Université de Montréal

Comparaison d'estimateurs de la variance du TMLE

par

Laurence Boulanger

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

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

mai 2019

Université de Montréal

Faculté des études supérieures et postdoctorales

Ce mémoire intitulé

Comparaison d'estimateurs de la variance du TMLE

présenté par

Laurence Boulanger

a été évalué par un jury composé des personnes suivantes :

Christian Léger
(président-rapporteur)

Mireille Schnitzer
(directeur de recherche)

David Haziza
(co-directeur)

Alejandro Murua
(membre du jury)

Mémoire accepté le :
7 septembre 2018

Sommaire

Dans un contexte d'inférence causale où l'on cherche à estimer l'effet causal moyen d'une exposition dichotomique sur une variable issue, le TMLE de M. Van der Laan et D. Rubin est une technique qui applique à une estimation initiale de l'espérance conditionnelle de l'issue une modification ayant pour but de réduire le biais dans l'estimation correspondante de l'effet causal moyen. L'étalon-or en ce qui concerne l'estimation de la variance du TMLE est un estimateur de type sandwich. Divers estimateurs alternatifs sont identifiés et comparés à l'étalon-or lorsque la taille échantillonnale est petite, lorsque les scores de propension sont grands ou petits et lorsque les estimations initiales pour le score de propension ou pour l'espérance conditionnelle de l'issue sont mauvais. À chacun des estimateurs de la variance, un intervalle de confiance pour le TMLE est construit. On trouve que l'estimateur Jackknife donne des taux de couverture équivalents ou supérieurs aux autres estimateurs dans toutes ces situations.

Mots clefs

Inférence causale, TMLE, estimation de la variance, estimateur sandwich, Jackknife

Summary

In the context of causal inference where we seek to estimate the average causal effect of a binary exposure on an outcome variable, the TMLE of M. Van der Laan and D. Rubin is a technique which applies to an initial estimate of the conditional expectation of the outcome a modification whose purpose is to reduce bias in the corresponding estimation of the average causal effect. The gold standard for estimation of the variance of the TMLE is a sandwich-type estimator. Various alternative estimators are identified and compared to the gold standard when the sample size is small, when the propensity scores are high or low and when the initial estimate of the propensity score or of the conditional expectation of the outcome is bad. for each of the variance estimator, a confidence interval for the TMLE is constructed. We find that the Jackknife estimator yields coverage rates equivalent or better than all the other estimators in all these situations.

Key words

Causal inference, TMLE, variance estimation, sandwich estimator, Jackknife

Table des matières

Liste des tableaux

Liste des figures

Remerciements

Je remercie mes superviseurs Mireille Schnitzer et David Haziza pour leur aide précieuse dans l'élaboration de ce mémoire.

Introduction

Dans un monde où les ensembles de données sont de plus en plus massifs, avec parfois plus de variables mesurées pour chaque individus que d'individus, les méthodes statistiques classiques s'appuyant sur un modèle paramétrique ne suffisent plus. De même, dès que la relation fonctionnelle entre une variable issue et les variables mesurées est complexe, l'utilisation d'un modèle paramétrique n'est pas sans risque. C'est souvent le cas dans un contexte d'inférence causale où l'on cherche à estimer l'effet causal moyen d'une variable exposition dichotomique sur une variable issue Ėn effet, la relation entre les variables est typiquement complexe et inconnue de sorte que l'utilisation d'un modèle paramétrique entraînerait une sévère erreur au niveau de l'estimation et de l'intervalle de confiance associé en cas de mauvaise spécification du modèle. Il devient alors nécessaire de développer de nouvelles méthodes qui, étant donné un modèle semi-paramétrique ou non-paramétrique pour la distribution des données, estiment un paramètre d'intérêt ciblé avec un maximum d'efficacité. Le TMLE, pour Targeted Maximum Likelihood Estimator de M. van der Laan et D. Rubin [**45, 43**] est une méthode générale qui, lorsqu'appliquée au problème de l'estimation de l'effet causal moyen, conduit à un estimateur non biaisé et efficace par rapport à la classe des estimateurs asymptotiquement linéaires [**43**]. Afin d'obtenir un intervalle de confiance, il faut également estimer la variance du TMLE et l'étalon-or pour l'estimation de la variance est un estimateur de type sandwich [**43, 35**]. Cependant, il est connu que l'étalon-or n'est pas sans défaut car il tend à converger trop lentement sous certaines conditions (cf. [**19**]). Le présent mémoire comparera l'étalon-or à des estimateurs de la variance alternatifs dans diverses situations.

Le chapitre 1 se veut une exposition cohérente des divers éléments permettant de comprendre la nature et les propriétés du TMLE appliqué à l'estimation de l'effet causal moyen. À la section 1.1, nous présentons les bases de l'inférence causale suivant D. Rubin [**28, 29, 30**]. L'effet causal moyen, qui permet d'identifier la présence d'une relation de

causalité entre une variable issue et une variable exposition dichotomique, est défini et les présuppositions causales permettant son calcul sont exposées. La section 1.2 introduit un premier exemple d'estimateur de l'effet causal moyen qui peut être vu comme un ancêtre du TMLE. La section 1.3 donne la définition d'un estimateur semi paramétrique asymptotiquement linéaire et de la fonction d'influence qui en détermine la variance asymptotique. En exemple, on trouve la classe des m-estimateurs pour lesquels l'estimateur sandwich de la variance est développé. La section 1.4 s'attaque à la théorie de l'efficacité concernant les estimateurs asymptotiquement linéaires. Le concept clé est ici celui de la fonction d'influence efficace et la construction géométrique permettant d'obtenir celle-ci par projection orthogonale d'une quelconque fonction d'influence. La section 1.5 commence par le calcul de la fonction d'influence efficace pour le modèle non-paramétrique. Un nouvel estimateur de l'effet causal moyen est ensuite introduit comme choix naturel d'un estimateur asymptotiquement linéaire ayant comme fonction d'influence la fonction d'influence efficace. Une propriété de double robustesse est démontrée pour cet estimateur lorsque certaines fonctions y figurant, le score de propension et l'espérance conditionnelle de la variable issue, sont estimés. À la section 1.6, l'estimateur TMLE pour l'effet causal moyen est présenté comme un cas particulier de l'estimateur de la section 1.5 où le score de propension et l'espérance conditionnelle de la variable issue sont estimés de telle sorte qu'une estimation initiale de l'espérance conditionnelle de la variable issue subit une modification, l'étape de mise-à-jour TMLE, prenant en compte l'estimation du score de propension et ayant pour but de réduire le biais dans l'estimation de l'effet causal moyen. La procédure générale TMLE est ensuite expliquée et l'étape de mise-à-jour TMLE dans l'estimation de l'effet causal moyen est contextualisée. Enfin, nous montrons comment l'étalon-or en matière d'estimation de la variance du TMLE peut être vu comme un exemple d'estimateur sandwich.

Le chapitre 2 introduit les différents estimateurs de la variance du TMLE que nous nous proposons de comparer à l'étalon-or à l'aide de simulations Monte-Carlo. La section **??** introduit l'estimateur Monte-Carlo qui, dans une étude de simulations, permet d'approximer la vraie valeur du paramètre d'intérêt. La section 2.1 contient 3 estimateurs de la variance dans le cas où le score de propension est estimé par un modèle paramétrique. Dans ce cas, l'estimateur sandwich ayant permis de produire d'étalon-or prend généralement une forme plus compliquée. À celle-ci, on peut appliquer l'un de deux termes de correction, la

correction de Fay-Graubard et la correction conservative, lorsqu'un ajustement à la hausse est nécessaire. La section 2.2 présente l'estimateur Jackknife, un estimateur non-paramétrique basée sur le principe du ré échantillonnage.

Le chapitre 3 présente les résultats des simulations ayant pour but de comparer l'étalon-or aux estimateurs de la variance vus au chapitre 2. La section 3.1 décrit brièvement les différents scénarios dans lesquels les estimateurs seront comparés. La section 3.3 décrit en détails les 3 scénarios considérés. La section 3.4 présente les résultats sous forme de graphiques en termes du biais relatif Monte-Carlo des différents estimateurs et sous forme de tableaux contenant le biais relatif Monte-Carlo et le taux de couverture des intervalles de confiances associés. La section 3.5 se veut un résumé des conclusions tirées.

Chapitre 1

Revue de littérature

1.1. Inférence causale

L'inférence causale est la branche de la statistique dont le but est d'identifier et quantifier la présence de relations de causalité entre différentes variables. Il est bien connu que la présence d'une association entre deux quantités n'implique pas une relation de causalité entre elles. Une difficulté importante à laquelle on doit faire est la présence de variables de confusion. Un exemple anecdotique célèbre illustrant ce concept fait intervenir Sir Ronald Fisher qui, lorsque confronté à des données révélant que la probabilité de développer une maladie pulmonaire est supérieure chez les fumeurs que chez les non-fumeurs, rejeta ces dernières comme constituant une preuve de l'existence d'une relation causale entre le fait de fumer et le développement d'une maladie pulmonaire. Pour illustrer son point, il évoqua l'existence hypothétique d'un certain gène qui prédispose à la fois l'individu à fumer *et* à développer une maladie pulmonaire [**34**]. La prévalence observée de maladies pulmonaires chez les fumeurs ne serait ainsi due qu'à une tierce variable, soit l'existence de ce gène. Dans cet exemple, la présence du gène joue le rôle de variable de confusion, i.e. une variable qui est associée à la fois à l'exposition (i.e. le fait d'être fumeur ou non) et à l'issue (i.e. développer le cancer du poumon ou non) de telle sorte qu'une association non causale entre l'exposition et l'issue est observée. Il est certain que toute étude visant à détecter une relation de causalité entre le fait de fumer et le développement d'une maladie pulmonaire devrait prendre en compte l'effet d'une telle variable de confusion. Mais avant tout, il importe d'avoir une idée claire de ce qu'on entend par causalité. Nous présentons l'approche par les variables contrefactuelles mise de l'avant par J. Neyman [**23**] et développée par D. Rubin [**28, 29, 30**].

Considérons la situation générale où sont donnés Ω une population d'intérêt, A une variable aléatoire dichotomique représentant l'application de deux traitements (expositions) alternatifs et Y une variable aléatoire réelle représentant l'issue. On considère également les objets Y^0, Y^1 appelés *issues contrefactuelles* ([**23, 28**]), où Y^a représente l'issue lorsque le traitement $A = a$ est appliqué. Ainsi, $Y^1 - Y^0$ compare la valeur que prend l'issue si on applique l'un ou l'autre des traitements. Si $Y^1(\omega) - Y^0(\omega) \neq 0$ pour un individu $\omega \in \Omega$, alors c'est nécessairement que le traitement reçu est la cause de cette différence chez l'individu ω. L'espérance $E[Y^1 - Y^0]$ est appelée *effet causal moyen* et représente l'effet causal au niveau de la population.

Une fois ces définitions établies, le premier défi de l'inférence causale est d'établir un lien entre les variables contrefactuelles et la variable Y. En effet, pour un individu donné $\omega \in \Omega$, on observe toujours que l'un ou l'autre de $Y^0(\omega)$ et $Y^1(\omega)$ sous la forme de $Y(\omega)$, selon que $A(\omega) = 0$ ou $A(\omega) = 1$, respectivement. Nous allons exposer et discuter des quatre conditions communément utilisées pour établir ce lien dans la population, mais d'abord, illustrons par un exemple les concepts définis ci-haut.

Exemple 1.1.1. *Dans une étude menée récemment* [**11, 12**]*, la population à l'étude est constituée des femmes enceintes souffrant d'asthme modéré persistant. Pour ces patientes, les deux traitements les plus communs sont soit une faible dose de corticostéroïdes inhalés (0-250 μg) combinés à des agonistes β_2 de longue durée, soit une dose moyenne de corticostéroïdes inhalés (250-500 μg)* [**15**]*. L'objectif de l'étude est de déterminer si le traitement par corticostéroïdes inhalés et agonistes β_2 est la cause d'un poids moyen faible chez les bébés en comparaison avec l'autre traitement. Notons Y le poids du bébé à la naissance et A la variable indicatrice qui vaut 1 si une femme a reçu le traitement par corticostéroïdes inhalés et agonistes β_2 et 0 si elle a reçu l'autre traitement. Au niveau individuel, on dira que le traitement $A = 1$ est la* cause *d'une diminution du poids de son bébé si le poids Y^1 lorsqu'une femme se voit administrer le traitement $A = 1$ est inférieur au poids Y^0 lorsque cette même femme se voit administrer le traitement $A = 0$. Autrement dit, on peut s'imaginer qu'au moment de choisir un traitement, l'univers se sépare en deux univers parallèles. Dans le premier, la femme s'expose au traitement $A = 1$ et dans le second, elle s'expose au traitement $A = 0$. Dans l'univers $A = 1$, la femme accouche d'un bébé de poids $Y = Y^1$ tandis que*

dans le second univers, elle accouche d'un bébé de poids $Y = Y^0$. *Si* $Y^1 \neq Y^0$, *alors c'est nécessairement que le traitement est la cause de cette différence.*

Comme nous le verrons à la section 1.2, les conditions (C1)-(C4) ci-bas sont suffisantes pour estimer l'effet causal moyen. Les conditions (C1) et (C2) sont appelées SUTVA (pour Stable Unit Treatment Value Assumption) par D. Rubin **[30, 40]**.

(C1) [Non-interférence **[5]**] L'issue pour un individu donné ne dépend pas du traitement reçu par un quelconque autre individu de la population.

Cette supposition est implicite dans la notation Y^a, qui suppose que seule la valeur a prise par la variable A détermine la valeur de l'issue.

(C2) [Cohérence **[23, 26]**] Il existe une unique version des traitements $A = 0$ et $A = 1$, de sorte que les issues contrefactuelles Y^0,Y^1 sont bien définies avec la propriété $A = a \Rightarrow Y^a = Y$.

L'exposition doit être définie de façon précise pour que les variables Y^1, Y^0 soient elles aussi bien définies. Une manière de définir précisément l'exposition dans l'exemple précédent serait de considérer une femme comme exposée au traitement $A = a$ à condition qu'elle ait commencé à prendre le traitement a au plus deux semaines après le début de la période gestationelle et à raison d'au moins 1 fois par jour jusqu'à l'accouchement. Il importe aussi que l'exposition soit définie de façon appropriée en relation avec la question de recherche, sans quoi on peut aboutir à des conclusions erronées. Par exemple, si A était définie comme représentant le traitement suivi par la mère *au jour de l'accouchement*, alors il se peut que certaines femmes aient suivi le traitement avec corticostéroïdes inhalés et agonistes β_2 pendant presque toute leur grossesse, puis ont changé de traitement peu de temps avant d'accoucher.

Plus généralement, s'il existe différentes versions du traitement pouvant affecter l'issue, alors les variables Y^1, Y^0 ne sont pas bien définies. Notons qu'il est tout de même possible d'établir une relation de causalité entre une issue et un traitement ayant plusieurs versions **[40]**.

Tous les types d'expositions ne se prêtent pas bien au jeu. Par exemple, l'interprétation de l'effet causal de la race, du sexe, de la croyance religieuse, de l'obésité, etc. d'un individu

est chose délicate dû au fait que ces caractéristiques ne correspondent pas à une intervention clairement définie. Plusieurs questions entourant ces caractéristiques comme cause font l'objet de débat dans la communauté scientifique. Pour ne citer qu'une poignée d'articles, on peut consulter **[16, 17, 38, 37, 41, 42]**.

(C3) [Interchangeabilité conditionnelle] Il existe un vecteur aléatoire $L = (L_1, \ldots, L_r)$ tel que $A \perp\!\!\!\perp Y^a | L$ (i.e. conditionnellement à L, Y^a est indépendant de A). Nous dirons que les variables aléatoires $L_1, \ldots, L_r$ forment un ensemble *suffisant*.

Afin de comprendre la nature de cette condition, considérons la condition plus forte d'interchangeabilité. Celle-ci stipule l'indépendance de A et Y^a. Le nom est dû au fait que si A et Y^a sont indépendants, alors $E[Y^a|A = 0] = E[Y^a|A = 1]$, i.e. on peut échanger les groupes $A = 0$ et $A = 1$ sans modifier l'espérance de Y^a, ce qui est souhaitable car, en vertu de (C1), elle implique $E[Y^a] = E[Y^a|A = a] = E[Y|A = a]$, ce qui permet d'exprimer l'effet causal moyen en terme des variables observées Y et A. Cependant, à moins d'être dans un contexte d'étude randomisée, la condition d'interchangeabilité est trop optimiste: il y aura généralement des facteurs de confusion L_i. On appelle *facteur de confusion* toute variable dont le processus d'exposition A et les issues contrefactuelles Y^0, Y^1 sont simultanément dépendantes.

Exemple 1.1.2. *Supposons que Y est une variable dichotomique qui vaut 1 si le patient meurt et 0 sinon. Supposons aussi que $A = 1$ représente un traitement en médecine qui comporte de grands risques et n'est utilisé qu'en dernier recours chez des patients proches de la mort. Dans cette situation, la probabilité de mourir est beaucoup plus grande chez les individus qui reçoivent le traitement que chez ceux qui ne le reçoivent pas: $E[Y^a|A = 1] > E[Y^a|A = 0]$.*

Par contre, si L satisfait à la condition (C3), on a

$$E[Y^a] = E[E[Y^a|L]] = E[E[Y^a|A = a,L]] = E[E[Y|A = a,L]]. \tag{1.1.1}$$

La condition (C3) est parfois appelée simplement pas de facteurs de confusion non mesurés car dire que la distribution conditionnelle de Y^a sur L et $A = 0$ coïncide avec la distribution conditionnelle de Y^a sur L et $A = 1$ est équivalent à dire que L contient tous les facteurs de confusion pour le couple (Y^a,A). La signification de (C3) est donc que, si on cherche à

estimer l'effet causal moyen, nous devons avoir accès en plus des valeurs observées pour Y et A, aux valeurs mesurées de tous les facteurs de confusion.

(C4) [Positivité] Pour L un ensemble suffisant de facteurs de confusion, la distribution conditionnelle $p_{A|L}(a|\ell)$ est positive dès que $p_L(\ell) > 0$.

Intuitivement, cette condition stipule que tous les individus dans la population à l'étude doivent avoir une chance non nulle de recevoir chacun des traitements. Supposons au contraire que $p_{A|L}(0|\ell) = 0$ pour une certaine strate $L = \ell$ non vide de la population Ω. Alors toute étude visant à estimer $E[Y^0]$ par $E[E[Y|A = 0,L]]$ négligerait nécessairement toute la strate $L = \ell$ et le résultat serait davantage une estimation de $E[Y^0]$ pour la population $\Omega\backslash(L = \ell)$ plutôt qu'une estimation pour Ω.

Avant de conclure cette section, mentionnons que l'effet causal moyen $E[Y^1 - Y^0]$ n'est pas la seule quantité pouvant servir à identifier une relation de causalité. Par exemple, si Y est une variable dichotomique, on préférera généralement le rapport de risque $P[Y^1 = 1]/P[Y^0 = 1]$ ou le rapport de cote $= (P[Y^1 = 1]/P[Y^1 = 0])/(P[Y^0 = 1]/P[Y^0 = 0])$.

1.2. Estimation de l'effet causal

Considérons à nouveau la situation la plus simple où l'exposition A est dichotomique et l'issue Y est une variable aléatoire réelle. Dans une étude randomisée, un échantillon de taille n est prélevé aléatoirement au sein de la population d'intérêt (ou de manière représentative de celle-ci). Les individus recrutés se voient alors assigner le traitement $A = 0$ ou $A = 1$ de façon complètement aléatoire. Dans ce contexte, la condition d'interchangeabilité conditionnelle (C3) est automatiquement vérifiée puisqu'il n'existe aucune variable de confusion. Le fait que le processus d'assignation du traitement est complètement aléatoire garantit que $P[A = 1]$ est une simple constante et donc ne dépend d'aucune variable extérieure. Du même coup, la condition de positivité (C4) automatiquement vérifiée elle aussi. Par conséquent, à condition que le protocole de l'étude ait été élaboré de manière à ce que les conditions (C1) et (C2) soient vérifiées, alors l'effet causal moyen peut être estimé simplement par

$$n_1^{-1} \sum_{A_i=1} Y_i - n_0^{-1} \sum_{A_i=0} Y_i, \tag{1.2.1}$$

où $n_a = \#(A_i = a)$ et où $\sum_{A_i=a}$ signifie que l'on somme sur les indices i tels que $A_i = a$.

Mais les essais randomisés contrôlés ont leurs limites, qu'elles soient de nature pratique, financière ou éthique. Par exemple, l'industrie pharmaceutique et les chercheurs sont réticents à mener des études sur les femmes enceintes de peur de mettre à risque la santé de deux populations vulnérables (mère et fœtus) **[31]**. Dans de telles situations, il faut se rabattre sur l'analyse de données préexistantes. On parle alors d'étude observationelle. Contrairement aux études randomisées, on ne peut s'attendre à ce que les facteurs de confusion soient distribués identiquement dans les deux groupes de traitement. Dans ce cas, il est clair que l'estimateur (1.2.1) ne peut être utilisé. On peut remédier à cela en pondérant l'issue Y_i d'un individu du groupe $A = a$ par l'inverse de la probabilité de traitement $P[A = a|L = L_i]$, où $L = (L_1, \ldots, L_r)^T$ est un ensemble de facteurs de confusion vérifiant les conditions (C3), (C4) et $L_i = (L_{i1}, \ldots, L_{ir})^T$ est la valeur de la variable L pour le i-ème individu. De cette façon, on élargit virtuellement les deux groupes de traitement de telle sorte que, pour $n = n_0 + n_1$ suffisamment grand, on s'attend à ce que les groupes virtuels soient de taille n et que la variable L y soit distribuée identiquement. En effet, supposons que $P[A = 1|L = \ell] = p$. Alors, $P[A = 0|L = \ell] = 1 - p$ et un individu dans le groupe $A = 1$ avec $L = \ell$ reçoit la pondération p^{-1} tandis qu'un individu dans le groupe $A = 0$ avec $L = \ell$ reçoit la pondération $(1 - p)^{-1}$. Sur k individus avec $L = \ell$, on s'attend à en retrouver pk dans le groupe $A = 1$ et $(1 - p)k$ dans le groupe $A = 0$, si bien que dans les groupes virtuels, les individus avec $L = \ell$ sont représentés au nombre de $pk \times p^{-1} = k$ dans le groupe $A = 0$ et au nombre de $(1 - p)k \times (1 - p)^{-1} = k$ dans le groupe $A = 1$.

L'analogue naturel de l'estimateur (1.2.1) pour une étude observationnelle est donc

$$\begin{aligned} & n^{-1} \sum_{A_i=1} \frac{Y_i}{P[A = 1|L = L_i]} - n^{-1} \sum_{A_i=0} \frac{Y_i}{P[A = 0|L = L_i]} \\ & = n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i}{P[A = 1|L = L_i]} - \frac{(1 - A_i)Y_i}{1 - P[A = 1|L = L_i]} \end{aligned}$$

et porte le nom d'*estimateur IPTW* pour Inverse Probability of Treatment Weighting ou encore *estimateur de Horvitz-Thompson* **[18]**. La quantité $P[A = 1|L]$ est appelée *score de propension* et sera notée $g(L)$.

Proposition 1.2.1. *Sous les présuppositions causales (C1)-(C4), l'estimateur IPTW est convergent.*

DÉMONSTRATION. Par la loi des grands nombres, $n^{-1}\sum_{i=1}^{n}\frac{A_iY_i}{g(L_i)}$ converge en probabilité vers $E[g(L)^{-1}AY]$. Or,

$$\begin{aligned} E[g(L)^{-1}AY] &= E[E[g(L)^{-1}AY|L]] \\ &= E[g(L)^{-1}E[AY|L]] \\ &= E[E[Y|L,A=1]] \\ &= E[Y^1] \quad \text{par (1.1.1).} \end{aligned}$$

On passe de la deuxième ligne à la troisième ligne en développant:

$$\begin{aligned} E[AY|L] &= E[E[AY|L,A]|L] \\ &= E[A[E[Y|L,A]|L] \\ &= 1\cdot E[Y|L,A=1]P[A=1|L] + 0\cdot E[Y|L,A=0]P[A=0|L] \\ &= E[Y|L,A=1]g(L). \end{aligned} \tag{1.2.2}$$

On procède de même pour montrer que $n^{-1}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-g(L_i)}$ converge en probabilité vers $E[Y^0]$. □

On peut consulter **[6]** pour plus d'information sur l'estimateur IPTW. Bien que ce dernier soit sans doute le plus naturel, il n'est pas le plus *efficace*, en général. À la prochaine section, on introduit une classe très large d'estimateurs convergents auxquels l'estimateur IPTW appartient et on montre comment trouver celui ayant la plus petite variance.

1.3. Estimateurs asymptotiquement linéaires

Dans cette section, nous présentons la théorie des estimateurs asymptotiquement linéaires (AL) suivant **[35]**. La linéarité asymptotique est une propriété désirable chez un estimateur car elle signifie que l'estimateur se comporte asymptotiquement comme une moyenne empirique. En particulier, un estimateur AL exhibera un taux de convergence de $\sqrt{n}$ et une distribution asymptotiquement normale.

Considérons la situation générale où Ω est un espace de probabilité et Z est une variable aléatoire sur Ω à valeurs dans un espace mesuré $\mathcal{T}$. On choisit un *modèle* pour Z, soit une partie $\mathcal{M}$ de l'espace $\mathcal{P}$ de toutes les distributions de probabilité sur $\mathcal{T}$ que l'on sait contenir la distribution P_Z de Z. On considère également une application $\Psi : \mathcal{M} \to \mathbb{R}^d$ telle

que $\psi_0 := \Psi(P_Z)$ est le paramètre que l'on cherche à estimer (le paramètre *ciblé*). Il faut mentionner ici que tous les résultats de cette section ne font référence au modèle $\mathcal{M}$ que dans un voisinage de P_Z. Ainsi, on peut toujours rétrécir le modèle $\mathcal{M}$ au besoin pour qu'il satisfasse aux conditions de régularité requises.

Exemple 1.3.1. *Dans un problème d'inférence causale tel que discuté à la section précédente, on a $Z = (Y,A,L)$ où Y représente l'issue, A est une exposition binaire, $L = (L_1,\ldots,L_r)^T$ est un ensemble suffisant de facteurs de confusion, $\psi_0 = E[Y^1 - Y^0]$ est l'effet causal moyen et les conditions (C1)-(C4) sont en vigueur. Étant donnée une distribution de probabilité P sur $\mathcal{T}$ et une mesure dominante ν, on peut factoriser la fonction de densité $p(y,a,\ell) = \frac{dP}{d\nu}$ comme suit:*

$$p(y,a,\ell) = p_{Y|A,L}(y|a,\ell)p_{A|L}(a|\ell)p_L(\ell).$$

On peut prendre pour Ψ l'application

$$\Psi(P) = \int \left(\int y p_{Y|A,L}(y|a=1,\ell)dy - \int y p_{Y|A,L}(y|a=0,\ell)dy \right) p_L(\ell)d\ell \tag{1.3.1}$$

car pour $P = P_Z$, cette expression est égale à $E[E[Y|A=1,L] - E[Y|A=0,L]]$, ce qui par (1.1.1) *est bien l'effet causal moyen.*

Étant donné un échantillon iid $Z_1,\ldots,Z_n$ de Z, un estimateur $\hat{\psi}_n \equiv \hat{\psi}_n(Z_1,\ldots,Z_n)$ pour ψ_0 est appelé *estimateur asymptotiquement linéaire* (AL) s'il existe une fonction mesurable $\varphi : \mathcal{T} \to \mathbb{R}^d$ appelée *fonction d'influence* telle que

(i) $\hat{\psi}_n - \psi_0 = \sum_{i=1}^n \frac{\varphi(Z_i)}{n} + o_p(n^{-1/2})$,

(ii) $E[\varphi(Z)] = 0$,

(iii) $0 < |\det(Var[\varphi(Z)])| < +\infty$.

La condition (i) peut également s'écrire

$$\sqrt{n}(\hat{\psi}_n - \psi_0) = \sum_{i=1}^n \frac{\varphi(Z_i)}{\sqrt{n}} + o_p(1).$$

En vertu du théorème de la limite centrale, on a

$$\sum_{i=1}^n \frac{\varphi(Z_i)}{\sqrt{n}} \xrightarrow{d} N\left(0,Var[\varphi(Z)]\right).$$

Par le théorème de Slutzky, on a alors

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{d} N(0,Var[\varphi(Z)]). \tag{1.3.2}$$

Dans ce sens, la fonction d'influence gouverne le comportement asymptotique de l'estimateur. Par exemple, pour ψ_0 scalaire, un intervalle de confiance asymptotique de niveau $1-\alpha$ est

$$\hat{\psi}_n \pm z_{\alpha/2}\sqrt{\frac{Var[\varphi(Z)]}{n}}, \tag{1.3.3}$$

où $P[X > z_{\alpha/2}] = \alpha/2$ pour $X \sim N(0,1)$. Nous appellerons *variance asymptotique* de l'estimateur $\hat{\psi}_n$ la quantité $n^{-1}Var[\varphi(Z)]$.

Remarque 1.3.2. *(1) La fonction d'influence dépend en général du paramètre ψ_0. Ainsi, on écrira parfois $\varphi(Z;\psi_0)$ pour mettre ce fait en évidence.*

(2) La condition (ii) est une condition de normalisation: si $\hat{\psi}_n$ vérifie (i) avec fonction φ, alors $\hat{\psi}'_n := \hat{\psi}_n - E[\varphi(Z)]$ vérifie (i) avec fonction d'influence $\varphi' := \varphi - E[\varphi(Z)]$ et $E[\varphi'(Z)] = 0$.

(3) Si $\hat{\psi}_n$ et $\hat{\psi}'_n$ sont deux estimateurs AL pour ψ_0 et ψ'_0 avec fonction d'influence respectives φ et φ', alors $a\psi_n + b\psi'_n$ est un estimateur AL pour $a\psi_0 + b\psi'_0$ avec fonction d'influence $a\varphi + b\varphi'$.

(4) Deux estimateurs AL $\hat{\psi}_n$ et $\hat{\psi}'_n$ de ψ_0 ont (presque partout) la même fonction d'influence si et seulement s'ils sont asymptotiquement équivalents au sens où $\hat{\psi}'_n = \hat{\psi}_n + o_p(n^{-1/2})$. En particulier, la fonction d'influence d'un estimateur AL est presque sûrement unique. Pour démontrer cela, on note d'abord que si $\hat{\psi}_n$ et $\hat{\psi}'_n$ ont des fonctions d'influence qui coïncident presque partout, alors par soustraction des relations (i), on obtient directement $\hat{\psi}'_n = \hat{\psi}_n + o_p(n^{-1/2})$. Inversement, si $\hat{\psi}'_n - \hat{\psi}_n = o_p(n^{-1/2})$, alors on obtient une contradiction en supposant que les fonctions d'influence ne coïncident pas presque partout car dans ce cas, le théorème de la limite centrale veut que $n^{-1/2}\sum_{i=1}^n(\varphi-\varphi')(Z_i) \xrightarrow{d} N(0, Var[(\varphi-\varphi')(Z)])$. Par ailleurs, la différence des relations (i) donne $n^{-1}\sum_{i=1}^n(\varphi-\varphi')(Z_i) = o_p(n^{-1/2})$ donc en particulier $n^{-1/2}\sum_{i=1}^n(\varphi-\varphi')(Z_i) \xrightarrow{d} 0$.

(5) Un estimateur AL est convergent, i.e. $\hat{\psi}_n \xrightarrow{p} \psi_0$. En effet, si on écrit la condition (i) sous la forme $\hat{\psi}_n - \psi_0 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\varphi(Z_i)}{\sqrt{n}} + o_p(n^{-1/2})$, alors par (1.3.2) et le théorème de Slutzy, on en déduit $\hat{\psi}_n - \psi_0 \xrightarrow{d} 0$. Or, la convergence en distribution vers 0 est équivalent à la convergence en probabilité vers 0.

Exemple 1.3.3 (**[35]** p.22)**.** *Soit Z une variable aléatoire réelle avec $E[Z] = \mu$, $Var[Z] = \sigma^2$.*

(1) Un estimateur AL pour $\psi_0 = \mu$ est la moyenne empirique $\hat{\mu}_n = n^{-1}(Z_1 + \ldots + Z_n)$ avec fonction d'influence $\varphi(Z) = Z - \mu$.

(2) Un estimateur AL pour $\psi_0 = \sigma^2$ est la variance empirique $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (Z_i - \hat{\mu}_n)^2$. Quelques manipulations algébriques permettent d'écrire

$$\hat{\sigma}_n^2 - \sigma^2 = n^{-1}\sum_{i=1}^n ((Z_i - \mu)^2 - \sigma^2) - (\hat{\mu}_n - \mu)^2.$$

De la loi faible des grands nombres, il découle $\hat{\mu}_n - \mu \xrightarrow{p} 0$ tandis que le théorème de la limite centrale permet d'écrire $\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} N(0,\sigma^2)$. Par le théorème de Slutzky, on a donc $\sqrt{n}(\hat{\mu}_n - \mu)^2 \xrightarrow{p} 0$. Ainsi, $\hat{\sigma}_n^2$ est AL avec fonction d'influence $\varphi(Z) = (Z - \mu)^2 - \sigma^2$.

Exemple 1.3.4 (m-estimateurs)**.** */***[35]** §3.2] Soit $Z : \Omega \to \mathcal{T}$ une variable aléatoire. On propose un modèle paramétrique $\{P_\theta | \theta \in \Theta \subset \mathbb{R}^d\}$ pour $P_Z = P_{\theta_0}$. Soit $m : \mathcal{T} \times \mathbb{R}^N \to \mathbb{R}^N$ une fonction vérifiant la condition de normalisation $E_\theta[m(Z,\theta)] = 0$. Étant donné un échantillon iid, $Z_1,\ldots,Z_n$, de Z, un* m-estimateur *de θ_0 est une solution $\hat{\theta}_n = \hat{\theta}_n(Z_1,\ldots,Z_n)$ de l'*équation estimante

$$\sum_{i=1}^n m(Z_i,\theta) = 0.$$

Montrons qu'un m-estimateur est AL et calculons sa fonction d'influence. Pour ceci, on utilise un développement de Taylor à l'ordre 1 autour de θ_0:

$$0 = \sum_{i=1}^n m(Z_i;\hat{\theta}_n) = \sum_{i=1}^n m(Z_i,\theta_0) + \sum_{i=1}^n \frac{\partial m(Z_i;\hat{\theta}_n^*)}{\partial \theta^T}(\hat{\theta}_n - \theta_0), \tag{1.3.4}$$

où $\hat{\theta}_n^*$ se trouve sur la ligne droite joignant θ_0 à $\hat{\theta}_n$ dans $\mathbb{R}^N$. Sous certaines conditions de régularité (cf. **[35]** *section 3.2), $\hat{\theta}_n \xrightarrow{p} \theta_0$ et $n^{-1}\sum_{i=1}^n \frac{\partial m(Z_i;\hat{\theta}_n^*)}{\partial \theta^T} \xrightarrow{p} E_{\theta_0}\left[\frac{\partial m(Z;\theta_0)}{\partial \theta^T}\right]$. Par le théorème de l'application continue, on a donc*

$$\left(n^{-1}\sum_{i=1}^n \frac{\partial m(Z_i;\hat{\theta}_n^*)}{\partial \theta^T}\right)^{-1} \xrightarrow{p} E_{\theta_0}\left[\frac{\partial m(Z;\theta_0)}{\partial \theta^T}\right]^{-1},$$

et on peut donc écrire (1.3.4) *sous la forme*

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &= -\left(n^{-1}\sum_{i=1}^n \frac{\partial m(Z_i;\hat{\theta}_n^*)}{\partial \theta^T}\right)^{-1}\left(n^{-1/2}\sum_{i=1}^n m(Z_i;\theta_0)\right) \\ &= -E_{\theta_0}\left[\frac{\partial m(Z;\theta_0)}{\partial \theta^T}\right]^{-1}\left(n^{-1/2}\sum_{i=1}^n m(Z_i;\theta_0)\right) + o_p(1).\end{aligned}$$

Il découle que θ_n est AL de fonction d'influence

$$\varphi(z) = -E_{\theta_0}\left[\frac{\partial m(z;\theta_0)}{\partial \theta^T}\right]^{-1} m(z;\theta_0).$$

La variance asymptotique de $\hat{\theta}_n$ est

$$\begin{aligned} n^{-1}Var[\varphi(Z)] &= n^{-1}E[\varphi(Z)\varphi(Z)^T] \\ &= n^{-1}E\left[\frac{\partial m(Z;\theta_0)}{\partial \theta^T}\right]^{-1} Var[m(Z,\theta_0)]\left(E\left[\frac{\partial m(Z;\theta_0)}{\partial \theta^T}\right]^{-1}\right)^T. \end{aligned}$$

Sous certaines conditions de régularité, un estimateur convergent de la variance asymptotique de $\hat{\theta}_n$ est donné par

$$\hat{V} = n^{-1}\left(n^{-1}\sum_{i=1}^{n}\frac{\partial m(Z_i;\hat{\theta}_n)}{\partial \theta^T}\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} m(Z_i;\hat{\theta}_n)^{\otimes 2}\right)\left(n^{-1}\sum_{i=1}^{n}\frac{\partial m(Z_i;\hat{\theta}_n)}{\partial \theta}\right)^{-1}. \tag{1.3.5}$$

L'estimateur (1.3.5) *est communément appellé* estimateur sandwich.

Exemple 1.3.5. *Sous certaines conditions de régularité, le vecteur score*

$$S(Z;\theta) = \frac{\partial \log p_Z(Z;\theta)}{\partial \theta} = \left(\frac{\partial \log p_Z(Z;\theta)}{\partial \theta_1},\ldots,\frac{\partial \log p_Z(Z;\theta)}{\partial \theta_N}\right)^\top$$

vérifie $E_\theta[S(Z;\theta)] = 0$. Le m-estimateur correspondant est l'estimateur du maximum de vraisemblance:

$$\sum_{i=1}^{n} S_j(Z_i;\hat{\theta}_n) = \sum_{i=1}^{n}\frac{\partial \log p_Z(Z_i;\hat{\theta}_n)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\bigg|_{\hat{\theta}_n} \log p_{Z_1,\ldots,Z_n}(Z_1,\ldots,Z_n;\theta).$$

La fonction d'influence est

$$\varphi(z) = I(\theta_0)^{-1}S(z;\theta_0),$$

où $I(\theta_0) = -E_{\theta_0}\left[\frac{\partial S(Z;\theta_0)}{\partial \theta^\top}\right]$ est, sous certaines conditions de régularité, égal à l'information de Fisher $Var_{\theta_0}[S(Z;\theta_0)]$. En particulier, $Var[\varphi(Z)] = I(\theta_0)^{-1}$ et en vertu de (1.3.2)*, on retrouve le résultat classique voulant que la variance asymptotique de l'estimateur du maximum de vraisemblance atteigne la borne de Cramér-Rao.*

Exemple 1.3.6 (**[20]**). *L'estimateur IPTW*

$$\hat{\psi}_n = n^{-1}\sum_{i=1}^{n}\frac{A_iY_i}{P[A=1|L=L_i]} - \frac{(1-A_i)Y_i}{1-P[A=1|L=L_i]}$$

pour $\psi_0 = E[Y^1 - Y^0]$ *est AL avec fonction d'influence*

$$\varphi_{IPTW}(Z) = \frac{AY}{P[A=1|L]} - \frac{(1-A)Y}{1-P[A=1|L=L]} - E[Y^1 - Y^0].$$

Dans ce cas-ci, le terme $o_p(n^{-1/2})$ *est* 0.

Dans la majorité des applications, le score de propension $g(L)$ *n'est pas connu et il faut l'estimer. Si un modèle paramétrique* $g(L;\alpha)$, $\alpha \in \mathbb{R}^N$ *est utilisé et que le paramètre* α *est estimé à l'aide du m-estimateur*

$$0 = n^{-1} \sum_{i=1}^{n} m_g(Z_i; \hat{\alpha}_n)$$

(par exemple par maximum de vraisemblance), alors l'estimateur

$$n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i}{g(L;\hat{\alpha}_n)} - \frac{(1-A_i)Y_i}{1-g(L_i;\hat{\alpha}_n)}$$

est obtenu par résolution de l'équation estimante

$$0 = n^{-1} \sum_{i=1}^{n} m(Z_i; \theta),$$

où $\theta = (\psi, \alpha)$ *et*

$$m(Z;\theta) = \{m_{IPTW}(Z;\psi,\alpha), m_g(Z;\alpha)^\top\}^\top$$

avec

$$m_{IPTW}(Z;\psi,\alpha) = \frac{AY}{g(L;\alpha)} - \frac{(1-A)Y}{1-g(L;\alpha)} - \psi.$$

Par conséquent, la fonction d'influence est (cf. Exemple 1.3.4)

$$\varphi(Z) = -E\left[\frac{\partial m(Z;\theta_0)}{\partial \theta^\top}\right]^{-1} m(Z;\theta_0).$$

En utilisant le fait voulant que

$$\begin{pmatrix} A & B \\ 0 & C \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}BC^{-1} \\ 0 & C^{-1} \end{pmatrix},$$

on obtient

$$E\left[\frac{\partial m}{\partial \theta^\top}\right]^{-1} = \begin{pmatrix} -1 & E\left[\frac{\partial m_{IPTW}}{\partial \alpha^\top}\right] E\left[\frac{\partial m_g}{\partial \alpha^\top}\right]^{-1} \\ 0 & E\left[\frac{\partial m_g}{\partial \alpha^\top}\right]^{-1} \end{pmatrix},$$

d'où, en utilisant le fait que $m_{IPTW}(Z;\psi_0,\alpha_0) = \varphi_{IPTW}(Z)$,

$$\varphi(Z) = \varphi_{IPTW}(Z) \\ - E\left[\frac{\partial m_{IPTW}(Z;\psi_0,\alpha_0)}{\partial \alpha^\top}\right] E\left[\frac{\partial m_g(Z;\alpha_0)}{\partial \alpha^\top}\right]^{-1} m_g(Z;\alpha_0).$$

L'exemple précédent illustre un phénomène général, à savoir que si $\varphi(z;\psi,\eta)$ est une fonction d'influence potentielle (i.e. elle vérifie les conditions (ii) et (iii) de la Définition 1.3) dépendant du paramètre d'intérêt ψ et aussi d'un autre paramètre η (possiblement de dimension infinie), alors sous certaines conditions de régularité en plus de conditions sur une estimation $\hat{\eta}$ de η, l'estimateur $\hat{\psi}_n$ obtenu par résolution de l'équation estimante

$$0 = \sum_i^n \varphi(Z_i;\psi,\hat{\eta})$$

sera AL avec fonction d'influence relié à φ d'une manière dépendant de la forme de φ et de la façon dont η est estimé [**20**].

1.4. La fonction d'influence efficace

Dans les sections 1.4.1 et 1.4.2, on cherchera à identifier la fonction d'influence φ_{eff} qui minimise $Var[\varphi(Z)]$. Par (1.3.2), un estimateur AL ayant cette fonction d'influence est *efficace* au sens où sa variance asymptotique est minimale parmi la classe des estimateurs AL. En fait, nous verrons que la variance de la fonction d'influence efficace est la borne de Cramer-Rao.

1.4.1. Le cas d'un modèle paramétrique

Comme à la section précédente, on considère un espace probabilisé Ω et une variable aléatoire $Z : \Omega \to \mathcal{T}$. On choisit un modèle *paramétrique* pour Z, soit un modèle $\mathcal{M}$ de la forme $\{P_\theta | \theta \in \Theta\}$ pour $\Theta \subset \mathbb{R}^N$ et posons $P_Z = P_{\theta_0}$. Soit aussi $\Psi : \mathcal{M} \to \mathbb{R}^d$ une application telle que $\Psi(P_{\theta_0}) = \psi_0$. Si la correspondance $\theta \mapsto P_\theta$ est bijective dans un voisinage de θ_0 alors, on peut sans perdre de généralité identifier $\mathcal{M}$ à Θ et utiliser la notation $\Psi(\theta)$ au lieu de $\Psi(P_\theta)$.

Par souçi de simplicité, supposons pour le reste de cette section que le paramètre que l'on cherche à estimer est un scalaire, i.e. $d = 1$. Notons $(\mathcal{H}_{\theta_0},\langle\cdot,\cdot\rangle)$ le sous-espace des éléments

de moyenne nulle de l'espace de Hilbert $L^2(\mathcal{T},P_Z)$. Explicitement, on a

$$\mathcal{H}_{\theta_0} = \{\varphi : \mathcal{T} \to \mathbb{R} \mid E_{\theta_0}[\varphi(Z)] = 0, E_{\theta_0}[\varphi(Z)^2] < \infty\},$$

$$\langle \varphi_1,\varphi_2 \rangle = E_{\theta_0}[\varphi_1(Z)\varphi_2(Z)].$$

Le sous-espace $\mathcal{H}_{\theta_0}$ est fermé dans $L^2(\mathcal{T},P_Z)$, ce qui en fait un sous-espace de Hilbert avec produit intérieur

$$\langle \varphi_1,\varphi_2 \rangle = Cov_{\theta_0}[\varphi_1(Z),\varphi_2(Z)].$$

On définit l'*espace tangent* $T_{\theta_0}\mathcal{M}$ comme le sous-espace de $\mathcal{H}_{\theta_0}$ engendré par les composantes du vecteur score:

$$T_{\theta_0}\mathcal{M} = \left\{ \sum_{i=1}^{N} a_i S_i(z;\theta_0) \mid a_i \in \mathbb{R} \right\}, \quad S_i(z;\theta_0) = \left. \frac{\partial \log p(z;\theta)}{\partial \theta_i} \right|_{\theta=\theta_0}. \tag{1.4.1}$$

Le produit intérieur relativement à cette base est donné par la matrice d'information de Fisher:

$$\langle S_i(z;\theta_0), S_j(z;\theta_0) \rangle = I(\theta_0)_{ij}.$$

Prenons un instant pour situer cette observation dans le cadre de la géométrie différentielle auquel elle appartient. Si $\mathcal{F}$ dénote l'espace vectoriel des fonctions sur $\mathcal{T}$, alors un choix d'une mesure dominante ν sur $\mathcal{T}$ induit un plongement de $\mathcal{M}$ dans $\mathcal{F}$ via $P_\theta \mapsto \log p(z;\theta)$, où $p(z;\theta) = \frac{dP_\theta}{d\nu}$. Autrement dit, une distribution de probabilité P se voit associée au logarithme de sa dérivée de Radon-Nykodym. Sous certaines conditions de régularité, cette réalisation de $\mathcal{M}$ admet en tout point P_θ un espace tangent $T_{P_\theta}\mathcal{M}$ (ou plus simplement $T_\theta\mathcal{M}$) et leur réunion forme un fibré vectoriel $T\mathcal{M}$ sur $\mathcal{M}$ de fibre $T_\theta\mathcal{M}$ appelé fibré tangent **[21]**. Il existe un autre fibré vectoriel sur $\mathcal{M}$, soit le fibré en espaces de Hilbert $\mathcal{H}$ de fibre $\mathcal{H}_{P_\theta} = \mathcal{H}_\theta = \{\varphi \in L^2(\mathcal{T},P_\theta) \mid E_\theta[\varphi(Z)] = 0\}$, et sous les conditions de régularité habituelles garantissant que le vecteur score est d'espérance nulle, le fibré tangent est un sous-fibré de $\mathcal{H}$. En particulier, au point $P_{\theta_0} \in \mathcal{M}$, on a $\mathcal{H}_{\theta_0} = \mathcal{H}_0$ et $T_{\theta_0}\mathcal{M} = T_0$, ce qui explique la notation et la terminologie. Le produit intérieur sur $\mathcal{H}$ induit (par simple restriction) une métrique riemannienne g sur $\mathcal{M}$ donnée par l'information de Fisher et appelée *métrique de Fisher-Rao* ou *métrique d'information*:

$$g_\theta = \langle \cdot,\cdot \rangle|_{T_\theta\mathcal{M}} = \sum_{i,j=1}^{N} I(\theta)_{ij} S_i(z;\theta)^* \otimes S_j(z;\theta)^*.$$

Ceci signifie simplement que si $\varphi_1, \varphi_2 \in T_\theta\mathcal{M}$ sont deux vecteurs tangents représentés respectivement par les vecteurs v_1,$v_2 \in \mathbb{R}^N$ par rapport à la base $\{S_i(z;\theta)\}$ de $T_\theta\mathcal{M}$, alors

$$g_\theta(\varphi_1,\varphi_2) = v_1^\top I(\theta) v_2.$$

Remarque 1.4.1. *Ce point de vue géométrique consistant à considérer l'information de Fisher comme métrique riemannienne remonte à C. R. Rao* **[25]** *et a été développé par B. Efron* **[8]**, *S. Amari* **[1, 2, 3]** *et d'autres. Pour ne citer qu'un résultat de la théorie, la connexion de Lévy-Civita associée à g s'étend à une connexion sur le fibré hilbertien $\mathcal{H}$ dont la* courbure *caractérise l'appartenance de $\mathcal{M}$ à la famille exponentielle.*

En géométrie riemannienne, on retrouve la notion de *gradient riemannien* qui généralise la notion euclidienne bien connue. Si f est une fonction réelle sur $\mathcal{M}$, son gradient riemannien au point θ_0 est le vecteur $\mathrm{grad}_g f \in T_{\theta_0}\mathcal{M}$ tel que

$$g_{\theta_0}(\mathrm{grad}_g f, V) = \sum_{i=1}^{N} \frac{\partial f(\theta_0)}{\partial \theta_i} v_i \tag{1.4.2}$$

pour tout vecteur tangent $V = \sum_{i=1}^N v_i S_i(z;\theta_0) \in T_{\theta_0}\mathcal{M}$. Le membre de droite représente la dérivée directionnelle de f dans la direction de V. De plus, par l'inégalité de Cauchy-Schwarz appliquée à V unitaire, ont voit que le gradient riemannien pointe dans la direction de croissance maximale de f.

On peut montrer que, à certaines conditions de régularité près[1], les éléments φ de $\mathcal{H}_0$ qui sont des fonctions d'influence pour l'estimation du paramètre ψ_0 sont précisément ceux qui vérifient ([**35**] Theorem 3.2 et §3.3):

$$\langle \varphi, S_i(z;\theta_0)\rangle = \frac{\partial \Psi(\theta_0)}{\partial \theta_i} \quad \forall i = 1,\ldots,N. \tag{1.4.3}$$

Dans le language de la géométrie différentielle, ce résultat s'énonce simplement en disant que *les éléments φ de $\mathcal{H}_0$ qui sont des fonctions d'influence pour l'estimation du paramètre ψ_0 sont précisément ceux dont la projection orthogonale sur T_0 est donnée par le gradient riemannien de Ψ*. En effet, tout élément $\varphi \in \mathcal{H}_0$ admet la décomposition orthogonale $\varphi = \varphi_0 + \varphi_\perp$ où $\varphi_0 \in T_0$ est la projection orthogonale de φ sur T_0 et $\varphi_\perp \in T_0^\perp$ est le complément orthogonal, et on a $\langle \varphi, S_i(z;\theta_0)\rangle = \langle \varphi_0, S_i(z;\theta_0)\rangle$ $\forall i$. Mais puisque $\{S_i(z;\theta_0)\}$

[1]Parmi les conditions de régularité entrant dans l'énoncé du théorème, on demande que l'estimateur appartienne à la sous-classe des estimateur AL appelés estimateur asymptotiquement linéaires réguliers (ALR) [**35**] (Definition 1).

forme une base de T_0, les nombres $\langle \varphi_0, S_i(z;\theta_0)\rangle$ $(i = 1,\ldots,N)$ caractérisent le vecteur φ_0. Plus précisément, si

$$\varphi_0 = \sum_{i=1}^{N} \varphi_{0i} S_i(z;\theta_0),$$

alors

$$\varphi_{0i} = \sum_{j=1}^{N} I(\theta_0)^{ij} \langle \varphi_0, S_j(z;\theta_0)\rangle, \tag{1.4.4}$$

où $I(\theta_0)^{ij} = (I(\theta_0)^{-1})_{ij}$. Or, de la définition (1.4.2), on voit bien que le gradient riemannien de Ψ vérifie lui aussi les relations (1.4.3). Donc (1.4.3) est équivalent à $\varphi_0 = \mathrm{grad}_g \Psi$. Le gradient riemannien $\mathrm{grad}_g \Psi$ est appelé *fonction d'influence efficace* **[35]** ou *gradient canonique* **[43]**. On le note φ_{eff} ou encore $D^*(P_Z)$ lorsqu'on veut mettre l'emphase sur le fait qu'il s'agît de la fonction d'influence *en* P_Z. En combinant (1.4.3) à (1.4.4), on obtient la formule suivante pour la fonction d'influence efficace:

$$\varphi_{eff}(z) = \sum_{i,j=1}^{N} \frac{\partial \Psi(\theta_0)}{\partial \theta_j} I(\theta_0)^{ij} S_i(z;\theta_0) \tag{1.4.5}$$

$$= \frac{\partial \Psi(\theta_0)}{\partial \theta^\top} I(\theta_0)^{-1} S(z;\theta_0). \tag{1.4.6}$$

Comme corollaire de la caractérisation (1.4.3) des fonctions d'influences, on obtient la formule suivante pour l'espace $FI(\mathcal{M},\psi_0)$ de toutes les fonctions d'influences:

$$FI(\mathcal{M},\psi_0) = \varphi_{eff} + T_0^{\perp}.$$

Une autre conséquence immédiate du point de vue géométrique que nous avons adopté est que la fonction d'influence efficace est la fonction d'influence ayant la plus petite variance. En effet, si $\varphi = \varphi_{eff} + \varphi_{\perp}$ est une fonction d'influence quelconque, on a

$$\begin{aligned} Var[\varphi(Z)] &= \langle \varphi(Z), \varphi(Z)\rangle \\ &= \langle \varphi_{eff}(Z), \varphi_{eff}(Z)\rangle + 2\langle \varphi_{eff}(Z), \varphi^{\perp}(Z)\rangle + \langle \varphi^{\perp}(Z), \varphi^{\perp}(Z)\rangle \\ &= \langle \varphi_{eff}(Z), \varphi_{eff}(Z)\rangle + \langle \varphi^{\perp}(Z), \varphi^{\perp}(Z)\rangle \\ &= Var[\varphi_{eff}(Z)] + Var[\varphi^{\perp}(Z)] \\ &\geq Var[\varphi_{eff}(Z)]. \end{aligned}$$

En fait, en utilisant (1.4.6), on voit que la variance de φ_{eff} atteint la borne de Cramer-Rao:

$$\begin{aligned} Var[\varphi_{eff}(Z)] &= \left(\frac{\partial\psi}{\partial\theta^\top} I(\theta_0)^{-1}\right) I(\theta_0) \left(\frac{\partial\psi}{\partial\theta^\top} I(\theta_0)^{-1}\right)^\top \\ &= \frac{\partial\psi}{\partial\theta^\top} I(\theta_0)^{-1} I(\theta_0) I(\theta_0)^{-1} \frac{\partial\psi}{\partial\theta} \\ &= \frac{\partial\psi}{\partial\theta^\top} I(\theta_0)^{-1} \frac{\partial\psi}{\partial\theta}. \end{aligned}$$

Pour résumer, *la fonction φ_{eff} donnée par* (1.4.6) *est l'unique fonction d'influence qui minimise la variance et elle s'obtient d'une fonction d'influence quelconque φ par projection orthogonale sur l'espace tangent.*

Exemple 1.4.2. *La fonction d'influence pour l'estimateur par maximum de vraisemblance (cf. Exemple 1.3.5) est la fonction d'influence efficace puisqu'elle s'exprime comme un multiple du score.*

Pour clore cette sous-section, on introduit la notion de score efficace. De la définition de la fonction d'influence efficace en tant que gradient riemannien, on sait que cette dernière pointe dans la direction de croissance maximale de Ψ, soit dans une direction perpendiculaire aux surfaces de niveau de Ψ. Plus précisément, dès que $\Psi'(\theta_0) \neq 0$, la surface de niveau $\{\Psi = \psi_0\}$ admet un espace tangent $\Lambda \subset T_{\theta_0}\mathcal{M}$ en θ_0 et on a

$$\varphi_{eff} = \mathrm{proj}_{\Lambda^\perp} \varphi_{eff}, \tag{1.4.7}$$

où $\mathrm{proj}_{\Lambda^\perp}$ est l'opérateur de projection orthogonale sur $\Lambda^\perp$, le complément orthogonal de Λ dans $T_{\theta_0}\mathcal{M}$. Maintenant, pour un modèle paramétrique, dès que $\Psi'(\theta_0) \neq 0$, il est possible de reparamétrer le modèle par $\theta = (\psi,\eta)$ tel que $\Psi(\psi,\eta) = \psi$ **[21]**. Pour une telle paramétrisation, le vecteur score est de la forme

$$S(z;\theta_0) = (S_\psi(z;\theta_0), S_\eta(z;\theta_0)^\top),$$

où

$$S_\psi(z;\theta_0) = \frac{\partial \log p(z;\psi_0,\eta_0)}{\partial\psi}, \quad S_\eta(z;\theta_0)^\top = \frac{\partial \log p(z;\psi_0,\eta_0)}{\partial\eta^\top}$$

et Λ est simplement l'espace engendré par les composantes de $S_\eta(z;\theta_0)$, i.e. $\Lambda = \{c^\top S_\eta(z;\theta_0) | c \in \mathbb{R}^{N-1}\}$. De plus, puisque

$$\frac{\partial\Psi}{\partial\theta^\top} = (1,0,\ldots,0)$$

avec cette paramétrisation, la formule (1.4.7) lorsque combinée à (1.4.6) se réduit à

$$\varphi_{eff} = I(\theta_0)^{11} S_{eff}(z;\theta_0) \tag{1.4.8}$$

où

$$S_{eff}(z;\theta_0) = \mathrm{proj}_{\Lambda^\perp} S_\psi = S_\psi - \langle S_\psi, S_\eta^\top \rangle \langle S_\eta, S_\eta^\top \rangle^{-1} S_\eta$$

est appelé le *score efficace*. En fait, le score efficace caractérise complètement la fonction d'influence efficace car prenant le produit intérieur avec S_ψ de l'équation (1.4.8), on obtient $I(\theta_0)^{11} = \langle S_{eff}, S_{eff} \rangle^{-1}$. En effet, on a $\langle \varphi_{eff}, S_\psi \rangle = 1$ par (1.4.3) et $\langle S_{eff}, S_\psi \rangle = \langle S_{eff}, S_{eff} \rangle + \langle S_{eff}, S_{eff}^\perp \rangle = \langle S_{eff}, S_{eff} \rangle$.

1.4.2. Le cas d'un modèle non-paramétrique

Jusqu'ici nous n'avons considéré que le cas où $\mathcal{M}$ est un modèle paramétrique pour ψ_0. Supposons maintenant au contraire que $\mathcal{M}$ n'est *pas* un modèle paramétrique. Il contient cependant des sous-modèles paramétriques, i.e. des modèles paramétriques $\mathcal{M}'$ pour ψ_0 avec $\mathcal{M}' \subset \mathcal{M}$. Pour ces modèles, l'espace tangent $T_P\mathcal{M}'$ est défini et on définit l'espace tangent $T_P\mathcal{M}$ comme la fermeture dans $\mathcal{H}_P$ de la somme directe sur les espaces tangents de tous les sous-modèles paramétriques passant par P:

$$T_P\mathcal{M} = \overline{\bigoplus_{\substack{\mathcal{M}' \subset \mathcal{M} \\ \dim \mathcal{M}' < \infty}} T_P\mathcal{M}'}.$$

Intuitivement, on pense à $T_P\mathcal{M}$ comme à l'ensemble de toutes les combinaisons linéaires d'éléments des $T_P\mathcal{M}'$ et leurs points limite. On suppose que $T_P\mathcal{M}$ est un espace vectoriel, ce qui n'est pas garanti mais sera le cas dans toutes les applications raisonables (cf. **[35]**). Il est nécessaire de prendre la fermeture dans la définition de $T_P\mathcal{M}$ pour garantir l'existence d'un opérateur de projection orthogonale **[4]**.

Notons ensuite que si ψ_n est un estimateur AL pour $\Psi : \mathcal{M} \to \mathbb{R}$, alors par restriction, ψ_n est aussi un estimateur AL pour $\Psi|_{\mathcal{M}'}$ pour tous les sous-modèles paramétriques $\mathcal{M}'$. En particulier, la fonction d'influence de ψ_n est de variance supérieure ou égale au supremum sur la borne de Cramer-Rao de tous les estimateur AL pour un sous-modèle paramétrique. On appelle ce supremum la *borne d'efficacité*. On peut montrer (**[35]** Theorem 4.2-4.3, **[20]**) que la fonction d'influence φ d'un estimateur ALR (i.e. un estimateur AL dont la restriction à tout sous-modèle paramétrique est ALR) vérifie (1.4.3) pour tout sous-modèle paramétrique.

De plus, il existe un unique élément $\varphi_{eff} \in \mathcal{H}_{P_Z}$ appelé *fonction d'influence efficace* qui vérifie (1.4.3) pour tout sous-modèle paramétrique $\mathcal{M}' \subset \mathcal{M}$ et dont la variance atteint la borne d'efficacité. La fonction d'influence efficace s'obtient d'une fonction d'influence arbitraire par projection orthogonale sur l'espace tangent. En particulier, si une fonction d'influence appartient à l'espace tangent, alors c'est la fonction d'influence efficace.

1.5. L'estimateur AIPTW

Reprenons la situation de l'Exemple 1.3.1, soit un triplet $Z = (Y,A,L) : \Omega \to \mathcal{T}$ satisfaisant les conditions (C1)-(C4) de la section 1.1. On considère le modèle non-paramétrique $\mathcal{M} = \mathcal{P}$ (l'ensemble de toutes les distributions de probabilité sur $\mathcal{T}$). Un choix d'une mesure dominante ν pour un voisinage de P_Z permet d'identifier les éléments de $\mathcal{M}$ à leur fonction de densité $\frac{dP}{d\nu} = p(y,a,\ell)$ et la décomposition

$$p(y,a,\ell) = p_{Y|A,L}(y|a,\ell)p_{A|L}(a|\ell)p_L(\ell) \tag{1.5.1}$$

permet de voir $\mathcal{M}$ comme sous-ensemble de $\mathcal{M}_{Y|A,L} \times \mathcal{M}_{A|L} \times \mathcal{M}_L$, où $\mathcal{M}_{Y|A,L}$ est l'ensemble des fonctions positives de y,a,ℓ, etc.. De même, on peut voir le fibré tangent $T\mathcal{M}$ comme

$$T_P\mathcal{M} = \{U+V+W \in T_{p_{Y|A,L}}\mathcal{M}_{Y|A,L} \oplus T_{p_{A|L}}\mathcal{M}_{A|L} \oplus T_{p_L}\mathcal{M}_L \mid E_P[U+V+W] = 0\}. \tag{1.5.2}$$

En effet, étant donné $p(y,a,\ell;t) = p_{Y|A,L}(y|a,\ell;t)p_{A|L}(a|\ell;t)p_L(\ell;t)$ un sous-modèle paramétrique avec $p(y,a,\ell;0) = p(y,a,\ell)$, on obtient

$$\begin{aligned}\int p'_{Y|A,L}(y|a,\ell;0)p_{A|L}(a|\ell)p_L(\ell) + p_{Y|A,L}(y|a,\ell)p'_{A|L;0}(a|\ell)p_L(\ell)\\ + p_{Y|A,L}(y|a,\ell)p_{A|L}(a|\ell)p'_L(\ell;0)d\nu = 0.\end{aligned} \tag{1.5.3}$$

en derivant la relation

$$\int p_{Y|A,L}(y|a,\ell;t)p_{A|L}(a|\ell;t)p_L(\ell;t)d\nu = 1$$

en $t = 0$. Or,

$$\begin{aligned}U = \frac{d}{dt}\Big|_{t=0} \log p_{Y|A,L}(y|a,\ell;t) = \frac{p'_{Y|A,L}(y|a,\ell)}{p_{Y|A,L}(y|a,\ell)},\\ V = \frac{d}{dt}\Big|_{t=0} \log p_{A|L}(a|\ell;t) = \frac{p'_{A|L}(a|\ell)}{p_{A|L}(a|\ell)},\\ W = \frac{d}{dt}\Big|_{t=0} \log p_L(\ell;t) = \frac{p'_L(\ell)}{p_L(\ell)},\end{aligned}$$

donc (1.5.3) s'écrit également

$$\int p(y,a,\ell)(U+V+W)d\nu = 0.$$

Le paramètre ciblé est l'effet causal moyen $\psi_0 = E[Y^1 - Y^0]$, qui, tel qu'établit à l'Exemple 1.3.1, se calcule à partir de la distribution à l'aide de la fonction

$$\Psi(P) = \int \left(\int y p_{Y|A,L}(y|a=1,\ell)dy - \int y p_{Y|A,L}(y|a=0,\ell)dy \right) p_L(\ell)d\ell. \tag{1.5.4}$$

Dans cette section, on calcule la fonction d'influence efficace pour cette situation et on étudie l'estimateur AIPTW, soit un estimateur AL ayant la fonction d'influence efficace comm fonction d'influence.

Lemme 1.5.1 ([**35**] Theorem 4.4)**.** *Pour le modèle non-paramétrique $\mathcal{M} = \mathcal{P}$ (l'ensemble de toutes les distributions de probabilité sur $\mathcal{T}$), l'espace tangent $T_{P_Z}\mathcal{M}$ est l'espace $\mathcal{H}_{P_Z}$ tout entier.*

Démonstration. Soit $h \in \mathcal{H}_{P_Z}$ une fonction *bornée*, ν une mesure dominant P_Z et $p(z) = \frac{dP_Z}{d\nu}$ la fonction de densité. Pour $|t|$ suffisament petit,

$$p(z;t) = p(z)(1+th(z))$$

est un sous-modèle paramétrique passant par P_Z en $t=0$ et ayant h comme vecteur score. En effet, on a

$$\int p(z;t)dz = \int p(z)dz + t\int h(z)p(z)dz = 1 + tE[h(Z)] = 1$$

et puisque h est bornée, il existe $\epsilon > 0$ tel que $(1+\epsilon h(z)) > 0 \ \forall t \in (-\epsilon,\epsilon)$. Ceci montre que $p(z;t)$ est une fonction de densité pour tout $t \in (-\epsilon,\epsilon)$. De plus,

$$\left.\frac{d\log p(z;t)}{dt}\right|_{t=0} = \frac{p'(z;0)}{p(z;0)} = \frac{p(z)h(z)}{p(z)} = h(z),$$

ce qui montre que $h \in T_{P_Z}\mathcal{M}$. On conclut en invoquant le fait que les fonctions bornées sont denses dans l'espace des fonctions L^2 d'espérance nulle. En effet, ceci signifie que si $h \in \mathcal{H}_{P_Z}$ est une fonction *non* bornée, alors il existe une suite de fonctions bornées $h_k \in T_{P_Z}\mathcal{M}$ telle que $||h_k - h||_{L^2} \to 0$. Comme $T_{P_Z}\mathcal{M}$ est fermé dans $L^2(\mathcal{T},P_Z)$, ceci signifie que $h \in T_{P_Z}\mathcal{M}$. □

Lemme 1.5.2. *La décomposition* (1.5.2) *est orthogonale. En particulier, la fonction d'influence efficace φ_{eff} en P_Z se décompose en trois parties mutuellement orthogonales*

$$\varphi_{eff} = \varphi^Y_{eff} + \varphi^A_{eff} + \varphi^L_{eff}.$$

DÉMONSTRATION. Considérons les sous-modèles paramétriques unidimensionnels

$$p(y,a,\ell;r) = p_{Y|A,L}(y|a,\ell;r)p_{A|L}(a|\ell)p_L(\ell), \tag{1.5.5}$$

$$p(y,a,\ell;s) = p_{Y|A,L}(y|a,\ell)p_{A|L}(a|\ell;s)p_L(\ell), \tag{1.5.6}$$

$$p(y,a,\ell;t) = p_{Y|A,L}(y|a,\ell)p_{A|L}(a|\ell)p_L(\ell;t) \tag{1.5.7}$$

passant par P_Z en $r=0$, $s=0$ et $t=0$ respectivement. Les scores associés sont respectivement

$$S_Y(y|a,\ell) = \left.\frac{d}{dr}\right|_{r=0} \log p(y,a,\ell;r) = \frac{p'_{Y|A,L}(y|a,\ell;0)}{p_{Y|A,L}(y|a,\ell;0)} \in T_Y\mathcal{M}, \tag{1.5.8}$$

$$S_A(a|\ell) = \left.\frac{d}{ds}\right|_{s=0} \log p(y,a,\ell;s) = \frac{p'_{A|L}(a|\ell;0)}{p_{A|L}(a|\ell;0)} \in T_A\mathcal{M}, \tag{1.5.9}$$

$$S_L(\ell) = \left.\frac{d}{dt}\right|_{t=0} \log p(y,a,\ell;t) = \frac{p'_L(\ell;0)}{p_L(\ell;0)} \in T_L\mathcal{M}, \tag{1.5.10}$$

et tout élément de $T_Y\mathcal{M}$ (resp. $T_A\mathcal{M}, T_L\mathcal{M}$) est de cette forme. Utilisant le fait que $E[S_Y|L,A] = E[S_A|L] = E[S_L] = 0$ (propriété des scores), on calcule

$$\langle S_A,S_L\rangle = E[S_A(A|L)S_L(L)] = E[E[S_A(A|L)|L]S_L(L)] = 0,$$

$$\langle S_Y,S_L\rangle = E[S_Y(Y|A,L)S_L(L)] = E[E[S_Y(Y|A,L)|A,L]S_L(L)] = 0,$$

$$\langle S_Y,S_A\rangle = E[S_Y(Y|A,L)S_A(A|L)] = E[E[S_Y(Y|A,L)|A,L]S_A(A|L)] = 0.$$

□

Proposition 1.5.3. *La fonction d'influence efficace au point $P \in \mathcal{M}$ est*

$$\varphi_{eff}(z) = \frac{a(y-Q_1(\ell))}{g(\ell)} + Q_1(\ell) - \frac{(1-a)(y-Q_0(\ell))}{1-g(\ell)} - Q_0(\ell) - \Psi(P), \tag{1.5.11}$$

avec décomposition orthogonale

$$\varphi^Y_{eff} = \frac{a(y-Q_1(\ell))}{g(\ell)} - \frac{(1-a)(y-Q_0(\ell))}{1-g(\ell)},$$

$$\varphi^A_{eff} = 0,$$

$$\varphi^L_{eff} = Q_1(\ell) - Q_0(\ell) - \Psi(P),$$

où $Q_a(\ell) = E_P[Y|A=a,L=\ell]$, $g(\ell) = P[A=1|L=\ell] = E_P[A|L=\ell]$.

DÉMONSTRATION. Montrons que la fonction d'influence efficace en $P = P_Z$ pour

$$\Psi^1(P) = \int \left(\int y p_{Y|A,L}(y|a=1,\ell) dy \right) p_L(\ell) d\ell$$

est

$$\varphi^1_{eff}(z) = \frac{a(y - Q_1(\ell))}{g(\ell)} + Q_1(\ell) - \Psi^1(P_Z).$$

Puisque l'espace tangent est $\mathcal{H}_{P_Z}$, il suffit de montrer que φ^1_{eff} est une fonction d'influence. Clairement, il suffit pour cela de montrer que $E[\varphi^1_{eff}(Z)] = 0$ et que

$$\langle \varphi^1_{eff}, S_\epsilon \rangle = \left. \frac{d\Psi(\epsilon)}{d\epsilon} \right|_{\epsilon=0}$$

pour tout sous-modèle paramétrique *unidimensionnel* $\mathcal{M}' = \{P_\epsilon | \epsilon \in \mathbb{R}\}$ tel que $P_0 = P_Z$ et $S_\epsilon = \frac{d}{d\epsilon}\big|_{\epsilon=0} \log \frac{dP_\epsilon}{d\nu}$ pour ν une mesure dominante. Si

$$\frac{dP_\epsilon}{d\nu} = p(y,a,\ell;\epsilon) = p_{Y|A,L}(y|a,\ell;\epsilon) p_{A|L}(a|\ell;\epsilon) p_L(\ell;\epsilon),$$

alors

$$S_\epsilon(z) = S_Y(y|a,\ell) + S_A(a|\ell) + S_L(\ell),$$

où S_Y, S_A, S_L sont comme dans (1.6.9)-(1.5.10). Posons

$$\varphi^{1,Y}_{eff}(z) = \frac{a(y - Q_1(\ell))}{g(\ell)}, \quad \varphi^{1,A}_{eff}(z) = 0, \quad \varphi^{1,L}_{eff}(z) = Q_1(\ell) - \Psi^1(P). \tag{1.5.12}$$

On calcule

$$E[\varphi^{1,Y}_{eff}(Z)] = E\left[\frac{1}{g(L)} \left(E[AY|L] - Q_1(L) E[A|L] \right) \right] = 0$$

(car $E[AY|L] = Q_1(L) g(L)$ (cf. (1.2.2)) et $E[A|L] = g(L)$ par définition),

$$E[\varphi^{1,L}_{eff}(Z)] = E[Q_1(L)] - \Psi^1(P_Z) = 0$$

(car $E[Q_1(L)] = E[Y^1]$ (cf. (1.1.1)) et $\Psi^1(P_Z) = E[Y^1]$). Ensuite, calculons les produits intérieurs des composantes de φ^1_{eff} avec les composantes de S_ϵ.

$$\begin{aligned} \langle \varphi^{1,Y}_{eff}, S_Y \rangle &= E\left[\frac{AY}{g(L)} S_Y \right] - E\left[\frac{A Q_1(L)}{g(L)} E[S_Y|A,L] \right] \\ &= E\left[\frac{AY}{g(L)} S_Y \right] \end{aligned}$$

(car $E[S_Y|A,L]=0$ (propriété des scores)),

$$\begin{aligned}\langle\varphi^{1,Y}_{eff},S_L\rangle &= E\left[\frac{E[AY|L]}{g(L)}S_L\right]-E\left[\frac{Q_1(L)E[A|L]}{g(L)}S_L\right],\\ &= E\left[Q_1(L)S_L-Q_1(L)S_L\right]\end{aligned}$$

$$\langle\varphi^{1,L}_{eff},S_Y\rangle = E[Q_1(L)E[S_Y|A,L]]+\Psi(P_z)E[S_Y|A,L]]=0,$$

$$\begin{aligned}\langle\varphi^{1,L}_{eff},S_L\rangle &= E[Q_1(L)S_L]+\Psi(P_z)E[S_L]\\ &= E[Q_1(L)S_L].\end{aligned}$$

Ceci démontre que (1.5.12) est bien la décomposition orthogonale de φ^1_{eff} et aussi que

$$\langle\varphi^1_{eff},S_\epsilon\rangle = E\left[\frac{AY}{g(L)}S_Y+Q_1(L)S_L\right].$$

Par ailleurs, par la règle de dérivation d'un produit, on a

$$\begin{aligned}\left.\frac{d\Psi(\epsilon)}{d\epsilon}\right|_{\epsilon=0} &= \int\left(\int yp'_{Y|A,L}(y|a=1,\ell;0)dy\right)p_L(\ell;0)d\ell\\ &\quad+\int\left(\int yp_{Y|A,L}(y|a=1,\ell;0)dy\right)p'_L(\ell;0)d\ell\\ &= \int\left(\int yS_{Y|A,L}(y|a=1,\ell)p_{Y|A,L}(y|a=1,\ell)dy\right)p_L(\ell)d\ell\\ &\quad+\int\left(\int yp_{Y|A,L}(y|a=1,\ell)dy\right)S_L(\ell)p_L(\ell)d\ell\\ &= E[E[YS_Y|A=1,L]]+E[E[Y|A=1,L]S_L(L)]\\ &= E\left[E\left[\frac{AY}{g(L)}S_Y\middle|A=1,L\right]g(L)\right]+E[Q_1(L)S_L(L)]\\ &= E\left[E\left[\frac{AY}{g(L)}S_Y\middle|L\right]\right]+E[Q_1(L)S_L(L)]\\ &= E\left[\frac{AY}{g(L)}S_Y+Q_1(L)S_L(L)\right].\end{aligned}$$

De la même façon, on montre que la fonction d'influence efficace en $P=P_Z$ pour

$$\Psi^0(P)=\int\left(\int yp_{Y|A,L}(y|a=0,\ell)dy\right)p_L(\ell)d\ell$$

est

$$\varphi^0_{eff}(z)=\frac{(1-a)(y-Q_0(\ell))}{1-g(\ell)}+Q_0(\ell)-\Psi^0(P_Z).$$

Comme $\Psi = \Psi^1 - \Psi^0$, la fonction d'influence efficace en $P = P_Z$ pour Ψ est $\varphi^1_{eff} - \varphi^0_{eff} = \varphi_{eff}$. Il est facile de constater que la preuve n'est pas spécifique à la distribution P_Z et donc la fonction d'influence en un point P quelconque de $\mathcal{M}$ est

$$\varphi_{eff}(z) = \frac{a(y - Q_1(\ell))}{g(\ell)} + Q_1(\ell) - \frac{(1-a)(y - Q_0(\ell))}{1 - g(\ell)} - Q_0(\ell) - \Psi(P).$$

□

Clairement, un estimateur AL ayant φ_{eff} comme fonction d'influence est

$$n^{-1} \sum_{i=1}^{n} \frac{A_i(Y_i - Q_1(L_i))}{g(L_i)} + Q_1(L_i) - \frac{(1 - A_i)(Y_i - Q_0(L_i))}{1 - g(L_i)} - Q_0(L_i).$$

On l'appelle *estimateur AIPTW* (pour Augmented IPTW). Plus souvent, on écrira l'estimeur AIPTW sous la forme

$$n^{-1} \sum_{i=1}^{n} H(A_i,L_i)(Y_i - Q_{A_i}(L_i)) + Q_1(L_i) - Q_0(L_i), \tag{1.5.13}$$

où

$$H(a,\ell) = \frac{a}{g(\ell)} - \frac{1-a}{1 - g(\ell)}. \tag{1.5.14}$$

Dans la pratique, les fonctions $g(\ell)$, et $Q_a(\ell)$ ne sont généralement pas connues et il faut les estimer. L'estimateur résultant est *doublement robuste* au sens suivant:

Proposition 1.5.4. *Soit $\hat{g}$, $\hat{Q}_0$, $\hat{Q}_1$ des estimateurs de g, Q_0 et Q_1 convergents respectivement vers $\tilde{g}$, $\tilde{Q}_0$ et $\tilde{Q}_1$ en probabilité. Si $\tilde{g} = g$ ou si $\tilde{Q}_0 = Q_0$, $\tilde{Q}_1 = Q_1$, alors sous des conditions de régularité suffisantes, l'estimateur*

$$n^{-1} \sum_{i=1}^{n} \hat{H}(A_i,L_i)(Y_i - \hat{Q}_{A_i}(L_i)) + \hat{Q}_1(L_i) - \hat{Q}_0(L_i) \tag{1.5.15}$$

est convergent.

Démonstration. Supposons d'abord que $\tilde{g} = g$ est consistant et que $\hat{Q}_a = \tilde{Q}_a + o_p(1)$. On écrit l'estimateur AIPTW sous la forme

$$n^{-1} \sum_{i=1}^{n} \hat{H}(A_i,L_i)Y_i - n^{-1} \sum_{i=1}^{n} \left(\hat{Q}_1(L_i) \left(\frac{A_i}{\hat{g}(L_i)} - 1 \right) - \hat{Q}_0(L_i) \left(\frac{1 - A_i}{1 - \hat{g}(L_i)} - 1 \right) \right),$$

Le premier terme est simplement l'estimateur IPTW pour lequel g est estimé. Par conséquent, ce terme converge en probabilité vers ψ_0 à condition que l'estimeur $\hat{g}$ soit suffisament

régulier. Le second terme converge quant à lui en probabilité vers

$$\begin{aligned}
&E\left[\left(\tilde{Q}_1(L)\left(\frac{A}{g(L)}-1\right)-\tilde{Q}_0(L)\left(\frac{1-A}{1-g(L)}-1\right)\right)\right]\\
&=E\left[E\left[\left(\tilde{Q}_1(L)\left(\frac{A}{g(L)}-1\right)-\tilde{Q}_0(L)\left(\frac{1-A}{1-g(L)}-1\right)\right)\right]\Big|L\right]\\
&=E\left[\left(\tilde{Q}_1(L)\left(\frac{E[A|L]}{g(L)}-1\right)-\tilde{Q}_0(L)\left(\frac{1-E[A|L]}{1-g(L)}-1\right)\right)\right]\\
&=0
\end{aligned}$$

puisque $E[A|L]=g(L)$.

Si $\hat{Q}_0$, $\hat{Q}_1$ sont convergents et $\hat{g}(L)=\tilde{g}(L)+o_p(1)$, on considère la forme

$$n^{-1}\sum_{i=1}^{n}\hat{H}(A_i,L_i)(Y_i-\hat{Q}_{A_i}(L_i))+n^{-1}\sum_{i=1}^{n}\hat{Q}_1(L_i)-\hat{Q}_0(L_i)$$

de l'estimateur AIPTW. Le second terme converge en probabilité vers $E[Q_1(L)-Q_0(L)]=\psi_0$ tandis que le premier terme converge en probabilité vers

$$\begin{aligned}
E[\tilde{H}(A,L)(Y-Q_A(L))]&=E[E[\tilde{H}(A,L)(Y-Q_A(L))]|A,L]\\
&=E[\tilde{H}(A,L)(E[Y|A,L]-Q_A(L))],\\
&=0,
\end{aligned}$$

où $\tilde{H}$ est comme dans (1.5.14) mais avec g remplacé par $\tilde{g}$. □

Remarque 1.5.5. *En utilisant la théorie des processus empiriques, on peut dégager des conditions sur $\hat{g}$ et $\hat{Q}$ garantissant que l'estimateur* (1.5.15) *est AL avec fonction d'influence efficace (cf.* **[20]** *section 4).*

1.6. L'estimateur TMLE

À la sous-section 1.6.1, on présente le TMLE (pour targeted maximum likelihood estimator) **[43]** pour le cas particulier de l'Exemple 1.3.1 où le paramètre ciblé est l'effet causal moyen. On verra que dans ce contexte, l'estimateur TMLE se présente comme un estimateur de la forme (1.5.15), jouissant ainsi de la propriété de double robustesse (Proposition 1.5.4). À la sous-section 1.6.2, on présente la théorie générale du TMLE et on voit comment la procédure présentée à la sous-section 1.6.1 en est un exemple.

1.6.1. Estimateur TMLE pour l'effet causal moyen

Le TMLE est un processus qui, étant donné une estimation $\hat{P}$ de P_Z (ou plus généralement une estimation de la *partie* de P_Z qui figure dans le calcul de $\Psi(P_Z)$), résulte en une version corrigée $\hat{P}^*$ de l'estimation. L'idée est qu'en général, la technique d'estimation $\hat{P}$ optimise le compromis biais-variance *pour l'estimation de* P_Z*, et non de* $\Psi(P_Z)$. Le TMLE se veut une correction de $\hat{P}$ en vue de réduire le biais dans l'estimation résultante $\Psi(\hat{P}^*)$ du paramètre d'intérêt $\Psi(P_Z)$.

Dans la situation de l'Exemple 1.3.1, on suppose que $\hat{g}$, $\hat{Q}_a$ sont des estimations des fonctions $g(\ell) = P[A = 1|L = \ell]$, $Q_a(\ell) = E[Y|A = a,L = \ell]$ $(a = 0,1)$ calculées à partir d'un échantillon iid $Z_i = (Y_i,A_i,L_i)$ $(i = 1,\ldots,n)$ de Z. La méthode d'estimation utilisée est *a priori* sans importance et nous ne supposons pas la convergence. La distribution p_L est quant à elle estimée par la distribution empirique, soit la distribution discrète

$$\hat{p}_L(\ell) = n^{-1}\sum_{i=1}^{n}\mathbb{I}(L_i = \ell). \tag{1.6.1}$$

Notons qu'en terme des fonctions Q_a, le paramètre ciblé s'écrit

$$\psi_0 = E[Q_1(L) - Q_0(L)].$$

Par conséquent, nos estimations donnent lieu à l'estimation de ψ_0

$$\hat{\psi}_0 = n^{-1}\sum_{i=1}^{n}\hat{Q}_1(L_i) - \hat{Q}_0(L_i). \tag{1.6.2}$$

La correction TMLE consiste à modifier $\hat{Q}_a$ dans (1.6.2) en ajustant un modèle de régression logistique pour $Q_A(L)$ avec logit$\hat{Q}_a$ comme offset et l'inverse du score de propension comme variable indépendante. Plus précisément, si Y est bornée, alors en y appliquant une transformation affine appropriée, on peut supposer sans perte de généralité que $0 < Y_i,\hat{Q}_{A_i}(L_i) < 1$ $\forall i$. On ajuste le modèle de régression logistique

$$\text{logit}Q_A(L) = \text{logit}\hat{Q}_A(L) + \epsilon\hat{H}(A,L)$$

pour ϵ, où $\hat{H}(A,L)$ est obtenu de (1.5.14) en remplaçant g par $\hat{g}$. Par la procédure d'estimation des paramètres dans un modèle de régression logistique, ceci revient à résoudre

l'équation estimante

$$0 = \sum_{i=1}^{n} \hat{H}(A_i,Y_i) \left(Y_i - \text{expit} \left(\text{logit}\hat{Q}_{A_i}(L_i) + \epsilon \hat{H}(A_i,L_i) \right) \right) \tag{1.6.3}$$

pour ϵ, où expit est la fonction inverse de logit,

$$\text{expit}(x) = \frac{e^x}{1+e^x}. \tag{1.6.4}$$

De nouvelles estimations $\hat{Q}_a^*(L_i)$ de $Q_a(L_i)$ sont ensuite calculés à partir de la formule

$$\hat{Q}_a^*(L_i) = \text{expit} \left(\text{logit}\hat{Q}_a(L_i) + \hat{\epsilon}\hat{H}(a,L_i) \right), \quad a = 0,1.$$

On obtient au final une nouvelle estimation

$$\hat{\psi}_0^* = n^{-1} \sum_{i=1}^{n} \hat{Q}_1^*(L_i) - \hat{Q}_0^*(L_i) \tag{1.6.5}$$

de ψ_0. Et en fait, puisque les $\hat{Q}_{A_i}^*(L_i)$ vérifient

$$0 = \sum_{i=1}^{n} \hat{H}(A_i,Y_i) \left(Y_i - \hat{Q}_a^*(L_i) \right)$$

on voit que cet estimateur coïncide avec l'estimateur (1.5.13) obtenu en substituant $g(L_i)$, $Q_a(L_i)$ par $\hat{g}(L_i)$ et $\hat{Q}_a^*(L_i)$ respectivement. On peut montrer **[43]** que l'estimateur $\hat{Q}_a^*(\ell)$ de $Q_a(\ell)$ est convergent dès que $\hat{Q}_a(\ell)$ est convergent. Par conséquent, par la Proposition 1.5.4, le TMLE (1.6.5) est doublement robuste au sens où $\hat{\psi}_0^*$ est une estimateur convergent de ψ_0 dès que l'un des deux estimations initiales $\hat{Q}_a(\ell)$ ou $\hat{g}(\ell)$ est convergent.

1.6.2. Le TMLE en général

Soit $\mathcal{M}$ un modèle pour la distribution P_Z d'une variable aléatoire Z et $\Psi : \mathcal{M} \to \mathbb{R}$ dont on cherche à estimer la valeur en P_Z. Notons tout d'abord que P_Z *est la distribution qui maximise la fonction* $P \mapsto E_{P_Z}\left[\log \frac{dP}{d\nu}(Z)\right]$, où ν est une mesure dominante. En effet, pour un sous-modèle paramétrique P_θ avec $P_{\theta_0} = P_Z$ et $\hat{\theta}^{MLE}$ l'estimateur du maximum de vraisemblance, on a $\hat{\theta}^{MLE} \to \theta_0$ et $n^{-1}\sum_{i=1}^{n} \log \frac{dP_\theta}{d\nu}(Z_i) \to E_{P_Z}\left[\log \frac{dP_\theta}{d\nu}(Z)\right]$, donc sous certaines conditions de régularité, on a $n^{-1}\sum_{i=1}^{n} \log \frac{dP_{\hat{\theta}^{MLE}}}{d\nu}(Z_i) \to E_{P_Z}\left[\log \frac{dP_{\theta_0}}{d\nu}(Z)\right]$. Comme $\hat{\theta}^{MLE}$ est la valeur du paramètre qui maximise $n^{-1}\sum_{i=1}^{n} \log \frac{dP_{\hat{\theta}^{MLE}}}{d\nu}(Z_i)$, on obtient la conclusion.

L'idée de base derrière l'estimateur TMLE de M. van der Laan et D. Rubin **[45]** est que si $\hat{P}$ est une estimation de P_Z relativement près de P_Z, on s'attend à pouvoir améliorer

l'estimation en considérant un sous-modèle paramétrique unidimensionnel P_ϵ avec $P_0 = \hat{P}$ et en prenant $\hat{P}'$ l'élément du sous-modèle qui maximise

$$n^{-1} \sum_{i=1}^{n} \log \frac{dP_\epsilon}{d\nu}(Z_i). \tag{1.6.6}$$

Mais si $\Psi(\hat{P}) = \Psi(\hat{P}')$, alors cette amélioration de l'estimation de P_Z n'aura servi à rien en ce qui concerne l'estimation de $\Psi(P_Z)$. Donc, on exige du modèle paramétrique P_ϵ qu'il se dirige dans la direction de variation maximale de Ψ en $\hat{P}$, i.e. dans la direction de la fonction d'influence efficace:

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log \frac{dP_\epsilon}{d\nu} = \text{const.} \times \varphi_{eff}(\hat{P}). \tag{1.6.7}$$

On peut ensuite répéter ce processus en considérant un sous-modèle paramétrique unidimensionnel à travers $\hat{P}'$ pointant dans la direction de $\varphi_{eff}(\hat{P}')$ et en définissant $\hat{P}''$ la distribution de ce sous-modèle maximisant la moyenne empirique de la log-vraisemblance. On répète le processus jusqu'à ce qu'il y ait stabilisation de la distribution à $\hat{P}^*$. L'*estimateur TMLE* de $\Psi(P_Z)$ est alors défini comme $\Psi(\hat{P}^*)$. Puisque la distribution $\hat{P}^*$ maximise (1.6.6) pour P_ϵ pointant dans la direction de $\varphi_{eff}(\hat{P}^*)$, on a par (1.6.7),

$$0 = \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} n^{-1} \sum_{i=1}^{n} \log \frac{dP_\epsilon}{d\nu}(Z_i) = const. \times n^{-1} \sum_{i=1}^{n} \varphi_{eff}(\hat{P}^*)(Z_i). \tag{1.6.8}$$

Le fait que $\hat{P}^*$ vérifie

$$\sum_{i=1}^{n} \varphi_{eff}(\hat{P}^*)(Z_i) = 0$$

sert de point de départ pour montrer que, sous certaines conditions de régularité et de convergence sur $\hat{P}$ ([**43**] Theorem A.5), le TMLE est un estimateur AL (donc convergent) pour ψ_0 avec fonction d'influence la fonction d'influence efficace (donc le TMLE est asymptotiquement efficace).

Bien qu'il existe habituellement une infinité de choix possibles pour le sous-modèle paramétrique unidimensionnel P_ϵ à travers un point donné, il existe un choix canonique et optimal, soit la *courbe intégrale maximale* du champ vectoriel formé des fonctions d'influence efficaces sur $\mathcal{M}$. La courbe intégrale maximale à travers $P \in \mathcal{M}$ est le plus grand sous-modèle paramétrique unidimensionnel P_ϵ avec $P_0 = P$ et $\frac{d}{d\epsilon} \log \frac{dP_\epsilon}{d\nu} = \varphi_{eff}(P_\epsilon)$ *pour tout* ϵ [**21**]. Autrement dit, la courbe intégrale maximale pointe *en tout point* dans la direction de variation maximale de Ψ. De plus, avec ce choix, le processus TMLE se stabilise après

une seule itération. En effet, si $\hat{P}$ est l'estimation initiale de P_Z et si on choisit la courbe intégrale maximale P_ϵ comme sous-modèle paramétrique, alors la seconde estimation $\hat{P}'$ est le point le long de P_ϵ qui maximise (1.6.6). À la seconde itération du processus, on choisit encore la courbe intégrale maximale à travers $\hat{P}'$. Mais cette courbe n'est autre que P_ϵ (à une reparamétrisation près, de sorte que la courbe passe par $\hat{P}'$ en $\epsilon = 0$), donc la distribution $\hat{P}''$ qui maximise (1.6.6) est $\hat{P}'$. On a donc $\hat{P}^* = \hat{P}'$.

Remarque 1.6.1. *(1) Maximiser la moyenne empirique* (1.6.6) *est équivalent à maximiser la* vraisemblance $\prod_{i=1}^n \frac{dP_\epsilon}{d\nu}(Z_i)$*, ce qui explique le nom Targeted Maximum Likelihood Estimator pour l'estimateur.*

(2) Plus généralement, soit $L : \mathcal{M} \to \mathbb{R}^{\mathcal{T}}$ une fonction (appelée fonction de perte*) telle que P_Z minimise $P \mapsto E_{P_Z}[L(P)(Z)]$ et dont on se sert pour identifier $T_{\theta_0}\mathcal{M}$ à $\mathcal{H}_{\theta_0}$ via*

$$\left.\frac{dL(P_\theta)}{d\theta}\right|_{\theta=\theta_0}$$

*(comparer avec 1.4.1 où $L(P) = -\log \frac{dP}{d\nu}$). Étant donné une estimation initiale $\hat{P} = \hat{P}^{(0)}$ de P_Z, on note $\hat{P}^{(j)}$ la distribution le long d'un modèle paramétrique unidimensionnel dans la direction de $\varphi_{eff}(\hat{P}^{(j-1)})$ qui minimise $n^{-1}\sum_{i=1}^n L(P)(Z_i)$. Si $\hat{P}^{(k)} = \hat{P}^{(k+1)}$ pour un certain k, alors on défini $\hat{P}^* = \hat{P}^{(k)}$ et l'*estimateur TMLE *de $\Psi(P_Z)$ est $\Psi(\hat{P}^*)$.*

1.6.3. Le TMLE pour l'effet causal moyen (bis)

Dans cette section, on récupère les résultats de la section 1.6.1 en appliquant les principes de l'estimation TMLE exposés à la section 1.6.2.

On se place dans la situation de l'Exemple 1.3.1 et *on considère le cas où Y est une variable de Bernoulli.* Alors, la distribution conditionnelle de Y sachant A,L s'écrit

$$p_{Y|A,L}(y|a,\ell) = Q_a(\ell)^y(1 - Q_a(\ell))^{1-y},$$

où $Q_a(\ell) = E[Y|A = 1,L = \ell]$. De même, la distribution conditionnelle de A sachant L s'écrit

$$p_{A|L}(a|\ell) = g(\ell)^a(1 - g(\ell))^{1-a}$$

où $g(\ell) = E[A|L = \ell]$. En particulier, par (1.5.1), une estimation initiale $\hat{P}$ de P_Z est équivalent à une estimation $\hat{Q}_a(\ell)$ de la fonction $Q_a(\ell)$, une estimation $\hat{g}(\ell)$ de la fonction

$g(\ell)$ et une estimation de $p_L(\ell)$. Pour cette dernière, on considère la distribution empirique (1.6.1) de L. Par la Proposition 1.5.3, un modèle paramétrique à travers $\hat{P}$ dans la direction de la fonction d'influence efficace est un modèle $p(z;\epsilon) = p_{Y|A,L}(y|a,\ell;\epsilon)p_{A|L}(a|\ell;\epsilon)p_L(\ell;\epsilon)$ tel que

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log p_{Y|A,L}(y|a,\ell;\epsilon) = \hat{H}(a,\ell)(y - \hat{Q}_a(\ell)), \tag{1.6.9}$$

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log p_{A|L}(a|\ell;\epsilon) = 0,$$

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log p_L(\ell;\epsilon) = \hat{Q}_1(\ell) - \hat{Q}_0(\ell),$$

où

$$\hat{H}(a,\ell) = \frac{a}{\hat{g}(\ell)} - \frac{1-a}{1-\hat{g}(\ell)}.$$

Or, les auteurs de **[43]** (Chapter 5) argumentent que rien ne sert de faire varier l'estimation de $p_L(\ell)$ parce que la distribution empirique est une forme non-paramétrique de la méthode du maximum de vraisemblance et donc le processus de mise à jour du TMLE n'améliorerait pas cette estimation. De même, Ψ ne dépend pas de $p_{A|L}$ donc inutile de varier cette partie de $\hat{P}$. On prend donc pour P_ϵ la courbe intégrale maximale passant par $\hat{P}$ du champ vectoriel $H(a,\ell)(y - Q_a(\ell))$, soit

$$p_L(\ell;\epsilon) = n^{-1}\sum_{i=1}^{n} \delta_{L_i}(\ell),$$

$$p_{A|L}(a|\ell;\epsilon) = \hat{g}(\ell)^a(1-\hat{g}(\ell))^{1-a},$$

$$p_{Y|A,L}(y|a,\ell;\epsilon) = \hat{Q}_a(\ell;\epsilon)^y(1-\hat{Q}_a(\ell;\epsilon))^{1-y}$$

où $\hat{Q}_a(\ell;\epsilon)$ est tel que $\hat{Q}_a(\ell;0) = \hat{Q}_a(\ell)$ et

$$\frac{d}{d\epsilon}\log p_{Y|A,L}(y|a,\ell;\epsilon) = \hat{H}(a,\ell)(y - \hat{Q}_a(\ell;\epsilon)). \tag{1.6.10}$$

On calcule

$$\begin{aligned}\frac{d}{d\epsilon}\log p_{Y|A,L}(y|a,\ell;\epsilon) &= \frac{y\frac{d}{d\epsilon}\hat{Q}_a(\ell;\epsilon)}{\hat{Q}_a(\ell;\epsilon)} - \frac{(1-y)\frac{d}{d\epsilon}\hat{Q}_a(\ell;\epsilon)}{1-\hat{Q}_a(\ell;\epsilon)} \\ &= \frac{(y-\hat{Q}_a(\ell;\epsilon))\frac{d}{d\epsilon}\hat{Q}_a(\ell;\epsilon)}{\hat{Q}_a(\ell;\epsilon)(1-\hat{Q}_a(\ell;\epsilon))},\end{aligned}$$

donc on cherche $\hat{Q}_a(\ell;\epsilon)$ tel que $\hat{Q}_a(\ell;0) = \hat{Q}_a(\ell)$ et

$$\frac{d}{d\epsilon}\hat{Q}_a(\ell;\epsilon) = \hat{Q}_a(\ell;\epsilon)(1-\hat{Q}_a(\ell;\epsilon))\hat{H}(a,\ell).$$

Or, la solution de l'équation différentielle $y' = y(1-y)$ est la fonction expit (cf. (1.6.4)), donc $\hat{Q}_a(\ell;\epsilon) = \text{expit}(f(\epsilon))$ où $f(\epsilon)$ est une fonction obéissant aux conditions

$$\frac{df}{d\epsilon} = \hat{H}(a,\ell),$$

$$f(0) = \text{logit}(\hat{Q}_a(\ell)).$$

La solution à ce problème est

$$f(\epsilon) = \text{logit}(\hat{Q}_a(\ell)) + \epsilon\hat{H}(a,\ell),$$

donc

$$\hat{Q}_a(\ell;\epsilon) = \text{expit}(\text{logit}(\hat{Q}_a(\ell)) + \epsilon\hat{H}(a,\ell)).$$

Par (1.6.10), la valeur de ϵ qui maximise la log-vraisemblance $\sum_{i=1}^n \log p_{Y|A,L}(Y_i|A_i,L_i;\epsilon)$ est la solution de l'équation estimante

$$0 = \sum_{i=1}^n \hat{H}(A_i,L_i)(Y_i - \hat{Q}_{A_i}(L_i;\epsilon)). \tag{1.6.11}$$

Or, par la théorie générale des modèles linéaires généralisés (GLM), cette équation est la même utilisée pour l'ajustement du modèle de régression logistique

$$\text{logit}(E[Y|A,L]) = \text{logit}(\hat{Q}_A(L)) + \epsilon\hat{H}(A,L). \tag{1.6.12}$$

Par conséquent, la première amélioration $\hat{P}'$ de l'estimation de P_Z est représentée par la fonction de densité $p'(y,a,\ell) = \hat{p}'_{Y|A,L}(y|a,\ell)\hat{p}_{A|L}(a|\ell)\hat{p}_L(\ell)$ où

$$\hat{p}_L(\ell) = n^{-1}\sum_{i=1}^n \delta_{L_i}(\ell),$$

$$\hat{p}_{A|L}(a|\ell) = \hat{g}(\ell)^a(1-\hat{g}(\ell))^{1-a},$$

$$\hat{p}'_{Y|A,L}(y|a,\ell) = \hat{Q}'_a(\ell)^y(1-\hat{Q}'_a(\ell))^{1-y},$$

et

$$\hat{Q}'_a(\ell) = \text{expit}\left(\text{logit}\hat{Q}_a(\ell) + \epsilon'\hat{H}(a,\ell)\right),$$

pour ϵ' la solution de (1.6.11). Par l'argument du paragraphe précédant la Remarque 1.6.1, le processus TMLE se termine ici. On peut aussi le constater explicitement en notant que la seconde itération du processus consiste à substituer dans

$$\begin{aligned} Q'_a(\ell;\epsilon) &= \text{expit}(\text{logit}(\hat{Q}'_a(\ell) + \epsilon\hat{H}(a,\ell)) \\ &= \text{expit}(\text{logit}(\hat{Q}_a(\ell) + (\epsilon' + \epsilon)\hat{H}(a,\ell)) \end{aligned}$$

la solution ϵ'' de l'équation

$$0 = \sum_{i=1}^{n} \hat{H}(A_i,Y_i)(Y_i - \hat{Q}'_{A_i}(L_i;\epsilon)).$$

Mais, par définition de ϵ', on sait que la solution de cette équation vérifie $\epsilon' + \epsilon'' = \epsilon'$, d'où $\epsilon'' = 0$ et $\hat{Q}''_a(\ell) = \hat{Q}'_a(\ell)$. Ainsi,

$$\hat{Q}^*_a(\ell) = \hat{Q}'_a(\ell)$$

et l'estimateur TMLE pour l'effet causal moyen est

$$\Psi(\hat{P}^*) = n^{-1} \sum_{i=1}^{n} \hat{Q}^*_1(L_i) - \hat{Q}^*_0(L_i). \tag{1.6.13}$$

Si maintenant Y est une *variable réelle quelconque*, on peut montrer que la même procédure de mise à jour peut être utilisée, à condition de prendre soin au préalable d'appliquer l'ensemble des points $\{Y_i,\hat{Q}_{A_i}(L_i)\}$ dans l'intervalle (0,1) à l'aide d'une transformation *affine* T **[13]**. Autrement dit, on remplace le modèle à ajuster (1.6.12) par

$$\text{logit}(E[T(Y)|A,L]) = \text{logit}(T(\hat{Q}_A(L)) + \epsilon\hat{H}(A,L).$$

Comme $E[T(Y)|A,L] = T(E[Y|A,L])$ (pour une transformation affine!), ce modèle est équivalent à

$$T(Q_A(L)) = \text{expit}(T(\hat{Q}_A(L)) + \epsilon\hat{H}(A,L)). \tag{1.6.14}$$

On pose donc

$$\hat{Q}^*_a(\ell) = T^{-1}(\text{expit}(T(\hat{Q}_a(\ell)) + \epsilon'\hat{H}(a,\ell))),$$

où ϵ' est la valeur de ϵ obtenue par ajustement du GLM (1.6.14), et l'estimateur TMLE de l'effet causal moyen est à nouveau donné par (1.6.13).

Chapitre 2

Estimation de la variance

Dans ce chapitre, nous présentons les différentes techniques d'estimation alternatives de la variance du TMLE que nous comparons à l'étalon-or au chapitre suivant.

2.0.1. Estimation de la variance du TMLE: l'étalon-or

L'étalon-or [**43**] en ce qui concerne l'estimation de la variance de l'estimateur (1.6.13) est un exemple de l'estimateur sandwich (1.3.5). Cependant, il peut être jugé insatisfaisant dans certaines conditions extrêmes notament lorsque la taille échantillonnale n est petite, lorsque les scores de propensions sont très petits (ou très grand) ou lorsque $\hat{g}$ n'est pas convergent.

Selon [**43**] (p.96), si certaines conditions de régularité sont remplies et si $\hat{g}$ est un estimateur convergent de g, alors l'estimateur TMLE est AL avec fonction d'influence

$$\varphi = \varphi_{eff}(\hat{P}^*_\infty) - \text{proj}_{T\mathcal{M}_{A|L}}\varphi_{eff}(\hat{P}^*_\infty),$$

où $\hat{P}^*_\infty = \lim_{n\to\infty}\hat{P}^*$ et où $\text{proj}_{T\mathcal{M}_{A|L}}$ est l'opérateur de projection orthogonale sur l'espace tangent au modèle $\mathcal{M}_{A|L}$ utilisé pour l'estimation $\hat{g}$. En particulier, $Var[\varphi(Z)] \leq Var[\varphi_{eff}(\hat{P}^*_\infty)(Z)]$, donc $Var[\varphi_{eff}(\hat{P}^*_\infty)(Z)]$ est une approximation *conservatrice* de $Var[\varphi(Z)]$. Si on estime $\varphi_{eff}(\hat{P}^*_\infty)$ par

$$\varphi_{eff}(\hat{P}^*) = \hat{H}(a,\ell)(y - \hat{Q}^*_a(\ell)) + \hat{Q}^*_1(\ell) - \hat{Q}^*_0(\ell) - \Psi(\hat{P}^*),$$

alors

$$n^{-1}\sum_{i=1}^{n}\varphi_{eff}(\hat{P}^*)(Y_i,A_i,L_i)^2$$

est une estimation asymptotiquement conservatrice de la variance de la fonction d'influence de l'estimateur TMLE. En particulier, une estimation asymptotiquement conservatrice de la

variance asymptotique de $\Psi(\hat{P}^*)$ est

$$n^{-2}\sum_{i=1}^{n}\varphi_{eff}(\hat{P}^*)(Y_i,A_i,L_i)^2. \tag{2.0.1}$$

Alternativement, on peut voir cette estimation comme un exemple de l'estimateur sandwich (cf. Exemple 1.3.3). En effet, si $\hat{g}$ est convergent, alors la variable aléatoire

$$m^*_{AIPTW}(Z;\psi) = \hat{H}(A,L)(Y-\hat{Q}^*_A(L)) + \hat{Q}^*_1(Y) - \hat{Q}^*_0(Y) - \psi, \tag{2.0.2}$$

où $\psi = E[Y^1 - Y^0]$ vérifie $E_\psi[m^*_{AIPTW}(Z;\psi)] = 0$ (voir la preuve de la Proposition 1.5.4). Le m-estimateur obtenu par résolution de l'équation estimante

$$0 = \sum_{i=1}^{n} m^*_{AIPTW}(Z_i;\psi)$$

est simplement le TMLE $\Psi(\hat{P}^*)$ et comme

$$\frac{dm^*_{AIPTW}(z;\psi)}{d\psi} = -1,$$

l'estimateur sandwich (1.3.5) prend la forme

$$\hat{V}_0 = n^{-2}\sum_{i=1}^{n} m^*_{AIPTW}(Z_i;\Psi(\hat{P}^*))^2, \tag{2.0.3}$$

ce qui est précisément l'étalon-or (2.0.1).

2.1. Deux estimateurs lorsque g est estimé paramétriquement

2.1.1. L'estimateur sandwich

L'étalon-or est une estimation de la variance de l'estimateur TMLE qui ne prend pas en compte la partie de la variance dûe au fait que les paramètres g et Q_a sont eux-même estimés à partir des données. Autrement dit, l'étalon-or traite $\hat{g}$ et $\hat{Q}_a$ comme fixes. Si $\hat{g}$ provient d'un modèle paramétrique $g(\ell;\alpha)$ tel que $g(\ell;\alpha_0) = g(\ell)$ et $\alpha \in \mathbb{R}^N$ est estimé par le m-estimateur $\hat{\alpha}_n$ résolvant l'équation estimante

$$0 = \sum_{i=1}^{n} m_g(Z_i;\alpha),$$

alors on peut procéder de façon analogue à l'Exemple 1.3.6 pour obtenir un estimateur sandwich prenant en compte la variabilité de $\hat{g}(\ell) = g(\ell;\hat{\alpha}_n)$. Spécifiquement, si on considère

la situation où l'issue Y est une variable dichotomique, alors le couple $(\Psi(\hat{P}^*),\hat{\alpha}_n^\top)$ est le m-estimateur résolvant l'équation estimante

$$0 = n^{-1}\sum_{i=1}^{n} m(Z_i;\theta)$$

où $\theta = (\psi,\alpha)$ et où

$$m(Z;\theta) = \{m^*_{AIPTW}(Z;\psi,\alpha), m_g(Z;\alpha)^\top\}^\top, \tag{2.1.1}$$

$$m_{AIPTW}(Z;\psi,\alpha) = H(A,L;\alpha)(Y - \hat{Q}^*_A(L;\alpha)) + \hat{Q}^*_1(L;\alpha) - \hat{Q}^*_0(L;\alpha) - \psi,$$

avec

$$H(A,L;\alpha) = \frac{A}{g(L;\alpha)} - \frac{1-A}{1-g(L;\alpha)},$$
$$\hat{Q}^*_A(L;\alpha) = \text{expit}(\text{logit}\hat{Q}_A(L) + \epsilon(\alpha)H(A,L;\alpha)) \tag{2.1.2}$$

et $\epsilon(\alpha)$ la solution de l'équation

$$0 = \sum_{i=1}^{n} H(A_i,Y_i,\alpha)\left(Y_i - \text{expit}\left(\text{logit}\hat{Q}_{A_i}(L_i) + \epsilon H(A_i,L_i,\alpha)\right)\right). \tag{2.1.3}$$

Si Y est une variable aléatoire réelle bornée quelconque, alors on remplace (2.1.2) par

$$\hat{Q}^*_A(L;\alpha) = T^{-1}\left(\text{expit}(\text{logit}T(\hat{Q}_A(L)) + \epsilon(\alpha)H(A,L;\alpha))\right),$$

où T est une fonction affine telle que $T(Y) \in (0,1)$. La fonction d'influence est (cf Exemple 1.3.4)

$$\varphi(Z) = -E\left[\frac{\partial m(Z;\theta_0)}{\partial\theta^\top}\right]^{-1} m(Z;\theta_0),$$

où $\theta_0 = (\psi_0,\alpha_0)$ pour α_0 tel que $g(\ell,\alpha_0) = g(\ell)$. On a

$$\frac{\partial m}{\partial\theta^\top} = \begin{pmatrix} -1 & \frac{\partial m^*_{AIPTW}}{\partial\alpha^\top} \\ 0 & \frac{\partial m_g}{\partial\alpha^\top} \end{pmatrix}.$$

On utilise les identités

$$\frac{d}{dx}\text{expit}(x) = \text{expit}(x)(1-\text{expit}(x)), \qquad \frac{d}{dx}\text{logit}(x) = \frac{1}{x(1-x)} \tag{2.1.4}$$

pour calculer

$$\frac{\partial m^*_{AIPTW}}{\partial\alpha^\top} = \frac{\partial H}{\partial\alpha^\top}(Y - \hat{Q}^*_A) + (v^*_1 - v^*_0 - Hv^*_A)\left(\frac{\partial\epsilon}{\partial\alpha^\top}H + \epsilon\frac{\partial H}{\partial\alpha^\top}\right),$$

$$\frac{\partial H}{\partial\alpha^\top} = -\left(\frac{A}{g^2} + \frac{1-A}{(1-g)^2}\right)\frac{\partial g}{\partial\alpha^\top},$$

avec la notation

$$v_a^* = \hat{Q}_a^*(1 - \hat{Q}_a^*).$$

Le terme $\partial\epsilon/\partial\alpha^\top$ se calcule par dérivation implicite de l'équation (2.1.3). On obtient

$$\frac{\partial\epsilon}{\partial\alpha^\top} = \left(\sum_{i=1}^n H_i^2 v_{A_i}^*\right)^{-1} \sum_{i=1}^n \left(Y_i - Q_{A_i}^* - v_{A_i}^*\epsilon(\alpha)H_i\right) \frac{\partial H_i}{\partial\alpha^\top}, \tag{2.1.5}$$

où

$$H_i = H(A_i,L_i;\alpha).$$

Ceci permet de calculer l'estimateur sandwich (1.3.5) de la matrice de variance-covariance asymptotique de $\hat{\theta} = (\Psi(\hat{P}^*), \hat{\alpha}_n)$. L'estimateur de la variance asymptotique de $\Psi(\hat{P}^*)$ est simplement l'élément (1,1) de cette matrice. On pose

$$\hat{V}_1 = (\hat{V})_{11}. \tag{2.1.6}$$

2.1.2. L'estimateur sandwich avec correction de Fay-Graubard

Bien que l'estimateur sandwich $\hat{V}_1$ considéré à la section 2.1.1 soit une amélioration sur l'étalon-or (2.0.1) (si g est estimé paramétriquement) dans la mesure où l'estimation $\hat{V}_1$ prend en compte la variabilité introduite par le fait que g est estimé, cet estimateur à tendance à sous-estimer la vraie valeur de la variance lorsque la taille échantillonnale n est petite. Dans l'optique de ce problème M. P. Fay et B. I. Graubard [**10**] adaptent une technique provenant de la théorie de l'échantillonage (cf. [**36**] p.141) et proposent une correction ayant pour but d'ajuster à la hausse la valeur de l'estimateur sandwich (1.3.5). Le contexte est le suivant. Soit $\hat{\theta}$ le m-estimateur obtenu par résolution de l'équation estimante

$$0 = \sum_{i=1}^n m(Z_i,\theta).$$

L'estimateur sandwich est de la forme

$$\hat{V} = n^{-1}\hat{\Gamma}^{-1}\hat{\Delta}(\hat{\Gamma}^{-1})^\top, \tag{2.1.7}$$

où

$$\hat{\Gamma} = n^{-1}\sum_{i=1}^n \left(\frac{\partial m(Z_i;\hat{\theta})}{\partial\theta^\top}\right), \quad \hat{\Delta} = n^{-1}\sum_{i=1}^n m(Z_i;\hat{\theta})^{\otimes 2}.$$

Lorsque la taille échantillonnale n est petite, Fay et Graubard [**10**] proposent de remplacer le terme $\hat{\Delta}$ dans (2.1.7) par

$$\hat{\Delta}_{FG} = n^{-1} \sum_{i=1}^{n} (F_i m(Z_i; \hat{\theta}))^{\otimes 2},$$

où F_i est une matrice diagonale avec

$$(F_i)_{jj} = \left[1 - \min \left\{ b, \left(\frac{\partial m(Z_i; \hat{\theta})}{\partial \theta^\top} \hat{\Gamma}^{-1} \right)_{jj} \right\} \right]^{-1/2},$$

et $b < 1$ est une constante choisie par l'utilisateur ayant pour but d'empêcher un ajustement extrême lorsque $\left(\frac{\partial m(Z_i; \hat{\theta})}{\partial \theta^\top} \hat{\Gamma}^{-1} \right)_{jj}$ est très près de 1. Ainsi, l'*estimateur sandwich avec correction de Fay-Graubard* pour la variance asymptotique de $\hat{\theta}$ est

$$\hat{V}_{FG} = n^{-1} \hat{\Gamma}^{-1} \hat{\Delta}_{FG} (\hat{\Gamma}^{-1})^\top.$$

Lorsque la fonction m utilisée est (2.1.1), on obtient un estimateur de la variance asymptotique de $\Psi(\hat{P}^*)$ en prenant l'élément (1,1) de la matrice $\hat{V}_{FG}$. On pose

$$\hat{V}_2 = (\hat{V}_{FG})_{11}. \tag{2.1.8}$$

2.1.3. L'estimateur sandwich avec correction conservatrice

Tout comme la correction de Fay-Graubard se veut une adaptation de [**36**] p.141, on adapte la correction conservatrice (cf. [**36**] p.141) en remplaçant le terme $\hat{\Delta}$ dans (2.1.7) par

$$\hat{\Delta}_C = n^{-1} \sum_{i=1}^{n} (F_i^2 m(Z_i; \hat{\theta}))^{\otimes 2}.$$

Ceci a comme effet d'ajuster à la hausse l'estimation de la variance de façon encore plus prononcée que la correction de Fay-Graubard. L'*estimateur sandwich avec correction conservatrice* pour la variance asymptotique de $\hat{\theta}$ est

$$\hat{V}_C = n^{-1} \hat{\Gamma}^{-1} \hat{\Delta}_C (\hat{\Gamma}^{-1})^\top.$$

Lorsque la fonction m utilisée est (2.1.1), on obtient un estimateur de la variance asymptotique de $\Psi(\hat{P}^*)$ en prenant l'élément (1,1) de la matrice $\hat{V}_C$. On pose

$$\hat{V}_3 = (\hat{V}_C)_{11}. \tag{2.1.9}$$

Exemple 2.1.1. *Dans un modèle de régression logistique*

$$g(L;\alpha) = \mathrm{expit}(\alpha_0 + \alpha_1 L_1 + \ldots + \alpha_r L_r),$$

le paramètre $\alpha = (\alpha_0, \ldots, \alpha_r)$ *est estimé par la méthode du maximum de vraisemblance. Autrement dit, par l'Exemple 1.3.5,* $\hat{\alpha}$ *est obtenu en résolvant l'équation estimante*

$$0 = \sum_{i=1}^{n} m_g(Z_i;\alpha)$$

où $m_g(z;\alpha)$ *est le score*

$$m_g(z;\alpha) = \frac{\partial}{\partial\alpha} \log p_{A|L}(a|\ell).$$

Puisque la fonction de masse $p_{A|L}(a|\ell)$ *est de la forme*

$$p_{A|L}(a|\ell) = g(\ell)^a (1 - g(\ell))^{1-a},$$

la log-vraisemblance s'écrit

$$\log p_{A|L}(a|\ell) = a \log g(L;\alpha) + (1-a)\log(1 - g(L;\alpha)).$$

En utilisant (2.1.10)*, on calcule*

$$m_g(z;\alpha) = (a - g(\ell;\alpha))\ell_+,$$

où $\ell_+ = (1 \ \ell_1 \ \ldots \ \ell_r)^\top$ *est la variable* ℓ *augmentée de* 1*. En utilisant* (2.1.4)*, on obtient*

$$\frac{\partial g(\ell;\alpha)}{\partial\alpha} = g(\ell;\alpha)(1 - g(\ell;\alpha))\ell^+, \tag{2.1.10}$$

et donc

$$\frac{\partial m_g}{\partial\alpha^\top} = -g(\ell;\alpha)(1 - g(\ell;\alpha))(\ell^+)^{\otimes 2}.$$

Ceci permet de calculer les trois estimateurs (2.1.6)*,* (2.1.8)*,* (2.1.9)*.*

2.2. Le Jackknife

Notons $\hat{\psi}^{(-i)}$ l'estimateur TMLE (1.6.13) calculé à partir du sous-échantillon $Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n$ (i.e. la i-ème observation est manquante). L'*estimateur Jackknife* **[36, 24]** de la variance du TMLE $\Psi(\hat{P}^*)$ (aucune observation manquante) est

$$\hat{V}_4 = \frac{n-1}{n} \sum_{i=1}^{n} (\mu_i - \overline{\mu})^2, \tag{2.2.1}$$

où

$$\mu_i = \frac{1}{n-1} \sum_{j \neq i} \hat{\psi}^{\hat{j}}$$

est la moyenne sur tous les TMLE $\hat{\psi}^{(-j)}$ calculés à partir d'un sous-échantillon contenant la i-ème observation, et où

$$\overline{\mu} = n^{-1} \sum_{i=1}^{n} \mu_i.$$

Notons que pour chaque i, le calcul de $\hat{\psi}^{(-i)}$ implique de ré-estimer g et Q_a à partir du sous-échantillon $Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n$.

Chapitre 3

Étude par simulation

3.1. Introduction

Dans ce chapitre, on présente les résultats d'une étude par simulation ayant pour but de comparer les estimateurs de la variance introduits au Chapitre 2 dans trois scénarios. L'estimateur $\hat{V}_0$ définit en (2.0.3) sera appelé Étalon-or , l'estimateur $\hat{V}_1$ définit en (2.1.6) sera appelé Sandwich , l'estimateur $\hat{V}_2$ définit en (2.1.8) sera appelé Corr. F-G , l'estimateur $\hat{V}_3$ définit en (2.1.9) sera appelé Corr. Cons. , l'estimateur $\hat{V}_4$ définit en (2.2.1) sera appelé Jackknife . Dans le premier scénario, on compare les estimateurs en fonction d'une suite décroissante de tailles échantillonnales. Dans le second scénario, on compare les estimateurs lorsque le score de propension $g(L)$ s'approche des valeurs extrêmes (0 et 1). Dans le troisième scénario, on compare les estimateurs dans diverses situations où les modèles pour g et/ou Q ne contiennent pas un ensemble suffisant de variables de confusion (voir 1.1 condition (C3)). Pour chacun de ces scénarios, on considère le cas où la variable réponse est une variable continue Y_c et obéit à un modèle de régression linéaire par rapport aux variables A,L et le cas où la variable réponse est une variable dichotomique Y_d obéissant à un modèle de régression logistique par rapport aux variables A,L. Dans les deux cas, on suppose que A obéit à un modèle de régression logistique par rapport à L. Chaque simulation comprend 20000 itérations et à chacune de ces itérations, les estimateurs de la variance sont calculés et comparés à la variance Monte-Carlo (introduite à la section 3.2) à l'aide de l'erreur relative

$$ER_k = \frac{\hat{V}_k - V_{MC}}{V_{MC}} \times 100, \tag{3.1.1}$$

où $\hat{V}$ est une notation générique désignant n'importe quel estimateur de la variance et $\hat{V}_k$ est l'estimation obtenue à la k-ième itération. Le biais relatif Monte-Carlo (en %) est donné par

$$BR = \frac{1}{20000} \sum_{k=1}^{20000} ER_k. \tag{3.1.2}$$

3.2. La variance Monte-Carlo

Pour chaque $k = 1, \ldots, K$, soit $Z_{k1}, \ldots, Z_{kn}$ un échantillon i.i.d. de taille n de la variable Z. On note $\hat{\psi}_n^{(k)}$ l'estimateur TMLE (1.6.13) calculé à partir du k-ème échantillon. Par *variance Monte-Carlo*, on entend la quantité

$$V_{MC} = (K-1)^{-1} \sum_{k=1}^{K} (\hat{\psi}_n^{(k)} - \overline{\psi}_n)^2, \tag{3.2.1}$$

où

$$\overline{\psi}_n = K^{-1} \sum_{k=1}^{K} \hat{\psi}_n^{(k)}$$

est la *moyenne Monte-Carlo*.

Proposition 3.2.1. *La variance Monte-Carlo converge vers la vraie variance* $Var(\Psi(\hat{P}^*))$.

Démonstration. Par la loi des grands nombres et le théorème de l'application continue, on a

$$\begin{aligned}
V_{MC} &= K^{-1} \sum_{k=1}^{K} (\hat{\psi}_n^{(k)} - \overline{\psi}_n)^2 \\
&= K^{-1} \sum_{k=1}^{K} (\hat{\psi}_n^{(k)})^2 - K^{-1} \sum_{k=1}^{K} 2\hat{\psi}_n^{(k)} \overline{\psi}_n + (\overline{\psi}_n)^2 \\
&= K^{-1} \sum_{k=1}^{K} (\hat{\psi}_n^{(i)})^2 - (\overline{\psi}_n)^2 \\
&\overset{K \to \infty}{\longrightarrow} E[\Psi(\hat{P}^*)^2] - E[\Psi(\hat{P}^*)]^2.
\end{aligned}$$

□

3.3. Scénarios considérés

3.3.1. Petites tailles échantillonnales

Dans ce scénario, le vecteur aléatoire L est constitué de deux composantes $L = (L_1,L_2)$ avec $L_1 \sim N(0;1)$, $L_2 \sim N(0;0{,}3)$. L'exposition A suit une loi de Bernoulli conditionnellement à L avec $g(L) = E[A|L] = \text{expit}\left(L_1 - \frac{1}{2}L_2\right)$. L'issue dichotomique Y_d suit une loi de Bernoulli conditionnellement à A et L avec $Q_A^d(L) = E[Y_d|A,L] = \text{expit}\,(0{,}3A - L_1 + L_2)$. L'issue continue Y_c suit une loi normale conditionnellement à A et L d'écart-type 1 et de moyenne $2A + L_1 - 10L_2$. Pour l'issue continue, la vraie valeur de l'effet causal moyen ψ_0^c est 2. Pour l'issue dichotomique, la vraie valeur de l'effet causal moyen ψ_0^d est estimé par simulation à 0,0609. Notons que les seules variables associées à la fois à A et à Y sont L_1 et L_2 donc tout ensemble L contenant L_1 et L_2 est un ensemble suffisant de facteurs de confusion. Pour chaque $n = 30{,}40{,}50,\ldots,130{,}140$, on prélève 20000 échantillons aléatoires $Z_1,\ldots,Z_n$ de taille n du vecteur aléatoire $Z = (Y,A,L)$ et l'estimateur TMLE de l'effet causal moyen est calculé. La variance Monte-Carlo est calculée à partir des 20000 valeurs du TMLE. Les estimations initiales $\hat{g}(L)$, $\hat{Q}_0(L)$, $\hat{Q}_1(L)$ utilisées dans le calcul du TMLE sont obtenues par ajustement des modèles de régression logistique et linéaire

$$\text{logit}(\hat{g}(L)) = \hat{\beta}_0^g + \hat{\beta}_1^g L_1 + \hat{\beta}_2^g L_2,$$

$$\text{logit}(\hat{Q}_A^d(L)) = \hat{\beta}_0^d + \hat{\beta}_1^d L_1 + \hat{\beta}_2^d L_2 + \hat{\beta}_3^d A,$$

$$\hat{Q}_A^c(L) = \hat{\beta}_0^c + \hat{\beta}_1^c L_1 + \hat{\beta}_2^c L_2 + \hat{\beta}_3^c A.$$

Pour les tailles échantillonales de $n = 20$ ou moins, il arrive que pour certaines itérations de la simulation, tous les sujets se voient assignés le même traitement $A = 0$ ou $A = 1$ (par pur hasard). Il en résulte les estimations $\hat{g}(L_i) = 0$, ce qui fait que la quantité $\hat{H}(A_i,L_i)$ apparaissant dans (1.6.11) n'est pas définie et donc l'étape de mise à jour TMLE ne peut être complétée. Par ailleurs, nous avons choisit de ne pas explorer les tailles échantillonales supérieures à 140 car nous jugeons que la plage $n = 30$ à $n = 140$ représente adéquatement le comportement relatif des différents estimateurs en présence de petites échantillons dans le présent contexte.

3.3.2. Scores de propension extrêmes

Dans ce scénario, le vecteur aléatoire L est constitué de deux composantes $L = (L_1,L_2)$ avec $L_1 \sim N(0;1)$, $L_2 \sim N(0;0{,}3)$. L'exposition A suit une loi de Bernoulli conditionnellement à L avec $g(L) = E[A|L] = \text{expit}\left(z + L_1 - \frac{1}{2}L_2\right)$, où z est un paramètre que nous faisons varier. Nous voyons donc que pour une valeur fixe de L, le score de propension apporche 0 lorsque z diminue et approche 1 lorsque z augmente. L'issue dichotomique Y_d suit une loi de Bernoulli conditionnellement à A et L avec $Q_A^d(L) = E[Y_d|A,L] = \text{expit}\,(2A - L_1 + L_2)$. L'issue continue Y_c suit une loi normale conditionnellement à A et L d'écart-type 1 et de moyenne $2A + L_1 - 10L_2$. Pour l'issue continue, la vraie valeur de l'effet causal moyen ψ_0^c est 2. Pour l'issue dichotomique, la vraie valeur de l'effet causal moyen ψ_0^d est estimé par simulation à 0,3417. Pour chaque valeur entière z entre -3 et 3, on prélève 20000 échantillons aléatoires $Z_1, \ldots, Z_n$ de taille $n = 1000$ du vecteur aléatoire $Z = (Y,A,L)$ et l'estimateur TMLE de l'effet causal moyen est calculé. La variance Monte-Carlo est calculée à partir des 20000 valeurs du TMLE. Les estimations initiales $\hat{g}(L)$, $\hat{Q}_0(L)$, $\hat{Q}_1(L)$ utilisés dans le calcul du TMLE sont obtenus par ajustement des mêmes modèles de régression logistique et linéaire qu'à la section 3.3.1. La taille d'échantillon de $n = 1000$ a été choisie dans le but de se placer dans le régime asymptotique, de manière à s'assurer qu'une différence dans le comportement des estimateurs soit attribuable aux scores de propension plutôt qu'à la taille échantillonale.

3.3.3. Modèles incomplets

Dans ce scénario, le vecteur aléatoire L est constitué de neuf composantes $L = (L_1,\ldots,L_9)$ avec $L_1 \sim N(0;1)$, $L_2 \sim N(0;0{,}3)$, $L_3 \sim N(0;0{,}6)$, $L_4 \sim N(0;1)$, $L_5 \sim N(0;0{,}9)$, $L_6 \sim N(0;0{,}2)$, $L_7 \sim N(0;0{,}5)$, $L_8 \sim N(0;1{,}1)$, $L_9 \sim N(0{,}5;0{,}3)$. L'exposition A suit une loi de Bernoulli conditionnellement à L avec $g(L) = E[A|L] = \text{expit}\left(-\frac{1}{4}L_4 + \frac{1}{5}L_5 - \frac{1}{6}L_6 + \frac{1}{7}L_7 - \frac{1}{8}L_8 + \frac{1}{9}L_9\right)$. L'issue dichotomique Y_d suit une loi de Bernoulli conditionnellement à A et L avec $Q_A^d(L) = E[Y_d|A,L] = \text{expit}\left(2A - L_1 + L_2 - \frac{1}{2}L_3 + \frac{1}{2}L_4 - \frac{1}{3}L_5 + \frac{1}{3}L_6\right)$. L'issue continue Y_c suit une loi normale conditionnellement à A et L d'écart-type 1 et de moyenne $2A + L_1 - 10L_2 + 5L_3 - 7L_4 + 2L_5 - 6L_6$. Pour l'issue continue, la vraie valeur de l'effet causal moyen ψ_0^c est 2. Pour l'issue dichotomique, la vraie valeur de l'effet causal moyen

ψ_0^d est estimé par simulation à 0,3287. On prélève 20000 échantillons aléatoires $Z_1, \ldots, Z_n$ de taille $n = 200$ du vecteur aléatoire $Z = (Y,A,L)$ et l'estimateur TMLE de l'effet causal moyen est calculé en ajustant divers modèles pour obtenir les estimations initiales $\hat{g}(L)$, $\hat{Q}_0(L)$, $\hat{Q}_1(L)$. La taille d'échantillon de $n = 200$ se veut être un entre-deux entre le régime asymptotique $n = 1000$ de la section 3.3.2 et le régime des petits échantillon considéré à la section 3.3.1. En effet, nous ne voulons pas que les différence observées au niveau des estimateurs soient dûes à la petite taille des échantillons, mais nous ne voulons pas non plus entrer dans le régime asymptotique où nous savons que certaines des situations considérées ci-bas coincident (Proposition 1.5.4).

Rappelons que même si A et Y dépendent de toutes les variables $L_1, \ldots, L_9$, il suffit pour que l'estimateur TMLE soit convergent d'identifier un ensemble suffisant de facteurs de confusion $L' \subset L = (L_1, \ldots, L_9)$ et de modéliser $g(L') = E[A|L']$ et $Q_A(L') = E[Y|A,L']$ à l'aide de deux estimateurs $\hat{g}(L')$ $\hat{Q}_A(L')$ dont au moins un est convergent. Il sera donc intéressant de comparer les estimateurs de la variance dans les six situations suivantes:

Situation 1: $L' = L$ et les modèles pour $g(L')$ et $Q_a(L')$ sont exacts.

Situation 2: $L' = L$ et le modèle pour $g(L')$ omet certaines des variables de L' tandis que le modèle pour $Q_a(L')$ est exact.

Situation 3: $L' = L$ et le modèle pour $g(L')$ est exact tandis que le modèle pour $Q_a(L')$ omet certaines des variables de L'.

Situation 4: $L' = (L_4,L_5,L_6)$ (soit le plus petit ensemble suffisant de facteurs de confusion) et les modèles pour $g(L')$ et $Q_a(L')$ contiennent toutes les variables de L'.

Situation 5: $L' = (L_4,L_5,L_6)$ (soit le plus petit ensemble suffisant de facteurs de confusion) et le modèle pour $g(L')$ omet certaines des variables de L' tandis que le modèle pour $Q_a(L')$ contient toutes les variables de L'.

Situation 6: $L' = (L_4,L_5,L_6)$ (soit le plus petit ensemble suffisant de facteurs de confusion) et le modèle pour $g(L')$ contient toutes les variables de L' tandis que le modèle pour $Q_a(L')$ omet certaines des variables de L'.

Notons qu'en terme des modèles, les situations 4, 5, 6 sont une répétition des situations 1, 2, 3 respectivement, avec comme seule différence la nature des variables L'. Plus précisément, les situations considérées sont les suivantes.

3.3.3.1. *Situation 1*

Dans cette situation où $L' = L$ et les modèles pour $g(L')$ et $Q_a(L')$ sont exact, les modèles suivants sont utilisés:

$$\text{logit}(\hat{g}(L)) = \hat{\beta}_0^g + \hat{\beta}_1^g L_4 + \hat{\beta}_2^g L_5 + \hat{\beta}_3^g L_6 + \hat{\beta}_4^g L_7 + \hat{\beta}_5^g L_8 + \hat{\beta}_6^g L_9,$$

$$\text{logit}(\hat{Q}_A^d(L)) = \hat{\beta}_0^d + \hat{\beta}_1^d L_1 + \hat{\beta}_2^d L_2 + \hat{\beta}_3^d L_3 + \hat{\beta}_4^d L_4 + \hat{\beta}_5^d L_5 + \hat{\beta}_6^d L_6 + \hat{\beta}_7^d A,$$

$$\hat{Q}_A^c(L) = \hat{\beta}_0^c + \hat{\beta}_1^c L_1 + \hat{\beta}_2^c L_2 + \hat{\beta}_3^c L_3 + \hat{\beta}_4^c L_4 + \hat{\beta}_5^c L_5 + \hat{\beta}_6^c L_6 + \hat{\beta}_7^c A.$$

3.3.3.2. *Situation 2*

Dans cette situation où $L' = L$ et le modèle pour $g(L')$ omet certaines des variables de L' tandis que le modèle pour $Q_a(L')$ est exact, les modèles suivants sont utilisés:

$$\text{logit}(\hat{g}(L)) = \hat{\beta}_0^g,$$

$$\text{logit}(\hat{Q}_A^d(L)) = \hat{\beta}_0^d + \hat{\beta}_1^d L_1 + \hat{\beta}_2^d L_2 + \hat{\beta}_3^d L_3 + \hat{\beta}_4^d L_4 + \hat{\beta}_5^d L_5 + \hat{\beta}_6^d L_6 + \hat{\beta}_7^d A,$$

$$\hat{Q}_A^c(L) = \hat{\beta}_0^c + \hat{\beta}_1^c L_1 + \hat{\beta}_2^c L_2 + \hat{\beta}_3^c L_3 + \hat{\beta}_4^c L_4 + \hat{\beta}_5^c L_5 + \hat{\beta}_6^c L_6 + \hat{\beta}_7^c A.$$

3.3.3.3. *Situation 3*

Dans cette situation où $L' = L$ et le modèle pour $g(L')$ est exact tandis que le modèle pour $Q_a(L')$ omet certaines des variables de L', les modèles suivants sont utilisés:

$$\text{logit}(\hat{g}(L)) = \hat{\beta}_0^g + \hat{\beta}_1^g L_4 + \hat{\beta}_2^g L_5 + \hat{\beta}_3^g L_6 + \hat{\beta}_4^g L_7 + \hat{\beta}_5^g L_8 + \hat{\beta}_6^g L_9,$$

$$\text{logit}(\hat{Q}_A^d(L)) = \hat{\beta}_0^d + \hat{\beta}_1^d A,$$

$$\hat{Q}_A^c(L) = \hat{\beta}_0^c + \hat{\beta}_1^c A$$

3.3.3.4. *Situation 4*

Dans cette situation où $L' = (L_4, L_5, L_6)$ et les modèles pour $g(L')$ et $Q_a(L')$ contiennent toutes les variables de L', les modèles suivants sont utilisés:

$$\text{logit}(\hat{g}(L')) = \hat{\beta}_0^g + \hat{\beta}_1^g L_4 + \hat{\beta}_2^g L_5 + \hat{\beta}_3^g L_6,$$

$$\text{logit}(\hat{Q}_A^d(L')) = \hat{\beta}_0^d + \hat{\beta}_1^d L_4 + \hat{\beta}_2^d L_5 + \hat{\beta}_3^d L_6 + \hat{\beta}_4^d A,$$

$$\hat{Q}_A^c(L) = \hat{\beta}_0^c + \hat{\beta}_1^c L_4 + \hat{\beta}_2^c L_5 + \hat{\beta}_3^c L_6 + \hat{\beta}_4^c A.$$

3.3.3.5. *Situation 5*

Dans cette situation où $L' = (L_4, L_5, L_6)$ et le modèle pour $g(L')$ omet certaines des variables de L' tandis que le modèle pour $Q_a(L')$ contient toutes les variables de L', les modèles suivants sont utilisés:

$$\text{logit}(\hat{g}(L')) = \hat{\beta}_0^g,$$

$$\text{logit}(\hat{Q}_A^d(L')) = \hat{\beta}_0^d + \hat{\beta}_1^d L_4 + \hat{\beta}_2^d L_5 + \hat{\beta}_3^d L_6 + \hat{\beta}_4^d A,$$

$$\hat{Q}_A^c(L) = \hat{\beta}_0^c + \hat{\beta}_1^c L_4 + \hat{\beta}_2^c L_5 + \hat{\beta}_3^c L_6 + \hat{\beta}_4^c A.$$

3.3.3.6. *Situation 6*

Dans cette situation où $L' = (L_4, L_5, L_6)$ et le modèle pour $g(L')$ contient toutes les variables de L' tandis que le modèle pour $Q_a(L')$ omet certaines des variables de L', les modèles suivants sont utilisés:

$$\text{logit}(\hat{g}(L')) = \hat{\beta}_0^g + \hat{\beta}_4^g L_4 + \hat{\beta}_5^g L_5 + \hat{\beta}_6^g L_6,$$

$$\text{logit}(\hat{Q}_A^d(L')) = \hat{\beta}_0^d + \hat{\beta}_1^d A,$$

$$\hat{Q}_A^c(L) = \hat{\beta}_0^c + \hat{\beta}_1^c A.$$

3.4. Résultats

Pour chacune des 20000 itérations d'une simulation, la sortie consiste en une estimation TLME $\hat{\psi}$ de l'effet causal moyen ainsi qu'en les estimations de la variance $\hat{V}_1, \ldots, \hat{V}_4$ définies au Chapitre 2. Nous présentons les résultats sous forme d'un graphique du biais relatif (3.1.2) ainsi qu'un tableau contenant le biais relatif et le taux de couverture des intervalles de confiances associés. Le taux de couverture (en %) représente la fraction des itérations pour lesquelles l'intervalle de confiance $\hat{\psi} \pm 1{,}96\sqrt{\hat{V}_j/n}$ contient la vraie valeur de l'effet causal moyen. Rappelons qu'un intervalle de confiance asymptotique pour $\hat{\psi}$ est donné par (1.3.3). Donc à condition que la taille échantillonale soit suffisament grande pour que l'approximation asymptotique (1.3.3) soit valide et que $\hat{V}_j$ approxime bien la vraie valeur de la variance, alors on s'attend à observer un taux de couverture près de 95%. Nous rapportons également le biais relatif de l'estimateur TMLE que l'on s'attend à trouver petit en vertu de la Proposition 1.5.4.

3.4.1. Petites tailles échantillonnales

3.4.1.1. *Issue dichotomique*

À la Figure 3.1, on voit qu'à partir de $n = 40$, toutes les méthodes sauf le Jackknife affichent un comportement similaire. Notamment, elles sous-estiment la vraie valeur de la variance mais le biais relatif tend à s'approcher de 0 lorsque n croît. Dans **[36, 10]**, l'estimateur Corr. F-G est présenté comme une alternative à l'estimateur Sandwich ayant la propriété que d'être plus grand dans le cas de petits échantillons où il est connu que l'estimateur Sandwich sous-estime la vraie valeur de la variance. De même l'estimateur Corr. Cons. est présenté dans **[36]** comme une alternative à l'estimateur Corr. F-G ayant la propriété que d'être encore plus grand que Corr. F-G. Ainsi, nous ne somme pas surpris d'observer que l'estimateur Corr. Cons. est plus grand que l'estimateur Corr. F-G qui lui est à son tour plus grand que l'estimateur Sandwich. Nous ne somme pas non plus surpris de constater que l'estimateur Sandwich est moins biaisé que l'Étalon-or puisqu'il s'agit dans les deux cas d'estimateurs de type sandwich mais le premier tient compte du fait que le score de propension est estimé paramétriquement. Quant à lui, le Jackknife surestime la vraie valeur de la variance avec un biais relatif absolu supérieur aux autres méthodes mais qui tend lui aussi à s'approcher de 0 lorsque n croît.

Le Tableau 3.1 révèle quant à lui le fait remarquable que le taux de couverture de l'intervalle de confiance associé au Jackknife est virtuellement indépendant de la taille échantillonnale avec un taux se maintenant entre 94% et 95% pour toutes les valeurs de n, y compris $n = 30$ et $n = 40$ où le biais relatif Monte-Carlo est respectivement de 216% et 102%. Pour les autres méthodes, le taux de couverture se comporte de façon similaire au biais relatif Monte-Carlo, approchant lentement 95% par le bas.

Fig. 3.1. Biais relatif Monte-Carlo en fonction de la taille échantillonnale (issue dichotomique)

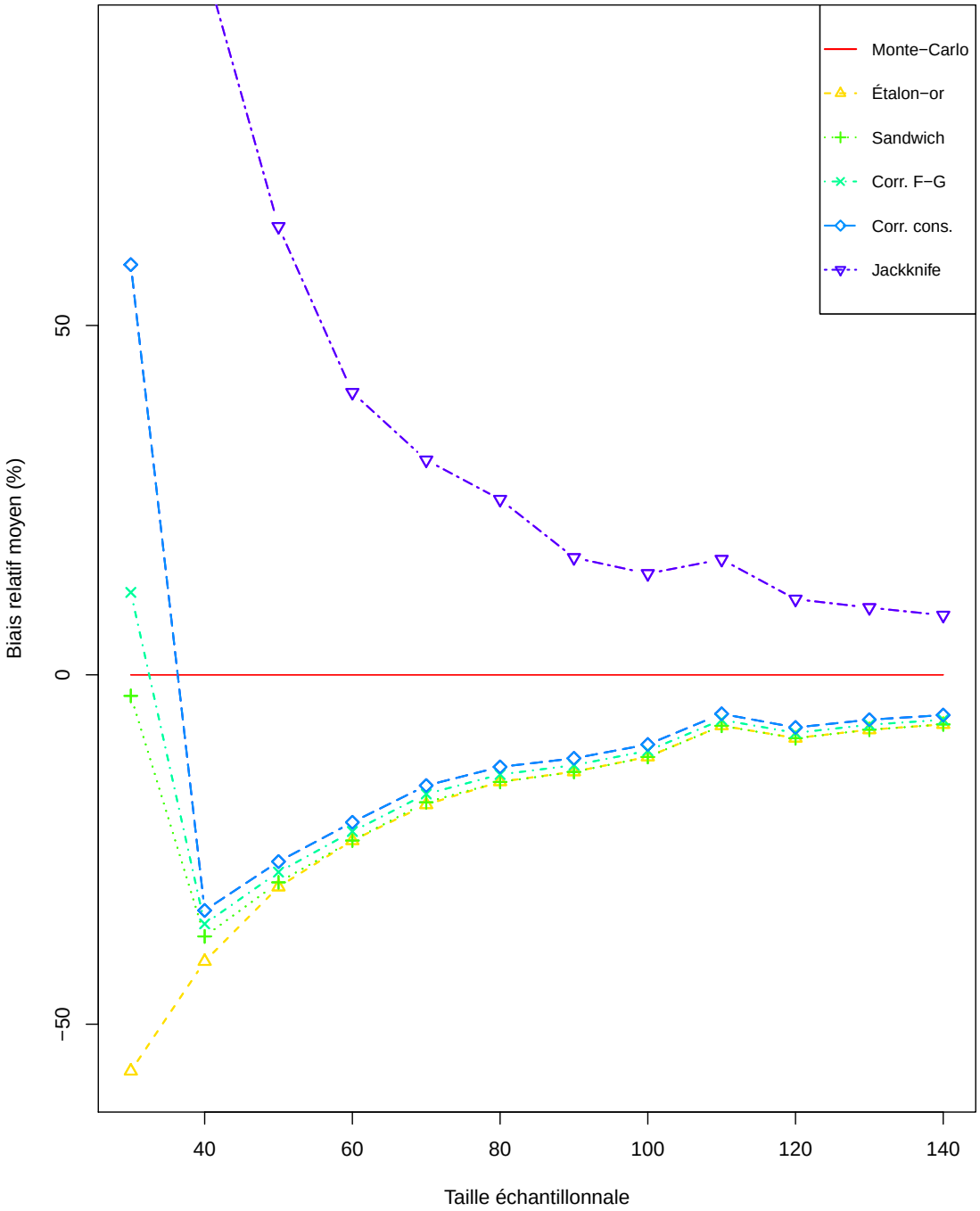

Tableau 3.1. Estimateurs de la variance en fonction de la taille échantillonnale n (issue dichotomique)

n	Biais relatif moyen TMLE	V_{MC} ($\times 10^{-3}$)	Biais relatif Monte-Carlo (taux de couverture)				
			Étalon-or	Sandwich	Corr, F-G	Corr, Cons,	Jackknife
30	-17	60	-57 (81,9)	-3 (81,9)	12 (82,5)	59 (83,2)	216 (94,7)
40	-9	36	-41 (86,3)	-37 (86,2)	-36 (86,6)	-34 (87,1)	102 (94,2)
50	-9	25	-30 (88,6)	-30 (88,6)	-28 (88,8)	-27 (89,2)	64 (94,2)
60	-4	20	-24 (90,0)	-24 (90,0)	-22 (90,2)	-21 (90,4)	40 (94,2)
70	-5	16	-19 (91,0)	-18 (91,0)	-17 (91,1)	-16 (91,4)	31 (94,3)
80	-4	14	-15 (91,9)	-15 (91,9)	-14 (92,1)	-13 (92,3)	25 (94,7)
90	-1	12	-14 (92,2)	-14 (92,2)	-13 (92,4)	-12 (92,5)	17 (94,5)
100	-2	11	-12 (92,7)	-12 (92,7)	-11 (92,8)	-10 (92,9)	14 (94,8)
110	0	9	-7 (93,1)	-7 (93,1)	-6 (93,2)	-6 (93,3)	17 (94,9)
120	0	9	-9 (92,9)	-9 (92,9)	-8 (93,0)	-8 (93,1)	11 (94,5)
130	0	8	-8 (93,2)	-8 (93,2)	-7 (93,4)	-6 (93,5)	10 (94,8)
140	1	7	-7 (93,6)	-7 (93,6)	-6 (93,7)	-6 (93,7)	9 (95,0)

3.4.1.2. *Issue continue*

À la Figure 3.2, on constate que les estimateur Étalon-or, Sandwich, Corr. F-G et Corr. Cons. se comportent de façon semblable au cas de l'issue dichotomique, soit en sous-estimant la vraie valeur de la variance mais en s'en approchant lorsque la taille échantillonnale augmente. Il en va de même pour le Jackknife, surestimant la vraie valeur de la variance mais s'en approchant lorsque la taille échantillonnale augmente.

En ce qui concerne le taux de couverture, le Tableau 3.2 révèle que les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons. donnent des taux de couvertures presque identiques au cas de l'issue dichotomique, soit approchant lentement 95% vers le bas. De façon similaire au cas de l'issue dichotomique, le Jackknife donne systématiquement un taux de couverture entre 95% et 96%, soit une légère sur-couverture.

Fig. 3.2. Biais relatif Monte-Carlo en fonction de la taille échantillonnale (issue continue)

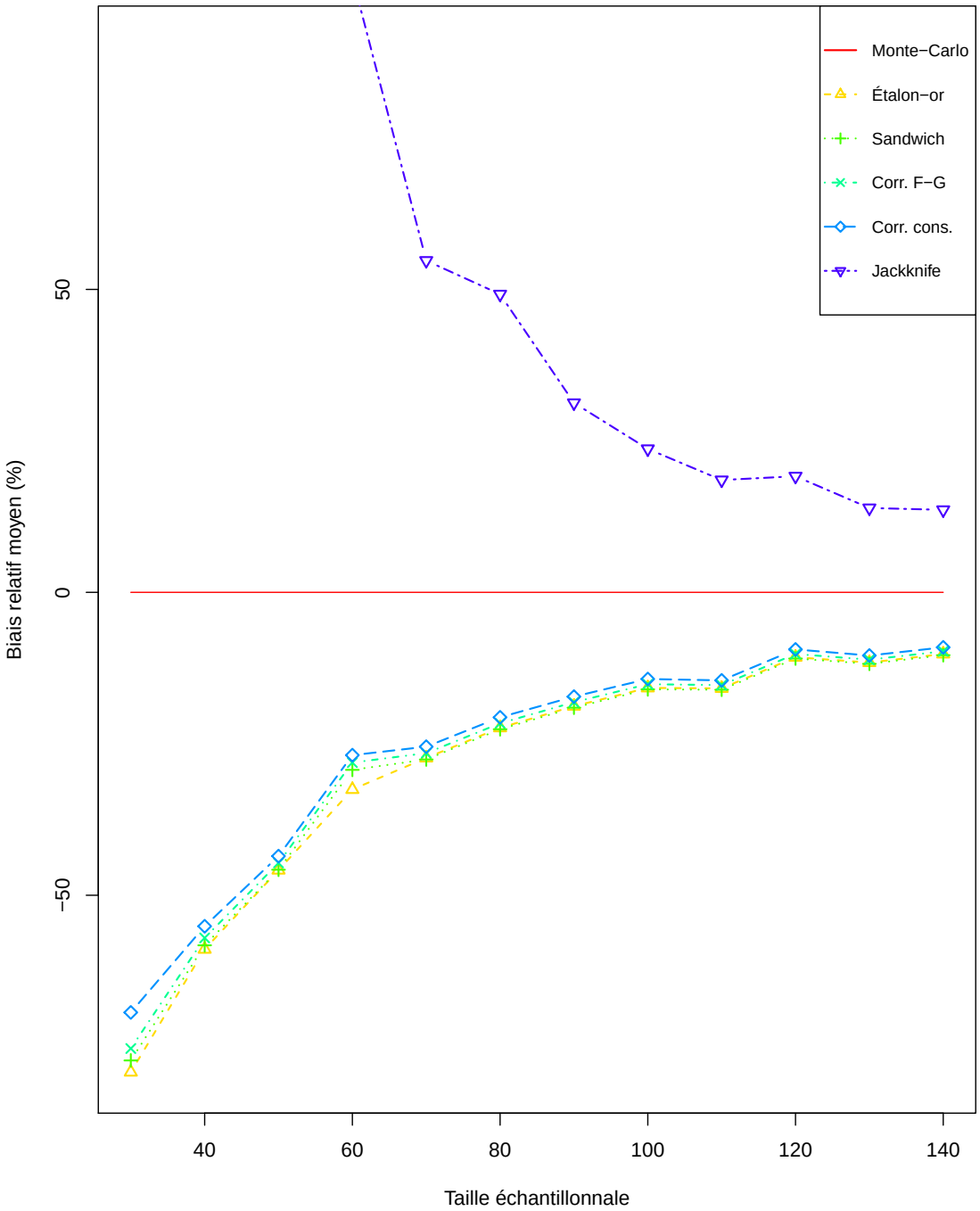

Tableau 3.2. Estimateurs de la variance en fonction de la taille échantillonnale n (issue continue)

n	Biais relatif Monte-Carlo TMLE ($\times 10^{-2}$)	V_{MC} ($\times 10^{-2}$)	Biais relatif Monte-Carlo (taux de couverture)				
			Étalon-or	Sandwich	Corr, F-G	Corr, Cons,	Jackknife
30	16	89	-79 (81,6)	-77 (81,6)	-75 (82,5)	-69 (83,2)	529 (96,2)
40	-3	29	-59 (85,5)	-58 (85,4)	-57 (86,0)	-55 (86,4)	309 (95,9)
50	2	17	-46 (87,8)	-46 (87,8)	-45 (88,2)	-44 (88,6)	182 (95,5)
60	-9	12	-33 (89,2)	-29 (89,1)	-28 (89,4)	-27 (89,7)	102 (95,2)
70	-24	9	-27 (89,9)	-28 (89,8)	-27 (90,1)	-25 (90,3)	55 (95,1)
80	-14	8	-22 (91,3)	-23 (91,3)	-22 (91,5)	-21 (91,7)	49 (95,6)
90	-5	7	-19 (91,6)	-19 (91,6)	-18 (91,8)	-17 (92,0)	31 (95,2)
100	-7	6	-16 (92,1)	-16 (92,0)	-15 (92,1)	-14 (92,3)	24 (95,3)
110	12	5	-16 (91,9)	-16 (91,9)	-15 (92,0)	-15 (92,1)	19 (94,9)
120	-2	5	-11 (92,6)	-11 (92,6)	-10 (92,8)	-9 (92,9)	19 (95,2)
130	-4	4	-12 (92,9)	-12 (92,9)	-11 (93,0)	-10 (93,1)	14 (95,4)
140	3	4	-10 (93,0)	-10 (93,0)	-10 (93,1)	-9 (93,2)	14 (95,2)

3.4.2. Scores de propensions extrêmes

3.4.2.1. *Issue dichotomique*

En observant la Figure 3.3, on constate que les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons. se suivent tous de près avec un biais relatif négatif. La courbe du Jackknife est essentiellement le reflet symétrique des courbes des estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons.. On note que la qualité de l'estimation s'améliore lorsque $|z|$ décroît, i.e. lorsque le score de propension s'éloigne de 0 et de 1.

Le Tableau 3.3 confirme que les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons. sont presque complètement confondus dans ce scénario et qu'ils sont comparables au Jackknife en terme de biais relatif absolu. Notons également que le Jackknife surestime la variance tandis que les autres méthodes la sous-estime. Au niveau du taux de couverture, le Jackknife performe strictement mieux que les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons..

Fig. 3.3. Biais relatif Monte-Carlo en fonction de la taille échantillonnale (issue dichotomique)

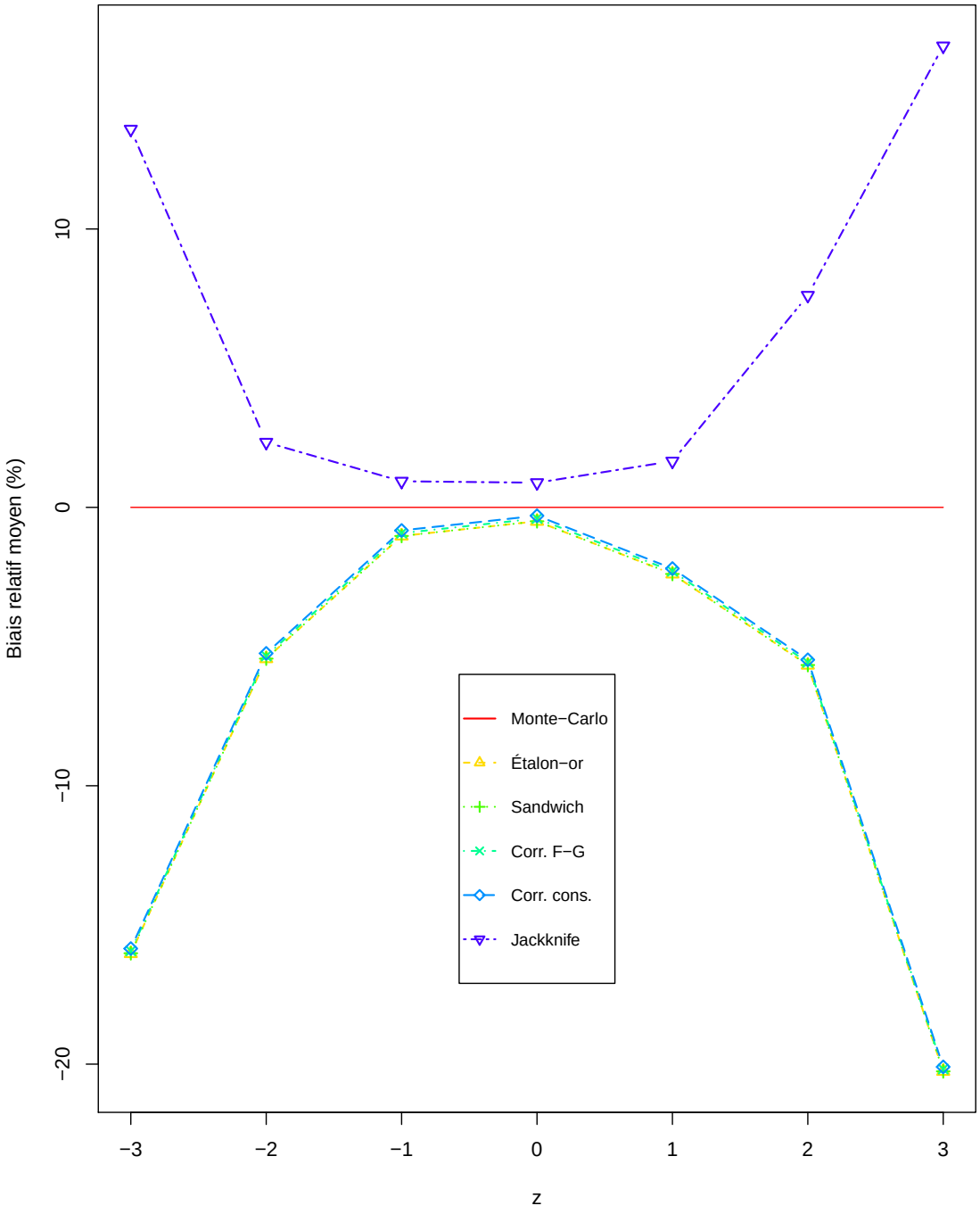

Tableau 3.3. Estimateurs de la variance en fonction du paramètre z (issue dichotomique)

z	Biais relatif Monte-Carlo TMLE ($\times 10^{-2}$)	V_{MC} ($\times 10^{-4}$)	Biais relatif Monte-Carlo (taux de couverture)				
			Étalon-or	Sandwich	Corr, F-G	Corr, Cons,	Jackknife
-3	-23	26	-16 (89,6)	-16 (89,6)	-16 (89,6)	-16 (89,6)	14 (91,2)
-2	-15	11	-5 (93,5)	-5 (93,5)	-5 (93,5)	-5 (93,5)	2 (94,1)
-1	-12	7	-1 (94,9)	-1 (94,9)	-1 (94,9)	-1 (94,9)	1 (95,0)
0	0	7	-1 (94,9)	0 (94,9)	0 (94,9)	0 (94,9)	1 (95,1)
1	7	12	-2 (94,4)	-2 (94,4)	-2 (94,5)	-2 (94,5)	2 (94,9)
2	-4	25	-6 (93,0)	-6 (93,0)	-6 (93,0)	-5 (93,0)	8 (94,5)
3	-16	67	-20 (88,9)	-20 (88,9)	-20 (88,9)	-20 (88,9)	17 (92,9)

3.4.2.2. *Issue continue*

Dans le cas de l'issue continue, la Figure 3.4 révèle que les estimateurs se comportent de manière en tous points similaire au cas de l'issue dichotomique.

Au Tableau 3.4, on voit qu'encore ici, le Jackknife donne lieu à des taux de couverture strictement supérieurs aux autres candidats. Un fait particulièrement remarquable est que pour $z = -3$ et $z = 3$, les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons. produisent un taux de couverture de 88% tandis que le Jackknife produit un taux de couverture de 94%.

Fig. 3.4. Biais relatif Monte-Carlo en fonction de la taille échantillonnale (issue continue)

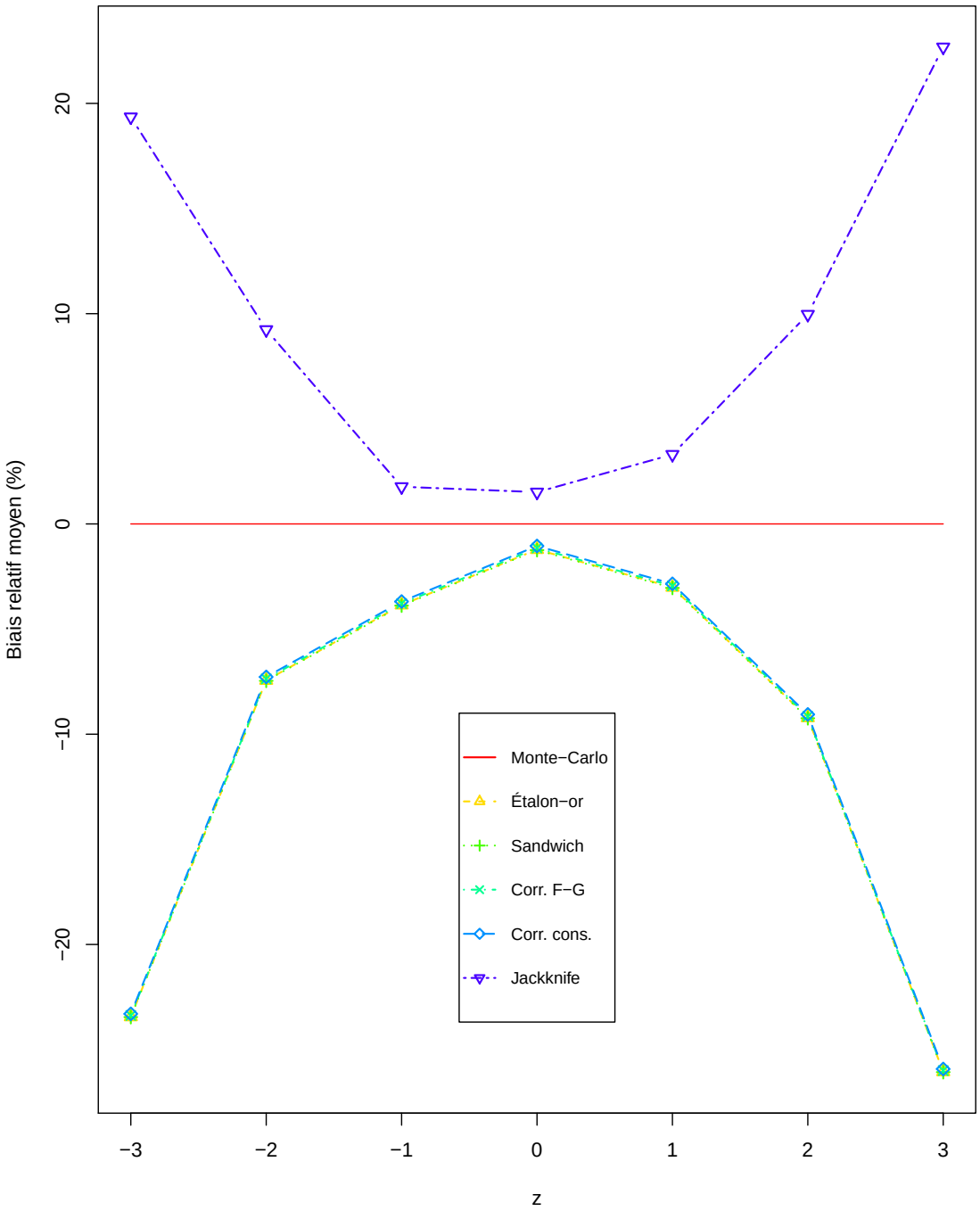

Tableau 3.4. Estimateurs de la variance en fonction du paramètre z (issue continue)

z	Biais relatif Monte-Carlo TMLE ($\times 10^{-3}$)	V_{MC} ($\times 10^{-2}$)	Biais relatif Monte-Carlo (taux de couverture)				
			Étalon-or	Sandwich	Corr, F-G	Corr, Cons,	Jackknife
-3	-77	36	-23 (88,4)	-23 (88,4)	-23 (88,4)	-23 (88,5)	19 (93,7)
-2	-5	14	-7 (92,6)	-7 (92,6)	-7 (92,6)	-7 (92,6)	9 (94,5)
-1	-26	7	-4 (94,1)	-4 (94,1)	-4 (94,1)	-4 (94,1)	2 (94,9)
0	25	5	-1 (94,7)	-1 (94,7)	-1 (94,7)	-1 (94,7)	2 (95,0)
1	-21	7	-3 (94,3)	-3 (94,3)	-3 (94,3)	-3 (94,3)	3 (95,1)
2	-56	14	-9 (92,3)	-9 (92,3)	-9 (92,3)	-9 (92,3)	10 (94,6)
3	3	37	-26 (87,6)	-26 (87,6)	-26 (87,6)	-26 (87,6)	23 (93,8)

3.4.3. Modèles incomplets

3.4.3.1. *Issue dichotomique*

La Figure 3.5 contient les boîtes à moustache pour la distribution de l'erreur relative de chaque estimateur. Le biais relatif Monte-Carlo s'y trouve représenté par un point rouge. Les valeurs aberrantes, définies comme ces valeurs qui font plus d'une fois et demi le troisième quartile ou moins d'une fois et demi le premier quartile, ne sont pas illustrées. Dans cette figure, on observe que les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons. ont tendance à sous-estimer la vraie valeur de la variance, sauf aux situations 3 et 6, qui correspondent au cas où le modèle pour Q est mal spécifié. De plus, il est intéressant de noter que les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons. performent mieux à la situation 4 qu'à la situation 1 et à la situation 5 qu'à la situation 2. Autrement dit, il est préférable de travailler avec un ensemble suffisant de facteurs de confusion qu'avec toutes les variables L. En terme de biais relatif absolu, la méthode du Jackknife ne donne des meilleurs résultats que la méthode du sandwich avec correction conservative qu'à la situation 6 mais elle se démarque par sa constance. En effet, le biais relatif du Jackknife est positif avec une valeur aux alentours de $6 \pm 3\%$ dans toutes les situations.

Le Tableau 3.5 confirme ce que nous observons des boîtes à moustaches. Au niveau du taux de couverture, l'estimateur sandwich avec correction conservative et le Jackknife donnent tous deux de bons résultats mais à la différence près que les invervalles de confiances associés à l'estimateur sandwich avec correction conservative ont tendance à sous-couvrir l'effet causal moyen tandis que ceux associés au Jackknife ont tendance à le sur-couvrir. Notons qu'en accord avec la Proposition 1.5.4, l'estimateur TMLE affiche un biais relatif absolu petit dans toutes les situations, soit inférieur ou égal à 0,2715%.

Fig. 3.5. Erreurs relatives des estimations de la variance (issue dichotomique)

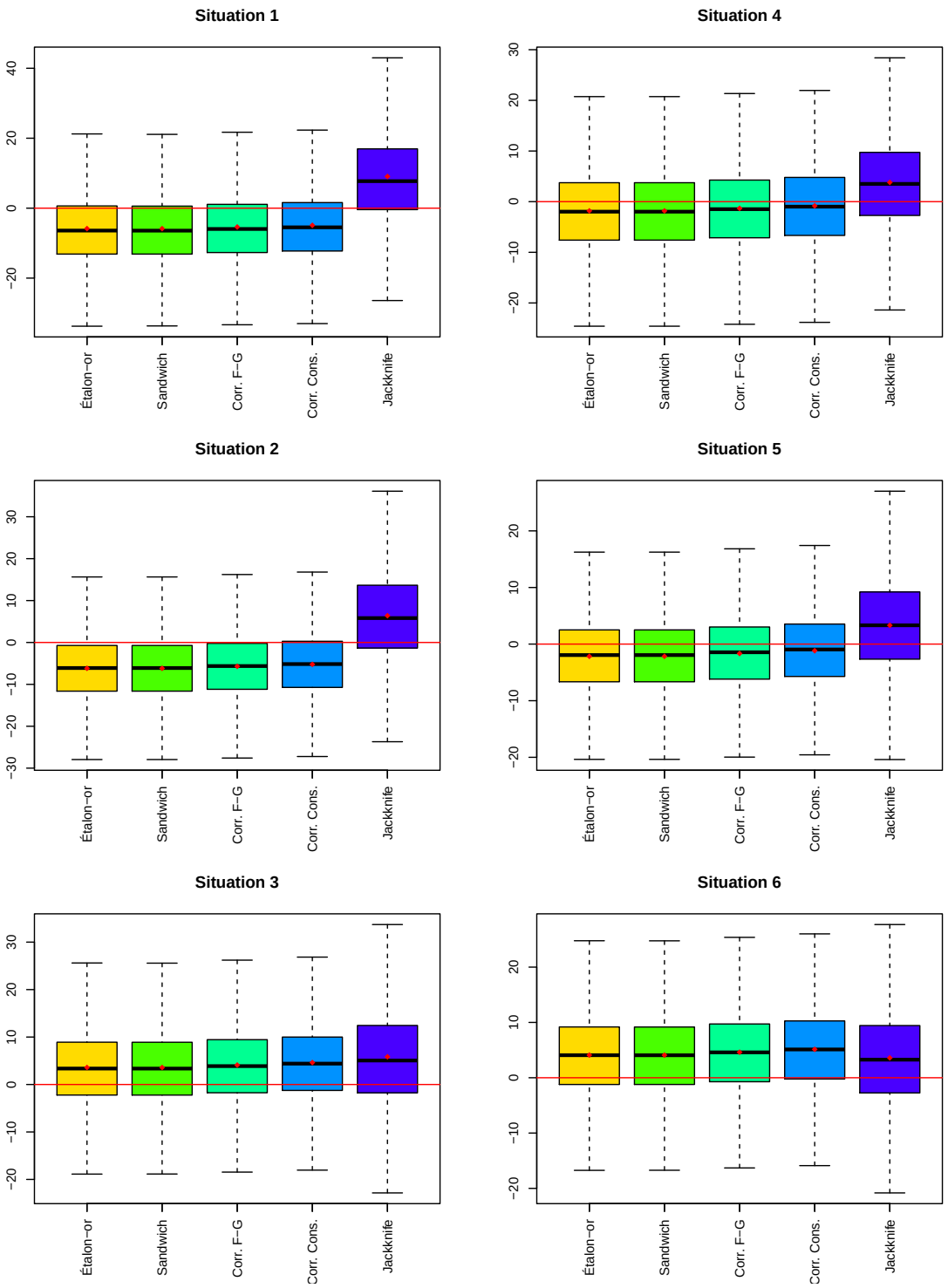

Tableau 3.5. Estimateurs de la variance en fonction du scénario (issue dichotomique)

Scénario	Biais relatif Monte-Carlo TMLE ($\times 10^{-2}$)	V_{MC} ($\times 10^{-3}$)	Biais relatif Monte-Carlo (taux de couverture)				
			Étalon-or	Sandwich	Corr, F-G	Corr, Cons,	Jackknife
1	-2191	34	-6 (94,0)	-6 (94,0)	-5 (94,1)	-5 (94,1)	9 (95,6)
2	-2715	33	-6 (94,6)	-6 (94,6)	-6 (94,6)	-5 (94,7)	6 (95,2)
3	62	40	4 (95,3)	4 (95,3)	4 (95,4)	5 (95,4)	6 (95,2)
4	-475	39	-2 (94,5)	-2 (94,5)	-1 (94,6)	-1 (94,7)	4 (95,2)
5	-1130	38	-2 (94,0)	-2 (94,0)	-2 (94,1)	-1 (94,1)	3 (95,3)
6	-4	39	4 (95,2)	4 (95,2)	5 (95,2)	5 (95,3)	4 (95,3)

3.4.3.2. *Issue continue*

Dans l'ensemble, le comportement des estimateurs de la variance affiché à la Figure 3.6 est le même que dans le cas de l'issue dichotomique, à l'exception des situations 3 et 6 où seul le Jackknife donne un biais relatif qui n'est pas de l'ordre de 250. Les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons. sont donc très affectés par une mauvaise spécification du modèle pour Q. La méthode du Jackknife étant de nature non-paramétrique, il n'en va pas de même pour elle. Notons qu'à la situation 3, la moyenne pour les estimateurs Sandwich, Corr. F-G et Corr. Cons. n'apparaît pas sur le graphique à cause d'une poignée de données aberrantes si grandes que la moyenne pour ces estimateurs se situe en dehors des bornes du graphique.

Pour les taux de couverture au Tableau 3.6, on observe la même tendance que pour le biais relatif avec le Jackknife qui donne un taux de couverture variant entre 95% et 96%, et ce même aux situations 3 et 6 où les autres méthodes donnent des intervalles de confiance de presque 100%.

Fig. 3.6. Erreurs relatives des estimations de la variance (issue continue)

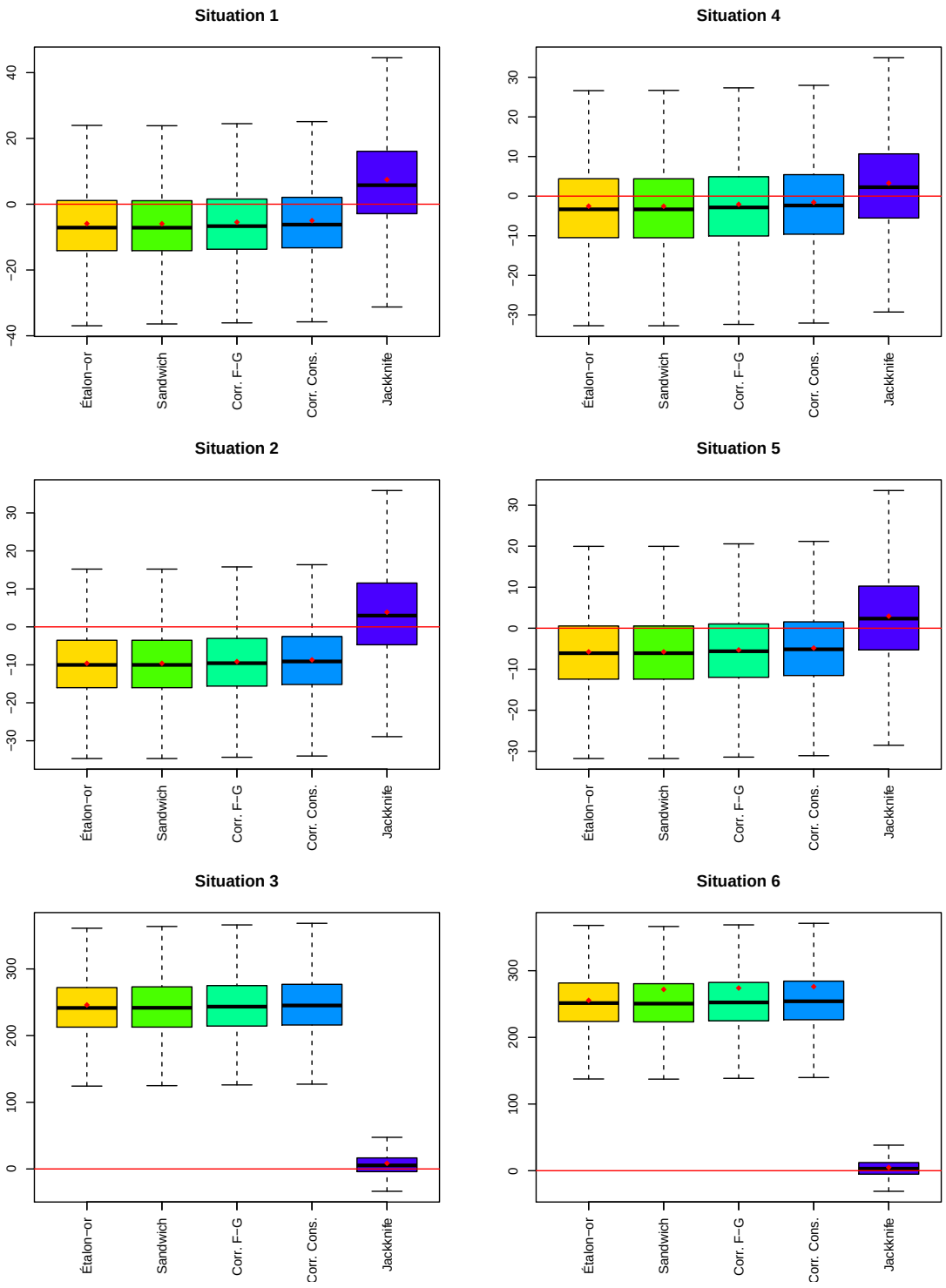

Tableau 3.6. Estimateurs de la variance en fonction du scénario (issue continue)

Scénario	Biais relatif Monte-Carlo TMLE ($\times 10^{-2}$)	V_{MC} ($\times 10^{-3}$)	Biais relatif Monte-Carlo (taux de couverture)				
			Étalon-or	Sandwich	Corr, F-G	Corr, Cons,	Jackknife
1	1	22	-6 (94,2)	-6 (94,2)	-5 (94,2)	-5 (94,3)	7 (95,6)
2	1	21	-10 (94,7)	-10 (94,7)	-9 (94,8)	-9 (94,8)	4 (95,4)
3	21	449	246 (100,0)	602 (100,0)	608 (100,0)	615 (100,0)	9 (95,3)
4	2	419	-3 (94,2)	-3 (94,2)	-2 (94,3)	-2 (94,4)	3 (95,3)
5	1	417	-6 (93,8)	-6 (93,8)	-5 (93,8)	-5 (93,9)	3 (95,4)
6	14	428	256 (100,0)	272 (100,0)	274 (100,0)	276 (100,0)	5 (95,6)

3.5. Conclusion

Dans tous les cas sauf au scénario 3 lorsque le modèle pour Q est mal spécifié, les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons. sous-estiment la vraie valeur de la variance, et donc parmi ces estimateurs, Corr. Cons. devrait être préféré puisqu'il est toujours le plus grand. Cependant, il est également apparent que parmi tous les candidats, c'est le Jackknife qui performe le mieux, tant au niveau du biais relatif que du taux de couverture. En fait, le Jackknife donne lieu à des taux de couverture de $95 \pm 1\%$ dans toutes les situations considérées sauf lorsque Y est dichotomique et que les scores de propension sont les plus extrêmes. De plus, le fait que les estimateurs Étalon-or, Sandwich, Corr. F-G et Corr. Cons. ont tendance à sous-estimer la vraie valeur de la variance se traduit par le fait qu'ils donnent lieu à des intervalles de confiance trop courts. Au contraire, le Jackknife à davantage tendance à surestimer la vraie valeur de la variance, ce qui donne lieu à des intervalles de confiance conservateurs, ce qui est généralement préférable.

Conclusion

Dans ce mémoire, nous avons exposé la procédure de mise à jour TMLE en mettant l'emphase sur l'estimation de l'effet causal moyen pour une exposition binaire. Afin de comprendre le comportement asymptotique et la propriété de double robustesse du TMLE, nous avons exposé la théorie des estimateurs asymptotiquement linéaires en utilisant autant que possible des arguments géométriques qui, nous l'espérons, rendent les résultats évidents. Ensuite, nous avons investigué diverses méthodes d'estimation de la variance du TMLE dans le but de les comparer à l'étalon-or [**43**]. Dans le cas où le score de propension est estimé paramétriquement, l'estimateur sandwich [**35**] est disponible ainsi que la correction à l'estimateur sandwich considérée par M. Fay et B. Graubard [**10**] et destinée à corriger à la hausse l'estimateur sandwich qui tend à sous-estimer la vraie valeur de la variance. Un second facteur de correction basé sur [**36**] a été défini et appelé correction conservative . À cette liste d'estimateurs de la variance, nous avons ajouter le Jackknife, une technique non-paramétrique classique.

Dans une étude par simulation, nous avons observé ces estimateurs sous plusieurs conditions et comparé leur biais relatif Monte-Carlo et le taux de couverture des intervalles de confiance associés. La conclusion générale est que le Jackknife donne des taux de couverture équivalents ou supérieurs aux autres estimateurs dans toutes les conditions.

Il est à noter qu'à l'exception de l'étalon-or et du Jackknife, tous les estimateurs de la variance que nous avons considérés ne sont valides que lorsque le score de propension ou l'espérance conditionnelle de l'issue est estimé paramétriquement. Certains progrès ont été accomplis récemment dans le domaine de l'estimation non-paramétrique de la variance du TMLE [**14**] mais c'est un domaine de recherche actif.

Annexe A

Notations

- p_X représente la fonction de densité de la variable aléatoire X si celle-ci est continue et la fonction de masse si X est discrète. Pour la densité conditionelle de X sachant $Y = y$ évaluée en x, on écrit $p_{X|Y}(x|y)$.
- Si v est un vecteur colonne, on écrit $v^{\otimes 2}$ au lieu de $vv^\top$. Cette notation est en accord avec le produit de Kronecker de matrices.
- L'indépendance entre variables aléatoires s'exprime par le symbole $\perp\!\!\!\perp$.
- La convergence en probabilité et la convergence en distribution sont notés respectivement $\xrightarrow{p}$ et $\xrightarrow{d}$. Si $X_n \xrightarrow{p} X$, on écrit aussi $\operatorname{plim}_n X_n = X$.
- Si X_n est une suite de variables aléatoires, nous écrirons $X_n = o_p(n^q)$ pour exprimer le fait que $n^{-q} X_n \xrightarrow{p} 0$.
- Pour une fonction $f(x_1, \ldots, x_n)$ et $x_0 \in \mathbb{R}^n$ nous écrirons $\frac{\partial f(x_0)}{\partial x_j}$ au lieu de $\left.\frac{\partial f(x_1,\ldots,x_n)}{\partial x_j}\right|_{(x_1,\ldots,x_n)=x_0}$.
- Nous identifions les éléments de $\mathbb{R}^n$ aux vecteurs colonnes. Ainsi, si par exemple $f(x_1, \ldots, x_N)$ est une fonction différentiable, alors $\frac{\partial f}{\partial x}$ représente le vecteur colonne

$$\begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{pmatrix}.$$

  Nous écrirons $\frac{\partial f}{\partial x^\top}$ pour dénoter le vecteur ligne $\left(\frac{\partial f}{\partial x}\right)^\top$.

Annexe B

Tableaux

Tableau B.1. Moyenne des estimations de la variance en fonction de la taille échantillonnale n (issue dichotomique)

n	TMLE moyen	V_{MC}	Étalon-or	Sandwich	Corr. F-G	Corr. Cons.	Jackknife
30	0,0508	0,0602	0,0261	0,0584	0,0673	0,0955	0,1903
40	0,0553	0,0356	0,021	0,0223	0,0229	0,0236	0,0718
50	0,0556	0,0251	0,0175	0,0177	0,018	0,0184	0,0413
60	0,0584	0,0195	0,0149	0,0149	0,0151	0,0154	0,0274
70	0,0581	0,0159	0,013	0,013	0,0132	0,0134	0,0208
80	0,0587	0,0136	0,0115	0,0115	0,0117	0,0118	0,017
90	0,0604	0,0119	0,0103	0,0103	0,0104	0,0105	0,0139
100	0,0594	0,0106	0,0093	0,0093	0,0094	0,0095	0,0121
110	0,0608	0,0092	0,0085	0,0085	0,0086	0,0086	0,0107
120	0,0608	0,0086	0,0078	0,0078	0,0079	0,0079	0,0095
130	0,0609	0,0079	0,0073	0,0073	0,0073	0,0074	0,0086
140	0,0616	0,0073	0,0067	0,0067	0,0068	0,0068	0,0079

Tableau B.2. Moyenne des estimations de la variance en fonction de la taille échantillonnale n (issue continue)

n	TMLE moyen	V_{MC}	Étalon-or	Sandwich	Corr. F-G	Corr. Cons.	Jackknife
30	2,0031	0,888	0,1847	0,2019	0,2193	0,2723	5,5866
40	1,9994	0,2899	0,1191	0,121	0,1245	0,1301	1,1854
50	2,0003	0,1722	0,0932	0,0933	0,0953	0,0972	0,4848
60	1,9983	0,1164	0,0785	0,0823	0,0837	0,0851	0,2346
70	1,9953	0,0936	0,068	0,0678	0,0687	0,0697	0,1448
80	1,9972	0,0782	0,0607	0,0605	0,0613	0,0621	0,1167
90	1,9991	0,0672	0,0546	0,0544	0,055	0,0556	0,0882
100	1,9986	0,0585	0,0493	0,0492	0,0497	0,0502	0,0724
110	2,0025	0,0535	0,045	0,0449	0,0453	0,0457	0,0634
120	1,9996	0,0465	0,0415	0,0414	0,0418	0,0421	0,0554
130	1,9992	0,0438	0,0387	0,0386	0,0389	0,0392	0,0499
140	2,0006	0,0402	0,0361	0,036	0,0363	0,0366	0,0457

Tableau B.3. Moyenne des estimations de la variance en fonction du paramètre z (issue dichotomique)

z	TMLE moyen	V_{MC}	Étalon-or	Sandwich	Corr. F-G	Corr. Cons.	Jackknife
-3	0,3409	0,0026	0,0022	0,0022	0,0022	0,0022	0,003
-2	0,3412	0,0011	0,0011	0,0011	0,0011	0,0011	0,0012
-1	0,3413	0,0007	0,0007	0,0007	0,0007	0,0007	0,0007
0	0,3417	0,0007	0,0007	0,0007	0,0007	0,0007	0,0007
1	0,3419	0,0012	0,0011	0,0011	0,0011	0,0011	0,0012
2	0,3416	0,0025	0,0024	0,0024	0,0024	0,0024	0,0027
3	0,3412	0,0067	0,0053	0,0053	0,0054	0,0054	0,0078

Tableau B.4. Moyenne des estimations de la variance en fonction du paramètre z (issue continue)

z	TMLE moyen	V_{MC}	Étalon-or	Sandwich	Corr. F-G	Corr. Cons.	Jackknife
-3	1,9985	0,0363	0,0278	0,0278	0,0278	0,0278	0,0433
-2	1,9999	0,0142	0,0132	0,0132	0,0132	0,0132	0,0156
-1	1,9995	0,0072	0,0069	0,0069	0,0069	0,0069	0,0073
0	2,0005	0,0054	0,0053	0,0053	0,0053	0,0053	0,0054
1	1,9996	0,0071	0,0069	0,0069	0,0069	0,0069	0,0073
2	1,9989	0,0144	0,0131	0,0131	0,0131	0,0131	0,0158

Tableau B.5. Moyenne des estimations de la variance en fonction du scénario (issue dichotomique)

Scénario	TMLE moyen	V_{MC}	Étalon-or	Sandwich	Corr. F-G	Corr. Cons.	Jackknife
1	0,328	0,0034	0,0032	0,0032	0,0032	0,0032	0,0037
2	0,3285	0,0039	0,0038	0,0038	0,0038	0,0038	0,004
3	0,3287	0,0039	0,004	0,004	0,004	0,0041	0,004
4	0,3283	0,0038	0,0038	0,0038	0,0038	0,0038	0,004
5	0,3278	0,0033	0,0031	0,0031	0,0031	0,0031	0,0035
6	0,3287	0,004	0,0041	0,0041	0,0041	0,0042	0,0042

Tableau B.6. Moyenne des estimations de la variance en fonction du scénario (issue continue)

Scénario	TMLE moyen	V_{MC}	Étalon-or	Sandwich	Corr. F-G	Corr. Cons.	Jackknife
1	2,0003	0,0219	0,0206	0,0206	0,0207	0,0208	0,0236
2	2,0004	0,4191	0,4085	0,4084	0,4104	0,4125	0,4329
3	2,0029	0,4275	1,52	1,5902	1,5991	1,608	0,4484
4	2,0003	0,4166	0,3925	0,3925	0,3945	0,3965	0,4288
5	2,0002	0,0214	0,0193	0,0193	0,0194	0,0195	0,0222
6	2,0043	0,4487	1,5523	3,1482	3,1788	3,21	0,487

Bibliographie

[1] S. Amari, Geometrical theory of asymptotic ancillarity and conditional inference, *Biometrika* **69**, 1-17 (1982).

[2] S. Amari, Differential geometry of curved exponential families – curvature and information loss, *Annals of Statistics* **10**, 357-387 (1982).

[3] S. Amari, *Differential Geometric Methods in Statistics*, Lecture notes in statistics **28**, Berlin, Springer-Verlag (1985).

[4] H. Brezis, *Analyse Fonctionelle – Théorie et application*, Sciences Sup, Dunod (2005).

[5] D. R. Cox, *Planning of Experiments*, New York: John Wiley & Sons (1958).

[6] J. K. Lunceford, M. Davidian, Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study, *Statistics in Medicine* **23**, 2937-2960 (2004).

[7] S. Yuan, H. H. Zhang, M. Davidian, Variable selection for covariate-adjusted semiparametric inference in randomized clinical trials, *Statistics in Medicine* **31**, 3789-3804 (2012).

[8] B. Efron, Defining the curvature of a statistical problem (with applications to second order efficiency), *Annals of Statistics*, **3**, 1189-1242 (1975).

[9] B. Efron, *The jackknife, the bootstrap, and other resampling plans*, Philadelphia, PA: Society for Industrial and Applied Mathematics (1982).

[10] M. P. Fay, B. I. Graubard, Small-sample adjustments for Wald-type tests using sandwich estimators, *Biometrics* **57**, 1198-1206 (2001).

[11] B. Cossette, A. Forget, M. Beauchesne et al., Impact of maternal use of asthma-controller therapy on perinatal outcomes, *Thorax* **68**, 724-730 (2013).

[12] S. Eltonsy, M. Beauchesne, L. Blais, Risk of congenital malformations for asthmatic pregnant women using a long-acting beta(2)-agonist and inhaled corticosteroid combination versus higher-dose inhaled corticosteroid monotherapy, *Journal of Allergy and Clinical Immunology* **135**, 123-130 (2015).

[13] S. Grüber, M. J. Van der Laan, A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome, *The International Journal of Biostatistics* **6**, Article 26 (2010).

[14] S. Grüber, *Collaborative Targeted Maximum Likelihood Estimation*, PhD thesis (2011).

[15] National asthma education and prevention program. Expert panel report 3 (EPR-3): Guidelines for the diagnosis and management of asthma-summary report 2007, *Journal of Allergy and Clinical Immunology* **120** (5 Suppl), S94-138 (2007).

[16] M. A. Hernán, S.L. Taubman, Does obesity shorten life? The importance of well-defined interventions to answer causal questions, *International Journal of Obesity* **32**, S8-S14 (2008).

[17] M. A. Hernán, A definition of causal effect for epidemiological research, *Journal of Epidemiology & Community Health* **58**, 265-271 (2004).

[18] D. G. Horvitz, D. J. Thompson, A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**, 663-685 (1952).

[19] G. Kauermann, R. J. Carroll, A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **96**, 1387-1396 (2001).

[20] E. H. Kennedy, Semiparametric theory and empirical processes in causal inference, arXiv:1510.04740.

[21] J. M. Lee, *Introduction to Smooth Manifolds*, Springer; 2nd edition (2013).

[22] M. Delgado-Rodriguez, Javier Llorca, Bias, *Journal of Epidemiology Community Health*, **58**, 635-614 (2004).

[23] J. Neyman, On the application of probability theory to agricultural experiments: Essay on principles. Section 9. (Translated and edited by D. M. Dabrowska and T. P. Speed), *Statistical Science* **5**, 465-472 (1990).

[24] M. H. Quenouilles, Notes on bias in estimation, *Biometrika* **43**, 353-360 (1956).

[25] C. R. Rao, Information and accuracy attainable in the estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society* **37**, 81-91.

[26] J. M. Robins, A new approach to causal inference in mortality studies with sustained exposure period - application to control of the healthy worker survivor effect, *Mathematical Modelling* **7**, 1393-1512 (1986).

[27] S. Greenland, J. Pearl, J. M. Robins, Causal diagrams for epidemiologic research, *Epidemiology* **10**, 37-48 (1999).

[28] D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688-701 (1974).

[29] P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41-55 (1983).

[30] D. B. Rubin, Comment on: "Randomization analysis of experimental data in the fisher randomization test" by D. Basu., *Journal of the American Statistical Association* **75**, 591-593 (1980).

[31] J. S. Sheffield, D. Siegel, M. Mirochnick,R. P. Heine, C. Nguyen, K. L. Bergman, R. M. Savic, J. Long, K. E. Dooley, M. Nesin, Designing drug trials: considerations for pregnant women, *Clinical Infectious Diseases* **59**, S437-S444 (2014).

[32] A. J. Stephens, E. J. Tchetgen Tchetgen, V. De Gruttola, Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level and individual-level covariates, *Statistics in Medicine* **31**, 915-930 (2012).

[33] E. H. Simpson, The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society. Series B (Methodological)* **2**, 238-241 (1951).

[34] P. D. Stolley, When genius errs: R. A. Fisher and the lung cancer controversy, *American Journal of Epidemiology* **133**, 416-425 (1991).

[35] A. A. Tsiatis, *Semiparametric Theory and Missing Data*, Springer, New York (2006).

[36] R. Valliant, A. H. Dorfman, R. M. Royall, *Finite Population Sampling and Inference, a Prediction Approach*, Wiley Inter-Science (2000).

[37] J. P. Vandenbroucke, A. Broadbent, N. Pearce, Causality and causal inference in epidemiology: the need for a pluralistic approach, *International Journal of Epidemiology* **45**, 1776-1786 (2016).

[38] T. J. VanderWeele, W. R. Robinson, On causal interpretation of race in regressions adjusting for confounding and mediating variables, *Epidemiology* **25**, 473-484 (2014).

[39] T. J. VanderWeele, Concerning the consistency assumption in causal inference, *Epidemiology* **20**, 880-883 (2009).

[40] T. J. VanderWeele, M. A. Hernán, Causal inference under multiple versions of treatment, *Journal of Causal Inference* **1**, 1-20 (2013).

[41] T. J. VanderWeele, Commentary: On causes, causal inference, and potential outcomes, *International Journal of Epidemiology* **45**, 1809-1816 (2016).

[42] T. J. VanderWeele, M. A. Hernán, *Causal effects and natural laws: Towards a conceptualization of causal counterfactuals for nonmanipulable exposures, with application to the effects of race and sex, In: Berzuini C, Dawid A, Bernardinelli L. (eds). Causality: Statistical Perspective and Applications.* John Wiley, 101-113 2012.

[43] S. Rose, M. van der Laan, *Targeted Learning: Causal inference for observational and experimental data*, Springer-Verlag (2011).

[44] M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, M. J. van der Laan, Diagnosing and responding to violations in the positivity assumption, *Statistical Methods in Medical Research* **21**, 31-54 (2010).

[45] M. J. van der Laan, D. Rubin, Targeted maximum likelihood learning, *U.C. Berkeley Division of Biostatistics Working Paper Series*, working paper 213 (2006).