homme	83,03	84,28	**0,026**
IMC|sexe=Femme	25,71	25,90	0,296
IMC|sexe=Homme	26,78	27,21	**0,011**
IMC≥30|sexe=Femme	0,23	0,22	0,329
IMC≥30|sexe=Homme	0,23	0,24	0,671
AP=APM ou AP=APE	0,55	0,59	**0,001**
AP=APE	0,18	0,20	0,051
(AP=APM ou AP=APE)|(F=FD ou F=FA)	0,53	0,57	**0,006**
PAM	95,25	94,54	**0,011**

*Valeur p provenant du test de Student pour les variables continues ou du test du χ^2 pour les variables catégoriques.

traitement augmente la CPT, mais d'une façon moindre qu'elle le fait réellement. En fait, en moyenne à travers les 5000 réplications de jeux de données de 10000 individus, nous observons $\hat{\mu}_{X=1} - \hat{\mu}_{X=0} = 0{,}131$, avec une variance de 0,0002. La coefficient utilisé devant la valeur d'exposition lors de la génération des données est plutôt de 0,25.

Nous étudions en premier lieu un cas où le vrai modèle fait partie de l'ensemble des modèles considérés puisque nous utilisons tous les modèles. Dans un deuxième temps, nous examinons différents sous-ensembles de modèles déterminés par le nombre de variables qui les forment.

4.3.1 Première situation

La première situation qui nous intéresse est le cas où l'ensemble de tous les modèles proposés est utilisé. Les Tableaux 4.20 et 4.21 présentent les résultats obtenus sous chacun des modèles individuellement ainsi qu'avec les estimateurs ECMP lorsque que nous considérons des échantillons de taille $n = 500$ et $n = 5000$ respectivement. Le premier fait à remarquer est que les modèles L2 à L8 et L20 produisent les plus grands biais. Ce sont des modèles qui omettent une ou plusieurs variables représentant le tabagisme. Tout comme nous le devinions dans le Tableau 4.19, le tabagisme est une variable confondante d'une très grande importance. Notons aussi qu'à l'exception de ces huit modèles, les biais sont petits (inférieurs à 0,02) et que l'EQM est alors très près de la variance.

Lorsque nous regardons les résultats obtenus avec $n = 500$, nous remarquons que la plupart des modèles présentent des variances du même ordre de grandeur soit d'environ 3×10^{-2}, mais les variances obtenues des modèles L4, L5, L6, L8 et L21 sont plus petites. Du côté des estimateurs ECMP, alors que les estimations basées sur le critère AIC sont construites à partir de nombreux modèles ayant chacun une probabilité ne dépassant pas 11%, le BIC accorde près de la moitié du poids au modèle L20. Ce modèle inclut les variables confondantes sous les formes qui ont un effet à la fois sur le traitement et la CPT, tel que représenté par la Figure 4.6. Peu de poids est accordé au modèle L1 malgré que ce soit le vrai modèle. Il ne semble pas y avoir de tendance

Tableau 4.20 Résultats obtenus en considérant tous les modèles (L1 à L22) avec $n = 500$ et $ech = 5000$

	$\hat{\theta}_{L1}$	$\hat{\theta}_{L2}$	$\hat{\theta}_{L3}$	$\hat{\theta}_{L4}$	$\hat{\theta}_{L5}$	$\hat{\theta}_{L6}$	$\hat{\theta}_{L7}$	$\hat{\theta}_{L8}$
Biais $\times 10^2$	0,165	-2,418	2,414	-9,777	-9,774	-17,022	-5,817	-3,840
Variance $\times 10^2$	3,252	3,839	3,239	0,691	0,630	0,394	3,697	1,556
EQM $\times 10^2$	3,252	3,897	3,298	1,647	1,586	3,292	4,036	1,703
$\overline{p_{AIC}} \times 10^2$	6,181	9,964	5,650	0,002	0,003	$< 0,001$	1,100	0,374
$\overline{p_{BIC}} \times 10^2$	0,858	9,915	6,853	0,015	0,124	0,002	1,114	0,539

	$\hat{\theta}_{L9}$	$\hat{\theta}_{L10}$	$\hat{\theta}_{L11}$	$\hat{\theta}_{L12}$	$\hat{\theta}_{L13}$	$\hat{\theta}_{L14}$	$\hat{\theta}_{L15}$	$\hat{\theta}_{L16}$
Biais $\times 10^2$	0,617	0,147	0,284	0,188	-0,056	0,796	1,734	0,192
Variance $\times 10^2$	3,318	3,326	3,124	3,420	3,170	3,033	2,846	3,241
EQM $\times 10^2$	3,322	3,327	3,125	3,421	3,170	3,039	2,876	3,242
$\overline{p_{AIC}} \times 10^2$	4,282	4,387	9,761	3,446	5,775	10,288	5,711	6,255
$\overline{p_{BIC}} \times 10^2$	0,019	0,121	9,657	0,015	0,900	11,583	6,553	0,867

	$\hat{\theta}_{L17}$	$\hat{\theta}_{L18}$	$\hat{\theta}_{L19}$	$\hat{\theta}_{L20}$	$\hat{\theta}_{L21}$	$\hat{\theta}_{L22}$
Biais $\times 10^2$	0,207	0,178	-0,573	-3,496	-0,517	0,508
Variance $\times 10^2$	3,250	3,347	3,325	3,565	1,451	3,879
EQM $\times 10^2$	3,250	3,347	3,328	3,687	1,454	3,881
$\overline{p_{AIC}} \times 10^2$	6,277	4,403	4,451	9,974	0,268	1,446
$\overline{p_{BIC}} \times 10^2$	0,879	0,107	0,118	49,739	0,021	$< 0,001$

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{inl,AIC}$	$\hat{\theta}_{inl,BIC}$
Biais $\times 10^2$	-0,196	-1,901	-0,060	-1,168	1,359	-0,033
Variance $\times 10^2$	3,457	3,438	3,228	3,246	3,079	3,124
EQM $\times 10^2$	3,457	3,474	3,228	3,260	3,098	3,124

générale dans la valeur du biais sous les estimateurs ECMP. Au niveau de la variance cependant, nous observons que les estimateurs moyennés ont une plus petite variance que les estimateurs employant le modèle ayant le plus petit AIC ou BIC. Tout comme nous l'avons observé dans le premier exemple, la variance obtenue de la pondération interne est plus petite que celle provenant de la pondération externe. Par conséquent ici, l'EQM des estimateurs moyennés à l'intérieur est également plus petite. Pour les trois types d'estimateurs ECMP, l'EQM des estimateurs avec la déclinaison AIC est plus petite qu'avec le BIC. Nous ne pouvons cependant pas accorder cet écart uniquement au biais ou à la variance puisque la situation est différente pour chacun des estimateurs. L'estimateur $\hat{\theta}_{AIC}$ présente un plus petit biais, mais une plus grande variance que $\hat{\theta}_{BIC}$. Avec la pondération externe, l'utilisation du AIC apporte à la fois un plus petit biais et une plus petite variance alors qu'avec la pondération interne, un plus grand biais, mais une plus petite variance sont associés au AIC.

Lorsque nous augmentons la taille des échantillons à 5000, plusieurs des modèles voient leurs probabilités diminuer considérablement. Entre autres, remarquons que le AIC et le BIC n'accordent presque plus d'importance au modèle L20. Les modèles L3 à L8, L15, L21 et L22 n'ont eux aussi pratiquement plus d'impact sur les estimateurs ECMP à cause de la petite probabilité qui leur est associée. Le vrai modèle, L1, gagne en importance, mais pas suffisamment pour que les estimateurs ECMP soient presque entièrement déterminés par ce modèle, tel que c'était le cas avec d'aussi grands échantillons dans le premier exemple. Le fait que les relations entre les variables soient plus complexes et que les modèles soient relativement semblables entre eux semble rendre la détection du bon modèle plus difficile. Pour l'ensemble des modèles, les biais ne sont pas très différents de ce qu'ils sont avec $n = 500$, mais les variances ont beaucoup diminué. La redistribution des probabilités accordées par le AIC et le BIC fait en sorte que les estimateurs ECMP ont tous un très petit biais. Leur EQM est donc maintenant égale (au niveau de précision que nous employons) à la variance dans tous les cas. Nous remarquons également que l'écart entre les EQM des estimations basées sur le AIC et sur le BIC est plus petit. Nous avons maintenant que pour les trois types d'estimateurs ECMP,

Tableau 4.21 Résultats obtenus en considérant tous les modèles (L1 à L22) avec $n = 5000$ et $ech = 5000$

	$\hat{\theta}_{L1}$	$\hat{\theta}_{L2}$	$\hat{\theta}_{L3}$	$\hat{\theta}_{L4}$	$\hat{\theta}_{L5}$	$\hat{\theta}_{L6}$	$\hat{\theta}_{L7}$
Biais $\times 10^2$	0,168	-2,277	2,396	-9,688	-9,720	-17,009	-6,181
Variance $\times 10^2$	0,213	0,272	0,210	0,044	0,042	0,025	0,275
EQM $\times 10^2$	0,213	0,323	0,268	0,983	0,987	2,918	0,657
$\overline{p_{AIC}} \times 10^2$	14,124	6,745	0,014	$< 0,001$	$< 0,001$	$< 0,001$	$< 0,001$
$\overline{p_{BIC}} \times 10^2$	4,174	27,205	0,114	$< 0,001$	$< 0,001$	$< 0,001$	$< 0,001$

	$\hat{\theta}_{L8}$	$\hat{\theta}_{L9}$	$\hat{\theta}_{L10}$	$\hat{\theta}_{L11}$	$\hat{\theta}_{L12}$	$\hat{\theta}_{L13}$	$\hat{\theta}_{L14}$	$\hat{\theta}_{L15}$
Biais $\times 10^2$	-3,581	0,212	0,124	0,317	0,177	-0,127	0,787	1,587
Variance $\times 10^2$	0,100	0,205	0,212	0,207	0,212	0,207	0,208	0,191
EQM $\times 10^2$	0,228	0,205	0,212	0,208	0,212	0,207	0,214	0,216
$\overline{p_{AIC}} \times 10^2$	$< 0,001$	7,599	8,302	6,811	5,447	2,819	6,994	0,033
$\overline{p_{BIC}} \times 10^2$	$< 0,001$	0,009	0,163	27,424	0,006	1,202	28,773	0,220

	$\hat{\theta}_{L16}$	$\hat{\theta}_{L17}$	$\hat{\theta}_{L18}$	$\hat{\theta}_{L19}$	$\hat{\theta}_{L20}$	$\hat{\theta}_{L21}$	$\hat{\theta}_{L22}$
Biais $\times 10^2$	0,193	0,219	0,161	-0,566	-3,516	-0,402	0,192
Variance $\times 10^2$	0,213	0,211	0,212	0,211	0,258	0,089	0,207
EQM $\times 10^2$	0,213	0,212	0,213	0,215	0,382	0,090	0,207
$\overline{p_{AIC}} \times 10^2$	14,254	12,572	8,941	3,606	0,051	$< 0,001$	1,688
$\overline{p_{BIC}} \times 10^2$	4,334	3,876	0,154	0,092	2,254	$< 0,001$	$< 0,001$

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais $\times 10^2$	0,118	-0,161	0,091	-0,177	0,247	0,002
Variance $\times 10^2$	0,214	0,226	0,211	0,219	0,210	0,218
EQM $\times 10^2$	0,214	0,226	0,211	0,219	0,210	0,218

l'estimation basée sur le AIC donne de meilleurs résultats au niveau de la variance et de l'EQM. Seul l'estimateur pondéré à l'interne produit un biais plus petit avec la version BIC que AIC. Avec cette taille échantillonnale, les EQM obtenues par la pondération externe et interne sont presque les mêmes lorsqu'elles découlent du même critère d'ajustement. Les estimateurs pondérés externes ou internes se démarquent cependant des estimateurs $\hat{\theta}_{AIC}$ et $\hat{\theta}_{BIC}$ par de plus petites EQM.

Des échantillons de 10000 individus ont aussi été simulés, mais les résultats obtenus ne sont pas présentés ici parce qu'ils sont semblables à ceux obtenus avec les échantillons de 5000 individus. Nous pouvons résumer les observations comme suit. Chacun des estimateurs présente une variance réduite d'environ la moitié de ce qu'elle était avec $n = 5000$. Le modèle L1 obtient 18,7% en probabilité par le AIC, étant désormais le modèle le plus important et 12,1% par le BIC, qui favorise plutôt les modèles L2, L11 et L14. Tout comme dans le cas précédent avec $n = 5000$, les déclinaisons AIC des estimateurs ECMP présentent de plus petites EQM. Pour chacun de ceux-ci cependant, nous avons maintenant que le biais engendré par la déclinaison AIC est plus grand que celui produit par le BIC, mais que la variance est plus petite.

4.3.2 Deuxième situation

Dans un second temps, nous voulons étudier le comportement des estimateurs à l'aide de différents sous-ensembles de modèles. Les sous-ensembles ont été conçus en fonction du nombre de variables incluses dans les modèles. Nous nous intéressons aux résultats obtenus lorsque nous considérons les sous-ensembles suivants :

(a) L4, L5, L6 et L20, qui incluent 9 variables et moins ;

(b) L2, L3, L7, L8, L11, L14 et L15, qui incluent 10 variables ;

(c) L1, L13, L16 et L17, qui incluent 11 variables ;

(d) L10, L18, L19 et L21, qui incluent 12 variables ;

(e) L9, L12 et L22, qui incluent 13 variables et plus.

Pour cette situation aussi, nous avons appliqué les estimateurs sur les échantillons de taille $n = 500$, $n = 5000$ et $n = 10000$. Toutefois, comme les résultats obtenus à partir des échantillons de 5000 individus suffisent à la compréhension des comportements pour n grand, nous ne présentons pas les résultats pour $n = 10000$.

Les modèles de l'ensemble (a) omettent deux variables (L4, L20) ou trois variables (L5, L6) par rapport au modèle L1. Tous omettent au moins une variable qui a trait au tabagisme et le modèle L20 laisse également tomber l'IMC sous sa forme continue. Les résultats obtenus avec ce sous-ensemble sont présentés dans les Tableaux B.1 et B.2 de l'appendice B. Avec aussi peu que 500 individus, le modèle L20 est largement favorisé par les deux critères AIC et BIC en obtenant près de 100% des probabilités. Ce modèle est celui qui présente le plus petit biais, mais la plus grande variance de l'ensemble. Conséquemment à l'attribution des probabilités, les six estimateurs ECMP ainsi que le modèle L20 ont des résultats très semblables. Nous pouvons toutefois remarquer que puisque le critère BIC associe un peu plus de probabilité aux modèles L4, L5 et L6 que le AIC, les estimateurs ECMP présentent une variance plus petite en déclinaison BIC qu'en déclinaison AIC. Pour ce sous-ensemble, $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$ sont associés à un biais et une variance plus petite que $\hat{\theta}_{ext,AIC}$, $\hat{\theta}_{ext,BIC}$, $\hat{\theta}_{AIC}$ et $\hat{\theta}_{BIC}$. Lorsque nous augmentons la taille échantillonnale à 5000, les résultats sont les mêmes quel que soit l'estimateur ECMP puisque pratiquement 100% des probabilités sont associées au modèle L20.

Le sous-ensemble (b) comprend tous les modèles qui n'ont qu'une seule variable de moins que le vrai modèle L1. Comme tous les modèles ont le même nombre de variables, le poids qui leur est associé par le critère AIC est le même qu'avec le BIC. C'est le cas également pour les sous-ensembles (c) et (d). En regardant les échantillons de $n = 500$ individus (Tableau B.3), nous remarquons tout d'abord que le modèle L7 présente un biais supérieur aux autres modèles et que L8 a la plus petite variance. Cependant, à ces deux modèles très peu de probabilités sont associées. Les modèles L2, L11 et L14 ont chacun une probabilité associée d'environ 24%, alors que L3 et L15 ont chacun 13%. Même s'ils présentent des biais environ dix fois plus grands que $\hat{\theta}_{ext,AIC}$, $\hat{\theta}_{ext,BIC}$,

$\hat{\theta}_{AIC}$ et $\hat{\theta}_{BIC}$, les estimateurs par pondération interne, $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$, sont les estimateurs qui ont les plus petites EQM. C'est la réduction de la variance engendrée par la pondération interne qui fait en sorte que nous arrivons à de tels résultats. Comme les modèles considérés dans l'ensemble présentent des variances assez différentes, l'impact de l'utilisation de la pondération interne est assez important ici. Avec les échantillons de $n = 5000$ individus (Tableau B.4), les modèles L2, L11 et L14 se partagent à part égale la presque totalité des probabilités. Le modèle L2, qui omet une variable reliée au tabagisme a une variance et un biais plus grand que L11 et L14. Les estimateurs par pondération interne sont ceux des six estimateurs ECMP qui présentent le plus petit biais et également la plus petite variance. Leur EQM est donc également la plus petite, mais pas très éloignée de ce qui est obtenu avec la pondération externe. Toutefois, l'écart est plus grand lorsque nous la comparons à l'EQM des estimateurs par le modèle ayant le plus petit AIC/ BIC.

Le troisième sous-ensemble (c) contient le vrai modèle, L1, de même que les modèles qui ont le même nombre de variables que celui-ci. Sous une taille échantillonnale de 500, les quatre modèles obtiennent environ un quart des probabilités (Tableau B.5). Hormis L13, qui présente un biais plus petit et négatif, ces modèles ont un biais, une variance et une EQM assez semblable. Bien que les estimateurs par pondération interne aient un biais plus grand que chacun des modèles individuels et que les autres estimateurs ECMP, leur petite variance leur permet d'avoir une EQM plus petite que tous les estimateurs, à l'exception de l'estimateur appliqué au modèle L13. Lorsque nous augmentons la taille échantillonnale à 5000 (Tableau B.6), le modèle L13 est encore celui qui présente la plus petite EQM, mais il perd en importance par rapport aux autres modèles pour la construction des estimateurs ECMP puisque seulement 6% lui sont maintenant associés. Le modèle L1 est maintenant celui auquel les plus grandes probabilités sont associées, mais il semble qu'une plus grande taille échantillonnale est nécessaire pour voir ces probabilités s'approcher davantage de 1. Avec des échantillons de 10000 individus, ces probabilités augmentent à 38%. Comme les modèles ont des résultats semblables, les estimateurs ECMP sont aussi semblables entre eux. En jugeant la performance par

l'EQM, nous ne pouvons pas dire qu'une des techniques considérant un ensemble de modèles est vraiment préférable dans ce cas avec autant d'individus.

Le sous-ensemble (d) est composé des modèles qui incluent une variable de plus que le vrai modèle L1. Parmi ces modèles, seul L18 emboîte le vrai modèle. Avec $n = 500$, le modèle L21, qui est le seul omettant des variables ayant trait au tabagisme, obtient peu de probabilité, alors que L10, L18 et L19 se partagent presque également 99% des probabilités (Tableau B.7). Bien que les estimateurs par pondération externe aient un biais presque nul, c'est encore la variance qui avantage la pondération interne dans ce cas. Nous remarquons aussi que le fait d'utiliser un estimateur moyenné (externe ou interne) entraîne une EQM plus petite que celle obtenue de $\hat{\theta}_{AIC}$ et $\hat{\theta}_{BIC}$, de même que des estimateurs appliqués aux modèles individuels L10, L18 et L19. En augmentant la taille échantillonnale à 5000, le modèle L19 n'obtient plus que 15% des probabilités et L21 une probabilité minime (Tableau B.8). Le modèle L18, qui emboîte L1, est celui auquel est associé le plus de poids. Les six estimateurs ECMP produisent donc des résultats presque identiques puisque L10 et L18 présentent la même variance et des biais similaires. Les EQM des estimateurs par pondération externe ou interne sont égales ou plus petites que ce qui est obtenu avec les trois modèles les plus probables (L10, L18 et L19).

Le dernier des sous-ensembles, (e), est composé des modèles L9, L12 et L22. Notons que ces trois modèles n'ont pas le même nombre de variables et donc que l'utilisation du critère AIC ou BIC a un impact sur la pondération. Notons également que L9 emboîte L1. Avec les échantillons de 500 individus, L9 est celui qui est privilégié par les deux critères, mais presque à égalité avec le modèle L12 (Tableau B.9). Le critère AIC accorde également un peu de poids à L22 (8%), mais le BIC pénalise ce modèle qui contient beaucoup plus de variables que les deux autres. Les biais de $\hat{\theta}_{ext,AIC}$, $\hat{\theta}_{ext,BIC}$, $\hat{\theta}_{AIC}$ et $\hat{\theta}_{BIC}$ sont similaires, alors que $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$ ont des biais plus grands. Le biais des quatre premiers estimateurs ECMP n'est pas très différent que nous utilisions le critère AIC ou BIC, mais pour la pondération interne le BIC entraîne un biais plus petit que le AIC. Pour ce qui est de la variance, le BIC, accordant moins de poids à

L22, procure aux estimateurs ECMP une plus petite variance que le AIC. L'estimateur ECMP qui présente la plus petite EQM est donc $\hat{\theta}_{int,BIC}$ suivi de $\hat{\theta}_{int,AIC}$. Les deux versions de l'estimateur par pondération interne affichent une EQM plus petite que celle obtenue sous chacun des modèles individuellement. En considérant les estimations de 5000 individus, les probabilités associées au modèle L9 augmentent légèrement, mais le AIC en laisse toujours une part importante à L12 (41%) et à L22 (8%) et le BIC accorde toujours 44% à L12. Comme c'est le cas avec plusieurs des sous-ensembles présentés avec autant d'individus, les EQM obtenues à partir des estimateurs ECMP se ressemblent beaucoup. Les estimateurs par pondération interne entraînent une EQM aussi petite que celle obtenue sous le modèle L9, qui lui présente la plus petite EQM des trois modèles lorsqu'ils sont employés individuellement dans l'estimateur.

Bien que les résultats obtenus avec ce second exemple ne soient pas aussi frappants que ceux observés dans le premier exemple, les conclusions globales demeurent sensiblement les mêmes. Considérer un ensemble de modèles plutôt qu'un seul pour faire nos analyses semble être, de manière générale, une bonne approche à adopter. Les estimateurs par pondération externe ou interne offrent souvent de meilleurs résultats que les estimateurs employant le modèle associé au plus petit AIC ou BIC. Cela est observable surtout lorsque nous considérons des échantillons de plus petite taille. De plus, à cause de la réduction de la variance engendrée par le moyennage du score de propension, l'estimateur par pondération interne présente les résultats les plus avantageux lorsque nous nous intéressons à l'EQM. Il est difficile avec les résultats obtenus à partir de cet exemple de déterminer si l'un des critères d'ajustement, AIC ou BIC, devrait être privilégié. Les observations faites pour la pondération interne sont les suivantes. Dans la première situation, où nous considérons l'ensemble de tous les modèles, l'estimateur par pondération interne basé sur le AIC fournit un plus grand biais, mais une plus petite variance et EQM que celui employant le BIC. Rappelons que dans la deuxième situation seulement les sous-ensembles (a) et (e) fournissent des résultats différents selon l'utilisation du AIC ou du BIC puisque ce sont les seuls sous-ensembles qui contiennent des modèles dont le nombre de variables n'est pas nécessairement égal d'un modèle à

l'autre. Pour ces sous-ensembles, c'est l'estimateur par pondération interne basé sur le BIC qui produit à la fois le plus petit biais et la plus petite variance.

Concluons en rappelant les principaux résultats obtenus à partir de nos simulations. Un premier exemple nous a permis d'analyser les résultats obtenus d'un cas assez simple. Nous en avons conclu que lorsque le vrai modèle fait partie de l'ensemble des modèles considérés pour l'estimation, les estimateurs par pondération interne offrent une très bonne performance au niveau de l'EQM. En utilisant une pondération basée sur le critère AIC, l'estimateur par pondération interne présentait même une EQM plus petite que celle observée sous le vrai modèle. Quand le sous-ensemble de modèles utilisé omet le vrai modèle, mais contient des modèles qui emboîtent le vrai modèle, l'estimateur $\hat{\theta}_{int,AIC}$ est encore celui qui fournit la plus petite EQM. Finalement, quand ni le vrai modèle, ni des modèles l'emboîtant ne font partie du sous-ensemble de modèles, nous obtenons encore que l'estimateur par pondération interne performe mieux que les autres estimateurs appliqués à un sous-ensemble de modèles ou aux modèles individuellement. Il est cependant moins clair de définir lequel des critères d'ajustement donne les meilleurs résultats.

Un deuxième exemple nous a permis d'examiner les mêmes estimateurs dans un cas plus complexe où la difficulté provenait des différentes définitions des variables confondantes et où il y avait de la dépendance entre certaines des variables. Dans ce cas, les estimateurs par pondération interne fournissent les plus petites EQM, mais la différence entre les EQM obtenues des estimateurs par pondération externe et interne est souvent moins grande que dans le premier exemple. Dans la plupart des sous-ensembles présentés dans cet exemple, la pondération interne engendre un plus grand biais, mais toujours une plus petite variance que la pondération externe. De plus, les estimateurs $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$ n'offrent pas la plus petite des EQM. Certains modèles employés de façon individuelle dans l'estimateur procurent une plus petite EQM. Toutefois, nous pouvons quand même conclure que l'utilisation des estimateurs pondérés offre de très bons résultats.

Lorsque nous décidons d'utiliser un estimateur qui considère un ensemble de modèles, ce que nous recommandons, les estimateurs pondérés produisent des résultats intéressants et plus particulièrement avec la pondération interne. L'avantage qu'apporte l'estimateur par pondération interne par rapport à l'estimateur par pondération externe est une réduction de la variance. Cela est remarquable surtout lorsque nous avons des échantillons de plus petite taille. Si le nombre d'individus est plus grand, les trois estimateurs qui considèrent un ensemble de modèles fournissent des résultats très semblables.

CONCLUSION

Nous voulions dans ce projet aborder le sujet de l'estimation causale en passant à travers divers volets. Nous avons tout d'abord exposé les principaux problèmes rencontrés lors de l'estimation causale en épidémiologie, soient la présence de variables confondantes et la difficulté de traiter ces dernières à l'intérieur d'études observationnelles. Une revue des techniques les plus couramment utilisées dans la pratique pour estimer les effets de type causal a ensuite été présentée. Nous nous sommes intéressés plus particulièrement à l'une de ces techniques : l'estimation par pondération par les probabilités inversées. Comme nous l'avons souvent répété dans ce mémoire, l'estimateur par pondération par les probabilités inversées nécessite la spécification d'un modèle de traitement. Ce modèle est d'une importance cruciale puisque s'il n'est pas bien déterminé, l'estimation peut s'en trouver grandement biaisée. Nous avons l'impression qu'en pratique l'incertitude liée à la sélection de ce modèle est souvent trop vite oubliée ou du moins sous-estimée.

Le premier de nos objectifs était de pallier l'incertitude liée à la sélection du modèle de traitement en considérant l'utilisation d'un ensemble de modèles, plutôt que d'en choisir un seul, pour obtenir les estimations. Nous pouvons alors espérer que soit le bon modèle fasse partie de l'ensemble ou, à plus forte raison, qu'il n'en fasse pas partie, mais qu'à partir des différents modèles, l'estimateur puisse construire une meilleure estimation que celle obtenue à partir d'un seul « mauvais » modèle. Pour ce faire, nous avons décrit trois différentes méthodes se déclinant chacune de deux façons suivant l'utilisation du critère d'ajustement AIC ou BIC : l'estimation à partir du modèle étant associé à la plus petite valeur de AIC/BIC, l'estimation par pondération externe et l'estimation par pondération interne. Le deuxième objectif était de déterminer si l'une de ces techniques devrait être privilégiée par rapport aux autres. Enfin, nous voulions également voir si un critère d'ajustement pouvait aussi être préféré.

Nous avons étudié les différentes questions de manière empirique à travers des échantillons simulés. Différentes tailles d'échantillon et plusieurs situations ont été présentées afin d'analyser le comportement des estimateurs. Dans chacun de nos exemples, nous avons supposé que nous connaissions toutes les variables confondantes, mais que nous ignorions sous quelle forme elles influençaient l'attribution au traitement et la variable résultat. Nos simulations nous ont permis de confirmer que dans la plupart des cas, l'utilisation d'un ensemble de modèles s'avère avantageuse. Nous avons observé que le moyennage des scores de propension, combiné dans nos exemples à l'utilisation d'un plafond appliqué à la valeur du poids d'un individu, réduit la variance de l'estimateur par pondération interne par rapport à l'estimateur par pondération externe. Les estimateurs $\hat{\theta}_{int,AIC}$ et $\hat{\theta}_{int,BIC}$ sont ceux des six estimateurs ECMP qui présentent les plus petites EQM. Nous avons vu que les estimateurs par pondération interne, plus particulièrement $\hat{\theta}_{int,AIC}$, offrent souvent de meilleures performances que les modèles employés individuellement dans l'estimateur par pondération par les probabilités inversées, et ce, même quand le modèle est le vrai. Les avantages que présentent les estimateurs par pondération interne sont observables lorsque nous avons des échantillons de plus petite taille. Cela nous amène à croire que leur utilisation dans la pratique pourrait être bénéfique.

Les observations faites à partir de ce projet nous amènent à vouloir approfondir davantage l'idée d'utiliser un ensemble de modèles plutôt qu'un seul pour l'estimation d'un effet causal. Plusieurs pistes méritent d'être explorées. Il serait intéressant, par exemple, d'intégrer la pondération externe et interne aux estimateurs doublement robustes. Nous pourrions aussi penser à une façon d'intégrer l'idée de considérer plusieurs modèles avant même l'estimation, mais lors de la sélection de ces modèles à partir d'un grand ensemble de variables confondantes ou non. Enfin, il serait important d'évaluer la performance des estimateurs moyennés sur de vraies données, où les résultats provenant de ceux-ci pourraient être comparés aux résultats découlant d'études expérimentales.

APPENDICE A

RÉSULTATS DU PREMIER EXEMPLE

Tableau A.1 Résultats complets obtenus en considérant tous les modèles (M1 à M14) avec $n = 300$ et $ech = 5000$

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
Biais	0,211	1,477	1,112	0,423	1,384	0,887	0,214
Variance	0,789	0,495	1,118	0,683	0,672	0,967	0,828
EQM	0,833	2,678	2,354	0,862	2,588	1,753	0,874
$\overline{p_{AIC}} \times 10^2$	31,491	2,824	6,528	14,053	2,868	5,972	12,922
$\overline{p_{BIC}} \times 10^2$	16,556	53,455	17,023	5,485	5,458	1,361	0,292

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
Biais	0,206	1,473	1,111	0,421	1,378	0,882	0,210
Variance	0,759	0,409	1,082	0,624	0,589	0,893	0,780
EQM	0,801	2,579	2,316	0,801	2,486	1,670	0,824
$\overline{p_{AIC}} \times 10^2$	9,514	0,868	1,900	4,359	0,894	1,794	4,013
$\overline{p_{BIC}} \times 10^2$	0,062	0,207	0,056	0,020	0,020	0,004	0,001

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,290	1,007	0,347	0,974	0,357	0,911
Variance	0,799	0,732	0,673	0,537	0,521	0,363
EQM	0,883	1,747	0,794	1,486	0,648	1,193

Tableau A.2 Résultats complets obtenus en considérant tous les modèles (M1 à M14) avec $n = 500$ et $ech = 5000$

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
Biais	0,143	1,467	1,187	0,367	1,369	0,964	0,144
Variance	0,579	0,277	1,021	0,507	0,335	1,019	0,592
EQM	0,599	2,428	2,430	0,642	2,210	1,949	0,612
$\overline{p_{AIC}} \times 10^2$	41,931	0,357	2,419	11,784	0,590	3,167	16,924
$\overline{p_{BIC}} \times 10^2$	40,135	28,630	16,806	8,779	3,662	1,365	0,411

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
Biais	0,150	1,474	1,193	0,375	1,374	0,971	0,151
Variance	0,546	0,227	0,944	0,464	0,284	0,946	0,555
EQM	0,569	2,398	2,368	0,604	2,173	1,889	0,578
$\overline{p_{AIC}} \times 10^2$	12,407	0,100	0,736	3,299	0,179	1,002	5,104
$\overline{p_{BIC}} \times 10^2$	0,091	0,058	0,036	0,013	0,009	0,003	0,001

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,170	0,622	0,206	0,629	0,220	0,605
Variance	0,572	0,626	0,508	0,455	0,419	0,326
EQM	0,601	1,014	0,550	0,850	0,467	0,692

Tableau A.3 Résultats complets obtenus en considérant tous les modèles (M1 à M14) avec $n = 1500$ et $ech = 5000$

	$\hat{\theta}_{M1}$	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
Biais	0,082	1,468	1,267	0,319	1,362	1,044	0,084
Variance	0,265	0,087	0,442	0,230	0,102	0,429	0,267
EQM	0,272	2,243	2,046	0,333	1,958	1,520	0,275
$\overline{p_{AIC}} \times 10^2$	53,318	$< 0{,}001$	0,001	1,373	$< 0{,}001$	0,040	21,994
$\overline{p_{BIC}} \times 10^2$	96,172	0,024	0,373	2,850	0,008	0,078	0,435

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
Biais	0,082	1,469	1,267	0,319	1,363	1,044	0,083
Variance	0,251	0,072	0,424	0,214	0,087	0,413	0,252
EQM	0,258	2,229	2,029	0,316	1,944	1,503	0,259
$\overline{p_{AIC}} \times 10^2$	16,187	$< 0{,}001$	$< 0{,}001$	0,415	$< 0{,}001$	0,012	6,658
$\overline{p_{BIC}} \times 10^2$	0,057	$< 0{,}001$	$< 0{,}001$	0,002	$< 0{,}001$	$< 0{,}001$	$< 0{,}001$

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,082	0,085	0,084	0,089	0,086	0,091
Variance	0,261	0,266	0,256	0,265	0,254	0,261
EQM	0,268	0,273	0,263	0,273	0,261	0,270

Tableau A.4 Résultats complets obtenus en considérant les modèles M2 à M14 avec $n = 300$ et $ech = 5000$

	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
Biais	1,477	1,112	0,423	1,384	0,887	0,214
Variance	0,495	1,118	0,683	0,672	0,967	0,828
EQM	2,678	2,354	0,862	2,588	1,753	0,874
$\overline{p_{AIC}} \times 10^2$	3,599	8,296	20,204	4,134	8,647	21,856
$\overline{p_{BIC}} \times 10^2$	58,729	19,882	9,798	7,263	2,414	1,288

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
Biais	0,206	1,473	1,111	0,421	1,378	0,882	0,210
Variance	0,759	0,409	1,082	0,624	0,589	0,893	0,780
EQM	0,801	2,579	2,316	0,801	2,486	1,670	0,824
$\overline{p_{AIC}} \times 10^2$	15,421	1,022	2,241	5,503	1,139	2,301	5,637
$\overline{p_{BIC}} \times 10^2$	0,270	0,227	0,063	0,031	0,025	0,007	0,003

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,355	1,164	0,406	1,130	0,409	1,041
Variance	0,789	0,656	0,635	0,508	0,474	0,339
EQM	0,915	2,010	0,800	1,784	0,642	1,421

Tableau A.5 Résultats complets obtenus en considérant les modèles M2 à M14 avec $n = 500$ et $ech = 5000$

	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
Biais	1,467	1,187	0,367	1,369	0,964	0,144
Variance	0,277	1,021	0,507	0,335	1,019	0,592
EQM	2,428	2,430	0,642	2,210	1,949	0,612
$\overline{p_{AIC}} \times 10^2$	0,525	3,452	18,741	1,041	5,457	32,875
$\overline{p_{BIC}} \times 10^2$	36,892	23,762	21,857	7,540	4,360	4,514

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
Biais	0,150	1,474	1,193	0,375	1,374	0,971	0,151
Variance	0,546	0,227	0,944	0,464	0,284	0,946	0,555
EQM	0,569	2,398	2,368	0,604	2,173	1,889	0,578
$\overline{p_{AIC}} \times 10^2$	22,761	0,132	0,926	4,536	0,265	1,418	7,871
$\overline{p_{BIC}} \times 10^2$	0,899	0,072	0,047	0,030	0,015	0,007	0,005

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,204	0,907	0,244	0,902	0,260	0,838
Variance	0,532	0,527	0,479	0,370	0,374	0,246
EQM	0,574	1,350	0,539	1,184	0,442	0,949

Tableau A.6 Résultats complets obtenus en considérant les modèles M2 à M14 avec $n = 1500$ et $ech = 5000$

	$\hat{\theta}_{M2}$	$\hat{\theta}_{M3}$	$\hat{\theta}_{M4}$	$\hat{\theta}_{M5}$	$\hat{\theta}_{M6}$	$\hat{\theta}_{M7}$
Biais	1,468	1,267	0,319	1,362	1,044	0,084
Variance	0,087	0,442	0,230	0,102	0,429	0,267
EQM	2,243	2,046	0,333	1,958	1,520	0,275
$\overline{p_{AIC}} \times 10^2$	$< 0{,}001$	0,003	2,683	$< 0{,}001$	0,100	50,828
$\overline{p_{BIC}} \times 10^2$	0,223	3,030	37,481	0,259	3,090	50,112

	$\hat{\theta}_{M8}$	$\hat{\theta}_{M9}$	$\hat{\theta}_{M10}$	$\hat{\theta}_{M11}$	$\hat{\theta}_{M12}$	$\hat{\theta}_{M13}$	$\hat{\theta}_{M14}$
Biais	0,082	1,469	1,267	0,319	1,363	1,044	0,083
Variance	0,251	0,072	0,424	0,214	0,087	0,413	0,252
EQM	0,258	2,229	2,029	0,316	1,944	1,503	0,259
$\overline{p_{AIC}} \times 10^2$	34,180	$< 0{,}001$	0,001	0,641	$< 0{,}001$	0,023	11,539
$\overline{p_{BIC}} \times 10^2$	5,783	$< 0{,}001$	0,001	0,011	$< 0{,}001$	0,001	0,009

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais	0,083	0,187	0,087	0,196	0,089	0,206
Variance	0,259	0,269	0,253	0,248	0,250	0,221
EQM	0,266	0,304	0,261	0,287	0,258	0,264

APPENDICE B

RÉSULTATS DU DEUXIÈME EXEMPLE

Tableau B.1 Résultats obtenus en considérant les modèles L4, L5, L6 et L20 avec $n = 500$ et ech=5000

	$\hat{\theta}_{L4}$	$\hat{\theta}_{L5}$	$\hat{\theta}_{L6}$	$\hat{\theta}_{L20}$
Biais$\times 10^2$	-9,777	-9,774	-17,022	-3,496
Variance$\times 10^2$	0,691	0,630	0,394	3,565
EQM$\times 10^2$	1,647	1,586	3,292	3,687
$\overline{p_{AIC}} \times 10^2$	0,177	0,111	0,017	99,695
$\overline{p_{BIC}} \times 10^2$	0,126	0,503	0,058	99,313

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	-3,503	-3,509	-3,501	-3,509	-3,419	-3,357
Variance$\times 10^2$	3,565	3,564	3,561	3,558	3,554	3,548
EQM$\times 10^2$	3,688	3,688	3,684	3,681	3,671	3,661

Tableau B.2 Résultats obtenus en considérant les modèles L4, L5, L6 et L20 avec $n = 5000$ et ech=5000

	$\hat{\theta}_{L4}$	$\hat{\theta}_{L5}$	$\hat{\theta}_{L6}$	$\hat{\theta}_{L20}$
Biais$\times 10^2$	-9,688	-9,720	-17,009	-3,516
Variance$\times 10^2$	0,044	0,042	0,025	0,258
EQM$\times 10^2$	0,983	0,987	2,918	0,382
$\overline{p_{AIC}} \times 10^2$	$< 0{,}001$	$< 0{,}001$	$< 0{,}001$	100,000
$\overline{p_{BIC}} \times 10^2$	$< 0{,}001$	$< 0{,}001$	$< 0{,}001$	100,000

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	-3,516	-3,516	-3,516	-3,516	-3,516	-3,516
Variance$\times 10^2$	0,258	0,258	0,258	0,258	0,258	0,258
EQM$\times 10^2$	0,382	0,382	0,382	0,382	0,382	0,382

Tableau B.3 Résultats obtenus en considérant les modèles L2, L3, L7, L8, L11, L14 et L15 avec $n = 500$ et ech=5000

	$\hat{\theta}_{L2}$	$\hat{\theta}_{L3}$	$\hat{\theta}_{L7}$	$\hat{\theta}_{L8}$	$\hat{\theta}_{L11}$	$\hat{\theta}_{L14}$	$\hat{\theta}_{L15}$
Biais$\times 10^2$	-2,418	2,414	-5,817	-3,840	0,284	0,796	1,734
Variance$\times 10^2$	3,839	3,239	3,697	1,556	3,124	3,033	2,846
EQM$\times 10^2$	3,897	3,298	4,036	1,703	3,125	3,039	2,876
$\overline{p_{AIC}} \times 10^2$	24,088	12,749	2,600	0,859	23,081	23,854	12,770
$\overline{p_{BIC}} \times 10^2$	24,088	12,749	2,600	0,859	23,081	23,854	12,770

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	0,113	0,113	0,091	0,091	1,036	1,036
Variance$\times 10^2$	3,295	3,295	3,192	3,192	3,097	3,097
EQM$\times 10^2$	3,295	3,295	3,193	3,193	3,107	3,107

Tableau B.4 Résultats obtenus en considérant les modèles L2, L3, L7, L8, L11, L14 et L15 avec $n = 5000$ et ech=5000

	$\hat{\theta}_{L2}$	$\hat{\theta}_{L3}$	$\hat{\theta}_{L7}$	$\hat{\theta}_{L8}$	$\hat{\theta}_{L11}$	$\hat{\theta}_{L14}$	$\hat{\theta}_{L15}$
Biais$\times 10^2$	-2,277	2,396	-6,181	-3,581	0,317	0,787	1,587
Variance$\times 10^2$	0,272	0,210	0,275	0,100	0,207	0,208	0,191
EQM$\times 10^2$	0,323	0,268	0,657	0,228	0,208	0,214	0,216
$\overline{p_{AIC}} \times 10^2$	32,624	0,189	$< 0,001$	$< 0,001$	33,100	33,765	0,323
$\overline{p_{BIC}} \times 10^2$	32,624	0,189	$< 0,001$	$< 0,001$	33,100	33,765	0,323

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	-0,163	-0,163	-0,201	-0,201	-0,034	-0,034
Variance$\times 10^2$	0,226	0,226	0,221	0,221	0,219	0,219
EQM$\times 10^2$	0,227	0,227	0,221	0,221	0,219	0,219

Tableau B.5 Résultats obtenus en considérant les modèles L1, L13, L16 et L17 avec $n = 500$ et ech=5000

	$\hat{\theta}_{L1}$	$\hat{\theta}_{L13}$	$\hat{\theta}_{L16}$	$\hat{\theta}_{L17}$
Biais$\times 10^2$	0,165	-0,056	0,192	0,207
Variance$\times 10^2$	3,252	3,170	3,241	3,250
EQM$\times 10^2$	3,252	3,170	3,242	3,250
$\overline{p_{AIC}} \times 10^2$	25,714	22,396	26,077	25,814
$\overline{p_{BIC}} \times 10^2$	25,714	22,396	26,077	25,814

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	0,129	0,129	0,141	0,141	0,393	0,393
Variance$\times 10^2$	3,343	3,343	3,254	3,254	3,236	3,236
EQM$\times 10^2$	3,343	3,343	3,254	3,254	3,238	3,238

Tableau B.6 Résultats obtenus en considérant les modèles L1, L13, L16 et L17 avec $n = 5000$ et ech=5000

	$\hat{\theta}_{L1}$	$\hat{\theta}_{L13}$	$\hat{\theta}_{L16}$	$\hat{\theta}_{L17}$
Biais$\times 10^2$	0,168	-0,127	0,193	0,219
Variance$\times 10^2$	0,213	0,207	0,213	0,211
EQM$\times 10^2$	0,213	0,207	0,213	0,212
$\overline{p_{AIC}} \times 10^2$	33,126	6,320	32,781	27,773
$\overline{p_{BIC}} \times 10^2$	33,126	6,320	32,781	27,773

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	0,156	0,156	0,168	0,168	0,222	0,222
Variance$\times 10^2$	0,213	0,213	0,212	0,212	0,212	0,212
EQM$\times 10^2$	0,214	0,214	0,213	0,213	0,213	0,213

Tableau B.7 Résultats obtenus en considérant les modèles L10, L18, L19 et L21 avec $n = 500$ et ech=5000

	$\hat{\theta}_{L10}$	$\hat{\theta}_{L18}$	$\hat{\theta}_{L19}$	$\hat{\theta}_{L21}$
Biais$\times 10^2$	0,147	0,178	-0,573	-0,517
Variance$\times 10^2$	3,326	3,347	3,325	1,451
EQM$\times 10^2$	3,327	3,347	3,328	1,454
$\overline{p_{AIC}} \times 10^2$	32,964	33,805	31,882	1,349
$\overline{p_{BIC}} \times 10^2$	32,964	33,805	31,882	1,349

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	0,119	0,119	0,037	0,037	0,583	0,583
Variance$\times 10^2$	3,419	3,419	3,319	3,319	3,275	3,275
EQM$\times 10^2$	3,419	3,419	3,319	3,319	3,278	3,278

Tableau B.8 Résultats obtenus en considérant les modèles L10, L18, L19 et L21 avec $n = 5000$ et ech=5000

	$\hat{\theta}_{L10}$	$\hat{\theta}_{L18}$	$\hat{\theta}_{L19}$	$\hat{\theta}_{L21}$
Biais$\times 10^2$	0,124	0,161	-0,566	-0,402
Variance$\times 10^2$	0,212	0,212	0,211	0,089
EQM$\times 10^2$	0,212	0,213	0,215	0,090
$\overline{p_{AIC}} \times 10^2$	39,232	46,199	14,569	$< 0{,}001$
$\overline{p_{BIC}} \times 10^2$	39,232	46,199	14,569	$< 0{,}001$

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	0,162	0,162	0,103	0,103	0,146	0,146
Variance$\times 10^2$	0,213	0,213	0,212	0,212	0,212	0,212
EQM$\times 10^2$	0,213	0,213	0,212	0,212	0,212	0,212

Tableau B.9 Résultats obtenus en considérant les modèles L9, L12 et L22 avec $n = 500$ et ech=5000

	$\hat{\theta}_{L9}$	$\hat{\theta}_{L12}$	$\hat{\theta}_{L22}$
Biais$\times 10^2$	0,617	0,188	0,508
Variance$\times 10^2$	3,318	3,420	3,879
EQM$\times 10^2$	3,322	3,421	3,881
$\overline{p_{AIC}} \times 10^2$	46,725	45,230	8,045
$\overline{p_{BIC}} \times 10^2$	50,884	49,102	0,015

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	0,476	0,471	0,431	0,436	1,614	1,337
Variance$\times 10^2$	3,587	3,462	3,413	3,323	3,272	3,208
EQM$\times 10^2$	3,589	3,464	3,415	3,325	3,298	3,226

Tableau B.10 Résultats obtenus en considérant les modèles L9, L12 et L22 avec $n = 5000$ et ech=5000

	$\hat{\theta}_{L9}$	$\hat{\theta}_{L12}$	$\hat{\theta}_{L22}$
Biais$\times 10^2$	0,212	0,177	0,192
Variance$\times 10^2$	0,205	0,212	0,207
EQM$\times 10^2$	0,205	0,212	0,207
$\overline{p_{AIC}} \times 10^2$	51,240	41,126	7,634
$\overline{p_{BIC}} \times 10^2$	55,752	44,248	$< 0,001$

	$\hat{\theta}_{AIC}$	$\hat{\theta}_{BIC}$	$\hat{\theta}_{ext,AIC}$	$\hat{\theta}_{ext,BIC}$	$\hat{\theta}_{int,AIC}$	$\hat{\theta}_{int,BIC}$
Biais$\times 10^2$	0,186	0,192	0,191	0,196	0,305	0,286
Variance$\times 10^2$	0,207	0,207	0,206	0,206	0,204	0,205
EQM$\times 10^2$	0,207	0,207	0,206	0,206	0,205	0,205

RÉFÉRENCES

Bang, H. et J. M. Robins. 2005. « Doubly robust estimation in missing data and causal inference models », *Biometrics*, vol. 61, p. 962–972.

Brookhart, A. M., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, et T. Stürmer. 2006. « Variable selection for propensity score models », *American Journal of Epidemiology*, vol. 163, no. 12, p. 1149–1156.

Burnham, K. P. et D. R. Anderson. 2004. « Multimodel inference ; understanding AIC and BIC in model selection », *Sociological Methods & Research*, vol. 33, no. 2, p. 261–304.

Capital Souffle. 2010. « La mesure du souffle ». En ligne. <http://www.capitalsouffle.fr/mesure_du_souffle.htm>. Consulté le 2 Septembre 2010.

Cochran, W. G. 1968. « The effectiveness of adjustment by subclassification in removing bias in observational studies », *Biometrics*, vol. 24.

Cole, S. R. et M. A. Hernan. 2008. « Constructing inverse probability weights for marginal structural models », *American Journal of Epidemiology*, vol. 168, no. 6, p. 656–664.

D'Agostino, R. B. J. 1998. « Propensity score methods for bias reduction in the comparaison of a treatment to a non-randomized control group », *Statistics in Medecine*, vol. 17, no. 19, p. 2265–2281.

Davidian, M. 2007. « Double robustness in estimation of causal treatment effects ». En ligne. <http://www4.stat.ncsu.edu/~davidian/double.pdf>. Consulté le 12 avril 2010.

Greenland, S., J. Pearl, et J. M. Robins. 1999. « Causal diagrams for epidemiologic research », *Epidemiology*, vol. 1, no. 10, p. 37–48.

Greenland, S., J. M. Robins, et J. Pearl. 1999. « Confounding and collapsibility in causal inference », *Statistical Science*, vol. 14, no. 1, p. 29–46.

Hernan, M. A. 2004. « A definition of causal effect for epidemiological research », *Journal of Epidemiology and Community Health*, vol. 58, no. 4, p. 265–271.

Kang, J. D. Y. et J. L. Schafer. 2007. « Demystifying double robustness : A comparaison of alternatives strategies for estimating a population mean from incomplete data », *Statistical Science*, vol. 22, no. 4, p. 523–539.

Kenneth, J. R. et S. Greenland. 1998. *Modern Epidemiology, second edition.* Lipincott, 3 édition.

Lunceford, J. K. et M. Davidian. 2004. « Stratification and weighitng via the propensity score in estimation of causal treatment effects : a comparative study », *Statistics in medicine*, vol. 23, no. 19, p. 2937–2960.

Rosenbaum, P. R. et D. B. Rubin. 1983. « The central role of the propensity score in observational studies for causal effects », *Biometrika*, vol. 70, no. 1, p. 41–55.

Rothman, K. J. 2002. *Epidemiology An introduction.* Oxford University Press.

Weitzen, S., K. L. Lapane, A. Y. Toledano, A. L. Hume, et V. Mor. 2004. « Principles for modeling propensity scores in medical research : a systematic literature review », *Pharmacoepidemiology and drug safety*, vol. 13, no. 12, p. 841–853.