Metrika
https://doi.org/10.1007/s00184-024-00962-4

Variance estimation for average treatment effects estimated by g-computation

Stefan Nygaard Hansen[1] · **Morten Overgaard**[1]

Received: 3 February 2023 / Accepted: 8 March 2024

Abstract
The average treatment effect is used to evaluate effects of interventions in a population. Under certain causal assumptions, such an effect may be estimated from observational data using the g-computation technique. The asymptotic properties of this estimator appears not to be well-known and hence bootstrapping has become the preferred method for estimating its variance. Bootstrapping is, however, not an optimal choice for multiple reasons; it is a slow procedure and, if based on too few bootstrap samples, results in a highly variable estimator of the variance. In this paper, we consider estimators of potential outcome means and average treatment effects using g-computation. We consider these parameters for the entire population but also in subgroups, for example, the average treatment effect among the treated. We derive their asymptotic distributions in a general framework. An estimator of the asymptotic variance is proposed and shown to be consistent when g-computation is used in conjunction with the M-estimation technique. The proposed estimator is shown to be superior to the bootstrap technique in a simulation study. Robustness against model misspecification is also demonstrated by means of simulations.

Keywords Average treatment effect · Causal inference · G-computation · Model misspecification · Variance estimation

1 Introduction

Many research questions involve quantifying the effect of intervening on some variable on an outcome. For example, policymakers might be interested in the effect of introducing a new treatment or new guidelines on certain health outcomes. This is

✉ Stefan Nygaard Hansen
stefanh@ph.au.dk

Morten Overgaard
moov@ph.au.dk

[1] Department of Public Health, Aarhus University, Bartholins Allé 2, 8000 Aarhus C, Denmark

Published online: 06 April 2024

usually quantified as an average effect over the entire population but in some situations it might also be useful to know if there are subgroups for which the intervention is particularly beneficial. The literature has focused much on assumptions needed to identify and estimate such effects based on observational data whereas the statistical properties of these estimators has had limited attention.

If Y denotes the outcome of interest, A the covariate we are interested in intervening on and $\mathbf{B}$ any other quantity we wish to include in our statistical model, for reasons to become clear in a moment, then we may quantify such an effect in two steps. First, we assume that the relationship between Y and $\mathbf{X} = (A, \mathbf{B})$ can be described by a conditional mean model

$$\mathrm{E}(Y \mid \mathbf{X}) = \mu(\boldsymbol{\beta}_0; \mathbf{X}) \tag{1}$$

for some mean function μ and true vector of parameters $\boldsymbol{\beta}_0$. Next, if $\mathbf{X}^a = (a, \mathbf{B})$ denotes the covariate vector obtained by replacing A by the fixed value a, then the *predictive margin* of the intervention a is defined as $\mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^a))$. Assuming (1) holds true, this quantity equals $\mathrm{E}(\mathrm{E}(Y \mid A = a, \mathbf{B}))$ justifying its name. Similarly, the *predictive contrast* between two interventions a_1 and a_2 is defined as $\mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_1}) - \mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_2}))$.

The predictive margin appears to quantify what would have happened in the underlying population if we had, contrary to fact, intervened on A and set it to a for all members of the population. However, such a causal interpretation is only warranted under further assumptions. One such assumption is exchangeability (Hernán and Robins 2020), also called (strong) ignorability (Rosenbaum and Rubin 1983). Assuming exchangeability conditional on $\mathbf{B}$, the predictive margin corresponds to the mean of a potential outcome variable after intervention and the predictive contrast corresponds to the average treatment effect comparing two interventions. Restricting these to subgroups may also be relevant, e.g. when the interest lies in the average treatment effect among the treated or untreated (Wang et al. 2017). In this context, the model in (1) is sometimes referred to as the Q-model and the estimating procedure described below is also known as (parametric) g-computation or standardization (Robins 1986; Snowden et al. 2011; Wang et al. 2017; Hernán and Robins 2020).

Standardization or g-computation can, in this situation, be described as a two-step estimating procedure: Obtain a reasonable estimate $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}_0$ and then estimate the potential outcome mean or average treatment effect by an appropriate sample average in which $\boldsymbol{\beta}_0$ has been replaced by $\hat{\boldsymbol{\beta}}_n$. Bootstrapping has become the predominant method for obtaining valid confidence intervals and statistical tests for these estimators (Snowden et al. 2011; Wang et al. 2017; Keil et al. 2014; Westreich et al. 2012; Nianogo et al. 2017; Chatton et al. 2020). However, as bootstrapping quickly becomes computationally intensive and, as we will demonstrate, may result in a highly variable estimator when the number of bootstrap samples are too few, alternative methods for estimating the variance seem valuable.

The asymptotic properties of these estimators have been studied in Dowd et al. (2014) and Qu and Luo (2015). In both papers, the authors disregard the sampling variation in $\mathbf{X}$ which has later been argued to be incorrect if causal parameters such

as the average treatment effect is the target (Terza 2016; Bartlett 2018). In both Terza (2016) and Bartlett (2018), these results have been extended to account for the sampling variability in $\mathbf{X}$ although the latter focus mostly on randomised trials. These properties have also been studied to some extent by Graubard and Korn (1999). Terza (2016) relies on results by Newey and McFadden (1994) but, in their derivation, they do however make an unnecessary simplification of the asymptotic variance that relies on a correctly specified conditional mean model. Bartlett (2018) also relies on Newey and McFadden (1994) and in their paper they discuss how the variance estimator proposed by Qu and Luo (2015) can be remedied by adding an extra term to their expression. The resulting expression coincides with that of Terza (2016). Therefore, the variance estimator proposed by Terza (2016) and Bartlett (2018) relies on a correctly specified conditional mean model and Terza (2016) argued that this is required for the analysis to be of interest. However, it appears that the results of Terza (2016) and Bartlett (2018) have been overlooked in the causal inference literature as is evident by the plethora of studies that use bootstrapping but also by the lack of availability in software. For example, a recent `R` package called `marginaleffects` (Arel-Bundock 2022) has implemented the variance estimator proposed by Dowd et al. (2014) and Qu and Luo (2015). Another `R` package called `riskCommunicator` relies entirely on bootstrapping (Grembi and McQuade 2022). Moreover, Terza (2016) and Bartlett (2018) only focus on estimating potential outcome means and average treatment effects in the entire population and, hence, they do not consider effects in subgroups nor do they consider relative treatment effects.

In this paper, we derive the asymptotic distribution of the g-computation estimator for potential outcome means in the entire population but also among subgroups. The asymptotic distribution is presented in a multivariate setting where multiple interventions are considered simultaneously which enables us to easily derive the asymptotic distribution of effect estimates such as the average treatment effect in the entire population and in subgroups. Relative treatment effects, obtained as ratios of two potential outcome means, are also considered. This is useful when one wants to express the treatment effect as a risk ratio or odds ratio, say (Rubin 2010). Based on these asymptotic properties, we propose an estimator for the asymptotic variance for the various estimators considered. We discuss when these variance estimators are consistent and argue that consistency is guaranteed when a type of M-estimation is used in the first step to estimate $\boldsymbol{\beta}_0$. The asymptotic variance for the average treatment effect derived here will have an extra term compared to that of Terza (2016) and Bartlett (2018) which, we argue, makes it robust against model misspecification. We compare coverage and variability of the proposed variance estimator against bootstrapping in a simulation study. We also compare the proposed variance estimator to that of Terza (2016) and Bartlett (2018) in a scenario where the conditional mean model is misspecified.

2 Potential outcome means

As above, we let Y denote the outcome of interest, A the variable we are interested in intervening on and $\mathbf{B}$ one or more variables we wish to include in our statistical model. We call A the treatment although A could, in principle, be any exposure variable

whose effect we might want to study. Let $Y(a)$ denote the outcome had the individual in question received treatment a. This is a potential outcome as, for a given individual, we are only able to observe one of them – all other variables are counterfactual. Let also $Y = Y(A)$ denote the observed outcome. This is often referred to as the consistency assumption. We denote by $\mathbf{X} = (A, \mathbf{B})$ the observed covariates and define $\mathbf{X}^a = (a, \mathbf{B})$ to be the covariates obtained by replacing A by the fixed value a. We also denote by $\mathbf{Z} = (Y, \mathbf{X})$ the observed outcome and covariates collectively.

In the following, we will work under the assumption of exchangeability conditional on $\mathbf{B}$. This assumption states that $Y(a)$ should be independent of A given $\mathbf{B}$ for all a. Under this assumption we have that

$$\begin{aligned} \mathrm{E}(Y(a)) &= \mathrm{E}(\mathrm{E}(Y(a) \mid \mathbf{B})) = \mathrm{E}(\mathrm{E}(Y(a) \mid A = a, \mathbf{B})) \\ &= \mathrm{E}(\mathrm{E}(Y \mid A = a, \mathbf{B})) = \mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^a)) \end{aligned} \tag{2}$$

where we have used that $Y\mathbf{1}(A = a) = Y(a)\mathbf{1}(A = a)$ by definition of Y.

By this chain of equalities, we conclude the following: The right-hand side, $\mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^a))$, corresponds to a predictive margin only when the conditional mean model is correctly specified and to a potential outcome mean only when, in addition, exchangeability holds. In the following, our focus will be on estimating the right-hand side and hence all of the results established in this paper holds even when the conditional mean model is misspecified although, in this case, the target parameter is simply $\mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^a))$ where $\boldsymbol{\beta}_0$ provides some sort of best fit of $\mu(\boldsymbol{\beta}; \mathbf{X})$ to $\mathrm{E}(Y \mid \mathbf{X})$. A similar argument as in (2) shows that $\mathrm{E}(Y(a) \mid \mathbf{V} = \mathbf{v}) = \mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^a) \mid \mathbf{V} = \mathbf{v})$ when $\mathbf{V}$ is a function of $\mathbf{X}$, meaning that we also have a causal interpretation when restricting the analysis to subgroups of the population.

Consider now the conditional mean model of (1) in which $\boldsymbol{\beta}_0$ is assumed to be a k-dimensional column vector. We assume throughout the paper that μ satisfies the following regularity conditions

- $\mathbf{x} \mapsto \mu(\boldsymbol{\beta}; \mathbf{x})$ is measurable for all $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$.
- $\boldsymbol{\beta} \mapsto \mu(\boldsymbol{\beta}; \mathbf{x})$ is continuously differentiable for all $\mathbf{x}$.

Consider also an i.i.d. sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ of $\mathbf{Z}$ and assume that, based on this sample, an estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}_0$ exists that is asymptotically linear in the sense that

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + \frac{1}{n} \sum_{i=1}^{n} \dot{\boldsymbol{\beta}}(\mathbf{Z}_i) + o_P(n^{-1/2}), \tag{3}$$

with an influence function $\dot{\boldsymbol{\beta}}$ such that $\dot{\boldsymbol{\beta}}(\mathbf{Z})$ is square integrable and $\mathrm{E}(\dot{\boldsymbol{\beta}}(\mathbf{Z})) = \mathbf{0}$. We discuss the existence of such an estimator in greater detail in Sect. 4.

Let a be a fixed intervention value and recall that $\mathbf{X}^a = (a, \mathbf{B})$ is the covariate vector obtained by replacing A with a. Consider now the potential outcome mean upon replacing A with a, $\theta^a = \mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^a))$, as our target parameter. It seems

natural to estimate this by

$$\hat{\theta}_n^a = \frac{1}{n}\sum_{i=1}^{n} \mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^a). \tag{4}$$

The procedure leading to the estimator in (4) is known as g-computation.

In the following, we say that a function $\theta \mapsto g(\theta; \mathbf{x})$ is dominated (square) integrable with respect to $\mathbf{X}$ around some θ_0 if there exists a function h such that $h(\mathbf{X})$ is (square) integrable and such that $\|g(\theta; \mathbf{x})\| \leq h(\mathbf{x})$ for all $\mathbf{x}$ and for θ in some compact set around θ_0. By a first-order Taylor expansion of μ we now obtain the following result.

Proposition 1 *Assume that $\mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated square integrable and that $\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated integrable with respect to $\mathbf{X}^a$ around $\boldsymbol{\beta}_0$. Then $\hat{\theta}_n^a$ is asymptotically linear with influence function*

$$\dot{\theta}^a(\mathbf{z}) = \mu(\boldsymbol{\beta}_0; \mathbf{x}^a) - \theta^a + \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}^a)\right)\dot{\boldsymbol{\beta}}(\mathbf{z}), \tag{5}$$

for $\mathbf{z} = (y, \mathbf{x})$. That is, for each n, we have the decomposition

$$\hat{\theta}_n^a = \theta^a + \frac{1}{n}\sum_{i=1}^{n} \dot{\theta}^a(\mathbf{Z}_i) + o_P(n^{-1/2}) \tag{6}$$

where $\dot{\theta}^a(\mathbf{Z})$ is square integrable with mean zero.

Proof The proof is given in the Appendix. □

With this decomposition at hand, the asymptotic distribution of $\hat{\theta}_n^a$ follows immediately. The following result is formulated in a multivariate setting as it allows us to study estimators obtained as functions of multiple potential outcome means. For a vector $\mathbf{a} = (a_1, \ldots, a_l)$ of interventions we let $\hat{\boldsymbol{\theta}}_n^{\mathbf{a}} = (\hat{\theta}_n^{a_1}, \ldots, \hat{\theta}_n^{a_l})^T$ denote the vector of estimators and $\boldsymbol{\theta}^{\mathbf{a}} = (\theta^{a_1}, \ldots, \theta^{a_l})^T$ the vector of target parameters. Furthermore, we let $\boldsymbol{\mu}(\boldsymbol{\beta}; \mathbf{X}^{\mathbf{a}})$ denote the l-dimensional column vector whose ith entry is $\mu(\boldsymbol{\beta}; \mathbf{X}^{a_i})$ and let $\mathrm{E}(\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{\mu}(\boldsymbol{\beta}; \mathbf{X}^{\mathbf{a}}))$ denote the $l \times k$ matrix whose ith row is $\mathrm{E}(\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}; \mathbf{X}^{a_i}))$. We then have the following result.

Theorem 2 *Let $\mathbf{a} = (a_1, \ldots, a_l)$ be a vector of interventions. Assume that $\mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated square integrable and that $\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated integrable with respect to $\mathbf{X}^{a_i}$ around $\boldsymbol{\beta}_0$ for $i = 1, \ldots, l$. Then $\hat{\boldsymbol{\theta}}_n^{\mathbf{a}} \xrightarrow{\mathrm{P}} \boldsymbol{\theta}^{\mathbf{a}}$ and*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{\mathbf{a}} - \boldsymbol{\theta}^{\mathbf{a}}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma}^{\mathbf{a}}) \tag{7}$$

for $n \to \infty$ with covariance matrix given by

$$\boldsymbol{\Gamma}^{\mathbf{a}} = \mathrm{Cov}\left(\boldsymbol{\mu}(\boldsymbol{\beta}_0; \mathbf{X}^{\mathbf{a}}) + \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu}(\boldsymbol{\beta}_0; \mathbf{X}^a)\right) \dot{\boldsymbol{\beta}}(\mathbf{Z})\right).$$

Proof Let $\dot{\boldsymbol{\theta}}^{\mathbf{a}} = (\dot{\theta}^{a_1}, \ldots, \dot{\theta}^{a_l})^T$ denote the vector of influence functions. By Proposition 1 we have

$$\hat{\boldsymbol{\theta}}^{\mathbf{a}}_n - \boldsymbol{\theta}^{\mathbf{a}} = \frac{1}{n} \sum_{i=1}^{n} \dot{\boldsymbol{\theta}}^{\mathbf{a}}(\mathbf{Z}_i) + o_P(n^{-1/2})$$

where the right-hand side is seen to converge to zero in probability by the law of large numbers as $\dot{\boldsymbol{\theta}}^{\mathbf{a}}(\mathbf{Z})$ has mean zero.

Moreover,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathbf{a}}_n - \boldsymbol{\theta}^{\mathbf{a}}) = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \dot{\boldsymbol{\theta}}^{\mathbf{a}}(\mathbf{Z}_i) + o_P(1)$$

where the right-hand side is seen to converge to a normal distribution with mean zero and covariance matrix $\boldsymbol{\Gamma}^{\mathbf{a}}$ by the central limit theorem. □

Assume now that an estimator $\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{z})$ of $\dot{\boldsymbol{\beta}}(\mathbf{z})$ exists for all $\mathbf{z}$. The asymptotic covariance matrix of Theorem 2 may then be estimated by the following plug-in estimator

$$\hat{\boldsymbol{\Gamma}}^{\mathbf{a}}_n = \frac{1}{n} \sum_{i=1}^{n} \left\{ \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_n; \mathbf{X}^{\mathbf{a}}_i) - \hat{\boldsymbol{\theta}}^{\mathbf{a}}_n + \left(\frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_n; \mathbf{X}^{\mathbf{a}}_j) \right) \hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i) \right\}^{\otimes 2} \tag{8}$$

where $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}^T$ for a column vector $\mathbf{x}$.

Under some mild regularity conditions on the estimator $\hat{\dot{\boldsymbol{\beta}}}_n$, this plug-in estimator will be consistent for the asymptotic covariance matrix as the following result shows.

Theorem 3 *Make the assumptions of Theorem 2 and assume furthermore that $\hat{\dot{\boldsymbol{\beta}}}_n$ satisfies*

$$\|\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{z}) - \dot{\boldsymbol{\beta}}(\mathbf{z})\| \leq g_n \cdot f(\mathbf{z}) \tag{9}$$

for a sequence of random variables $g_n \xrightarrow{\mathrm{P}} 0$ and a measurable function f with $\mathrm{E}(f(\mathbf{Z})^2) < \infty$. Then $\hat{\boldsymbol{\Gamma}}^{\mathbf{a}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\Gamma}^{\mathbf{a}}$.

Proof See the Appendix. □

The assumption (9) is just one example of an assumption which ensures consistency of the plug-in estimator of (8). Other assumptions might also do the job but this particular assumption can be seen to hold in the examples discussed in Sect. 4.

2.1 Potential outcome means in a subgroup

Sometimes we are interested in the mean of the potential outcome, not for the entire population, but only among individuals in a certain subgroup. Let $\mathbf{V} = h(\mathbf{X})$ be a partition of $\mathbf{X}$ and consider a subgroup $\mathbf{V} = \mathbf{v}$ with $P(\mathbf{V} = \mathbf{v}) > 0$. Define also $\psi^{a,\mathbf{v}} = \mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^a) \mid \mathbf{V} = \mathbf{v})$ to be the potential outcome mean corresponding to the intervention a among those with covariate vector $\mathbf{V}$ equal to $\mathbf{v}$.

Let $m = \#\{i : \mathbf{V}_i = \mathbf{v}\}$ be the number of observations compatible with $\mathbf{V} = \mathbf{v}$ in our sample. A natural estimator for $\psi^{a,\mathbf{v}}$ is then

$$\hat{\psi}_n^{a,\mathbf{v}} = \frac{1}{m} \sum_{i:\mathbf{V}_i = \mathbf{v}} \mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^a). \tag{10}$$

Similar to above we argue that this estimator is asymptotically linear.

Proposition 4 *Assume that $\mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated square integrable and that $\frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated integrable with respect to $\mathbf{X}^a$ around $\boldsymbol{\beta}_0$. Then $\hat{\psi}_n^{a,\mathbf{v}}$ is asymptotically linear with influence function*

$$\dot{\psi}^{a,\mathbf{v}}(\mathbf{z}) = (\mu(\boldsymbol{\beta}_0; \mathbf{x}^a) - \psi^{a,\mathbf{v}}) \frac{\mathbf{1}(h(\mathbf{x}) = \mathbf{v})}{P(\mathbf{V} = \mathbf{v})} + \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0; \mathbf{x}^a) \mid \mathbf{V} = \mathbf{v} \right) \dot{\boldsymbol{\beta}}(\mathbf{z}), \tag{11}$$

for $\mathbf{z} = (y, \mathbf{x})$. That is, for each n, we have the decomposition

$$\hat{\psi}_n^{a,\mathbf{v}} = \psi^{a,\mathbf{v}} + \frac{1}{n} \sum_{i=1}^{n} \dot{\psi}^{a,\mathbf{v}}(\mathbf{Z}_i) + o_P(n^{-1/2}) \tag{12}$$

where $\dot{\psi}^{a,\mathbf{v}}(\mathbf{Z})$ is square integrable with mean zero.

Arguments similar to those above yield the following asymptotic result.

Theorem 5 *Let $\mathbf{a} = (a_1, \ldots, a_l)$ be a vector of interventions. Assume that $\mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated square integrable and that $\frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated integrable with respect to $\mathbf{X}^{a_i}$ around $\boldsymbol{\beta}_0$ for $i = 1, \ldots, l$. Then $\hat{\boldsymbol{\psi}}_n^{\mathbf{a},\mathbf{v}} \xrightarrow{\mathrm{P}} \boldsymbol{\psi}^{\mathbf{a},\mathbf{v}}$ and*

$$\sqrt{n}(\hat{\boldsymbol{\psi}}_n^{\mathbf{a},\mathbf{v}} - \boldsymbol{\psi}^{\mathbf{a},\mathbf{v}}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Lambda}^{\mathbf{a},\mathbf{v}}) \tag{13}$$

for $n \to \infty$ with

$$\boldsymbol{\Lambda}^{\mathbf{a},\mathbf{v}} = \mathrm{Cov}\left(\left(\boldsymbol{\mu}(\boldsymbol{\beta}_0; \mathbf{X}^{\mathbf{a}}) - \boldsymbol{\psi}^{\mathbf{a},\mathbf{v}} \right) \frac{\mathbf{1}(\mathbf{V} = \mathbf{v})}{P(\mathbf{V} = \mathbf{v})} + \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu}(\boldsymbol{\beta}_0; \mathbf{X}^a) \mid \mathbf{V} = \mathbf{v} \right) \dot{\boldsymbol{\beta}}(\mathbf{Z}) \right)$$

The asymptotic covariance matrix $\mathbf{\Lambda}^{\mathbf{a},\mathbf{v}}$ may be estimated by

$$\hat{\mathbf{\Lambda}}_n^{\mathbf{a},\mathbf{v}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_n;\mathbf{X}_i^{\mathbf{a}}) - \hat{\boldsymbol{\psi}}_n^{\mathbf{a},\mathbf{v}}\right)\frac{\mathbf{1}(\mathbf{V}_i=\mathbf{v})}{m/n} + \left(\frac{1}{m}\sum_{j:\mathbf{V}_j=\mathbf{v}}\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_n;\mathbf{X}_j^{\mathbf{a}})\right)\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i)\right\}^{\otimes 2} \tag{14}$$

for which we have the following result.

Theorem 6 *Make the assumptions of Theorem 5 and assume furthermore that* $\hat{\dot{\boldsymbol{\beta}}}_n$ *satisfies*

$$\|\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{z}) - \dot{\boldsymbol{\beta}}(\mathbf{z})\| \leq g_n \cdot f(\mathbf{z})$$

for a sequence of random variables $g_n \xrightarrow{\mathrm{P}} 0$ *and a measurable function* f *with* $\mathrm{E}(f(\mathbf{Z})^2) < \infty$. *Then* $\hat{\mathbf{\Lambda}}_n^{\mathbf{a},\mathbf{v}} \xrightarrow{\mathrm{P}} \mathbf{\Lambda}^{\mathbf{a},\mathbf{v}}$.

3 Average treatment effects

Often we are not explicitly interested in the mean of a potential outcome but more so in a contrast between such two: the average treatment effect. Let a_1, a_2 be two interventions and define the average treatment effect between such two interventions by

$$\mathrm{ATE}(a_1,a_2) = \mathrm{E}\Big(\mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_1}) - \mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_2})\Big) = \theta^{a_1} - \theta^{a_2}$$

which we estimate by $\widehat{\mathrm{ATE}}_n(a_1,a_2) = \hat{\theta}_n^{a_1} - \hat{\theta}_n^{a_2}$. The asymptotic distribution of this estimator now follows immediately from Theorem 2 and an application of the delta method.

Corollary 7 *Let* a_1, a_2 *be two interventions such that* $\mu(\boldsymbol{\beta};\mathbf{x})$ *is dominated square integrable and that* $\frac{\partial}{\partial\boldsymbol{\beta}}\mu(\boldsymbol{\beta};\mathbf{x})$ *is dominated integrable with respect to* $\mathbf{X}^{a_i}$ *around* $\boldsymbol{\beta}_0$ *for* $i = 1, 2$. *Then* $\widehat{\mathrm{ATE}}_n(a_1,a_2) \xrightarrow{\mathrm{P}} \mathrm{ATE}(a_1,a_2)$ *and*

$$\sqrt{n}\Big(\widehat{\mathrm{ATE}}_n(a_1,a_2) - \mathrm{ATE}(a_1,a_2)\Big) \xrightarrow{d} N(0,\Gamma^{a_1,a_2}) \tag{15}$$

for $n \to \infty$ *with*

$$\Gamma^{a_1,a_2} = \mathrm{Var}\left(\mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_1}) - \mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_2}) + \mathrm{E}\Big(\frac{\partial}{\partial\boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_1}) - \frac{\partial}{\partial\boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_2})\Big)\dot{\boldsymbol{\beta}}(\mathbf{Z})\right) \tag{16}$$

The asymptotic variance may be consistently estimated by

$$\hat{\Gamma}^{a_1,a_2} = \frac{1}{n}\sum_{i=1}^{n}\left(\mu(\hat{\boldsymbol{\beta}}_n;\mathbf{X}_i^{a_1}) - \mu(\hat{\boldsymbol{\beta}}_n;\mathbf{X}_i^{a_2}) + \hat{\theta}_n^{a_2} - \hat{\theta}_n^{a_1} + \left(\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}}\mu(\hat{\boldsymbol{\beta}}_n;\mathbf{X}_j^{a_1}) - \frac{\partial}{\partial\boldsymbol{\beta}}\mu(\hat{\boldsymbol{\beta}}_n;\mathbf{X}_j^{a_2})\right)\right)\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i)\right)^2 \tag{17}$$

under assumption (9).

In many scenarios, it will be the case that $\mathrm{E}(\dot{\boldsymbol{\beta}}(\mathbf{Z}) \mid \mathbf{X}) = 0$ provided that the conditional mean model is correctly specified. Under this assumption, the asymptotic variances described so far may be simplified. For example, the asymptotic variance, Γ^{a_1,a_2}, for the average treatment effect reduces to

$$\mathrm{Var}\left(\mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_1}) - \mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_2})\right) + \mathrm{Var}\left(\mathrm{E}\left(\frac{\partial}{\partial\boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_1}) - \frac{\partial}{\partial\boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_2})\right)\dot{\boldsymbol{\beta}}(\mathbf{Z})\right) \tag{18}$$

An estimator of the variance based only on the second term is what was proposed by Dowd et al. (2014) and Qu and Luo (2015) and implemented in e.g. the `marginaleffects` package for `R` (Arel-Bundock 2022). An estimator of the asymptotic variance based on both terms is what was proposed by Terza (2016) and Bartlett (2018). In Sect. 5 we compare the estimator based on (18) with the estimator in (17).

3.1 Average treatment effects in a subgroup

Sometimes we are interested in the average treatment effect in a subgroup of the population, e.g. among those who received the treatment. More generally, if the effect of interest is between interventions a_1 and a_2 among those with $\mathbf{V} = \mathbf{v}$, then this corresponds to the parameter

$$\mathrm{ATE}(a_1, a_2 \mid \mathbf{v}) = \mathrm{E}(\mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_1}) - \mu(\boldsymbol{\beta}_0;\mathbf{X}^{a_2}) \mid \mathbf{V} = \mathbf{v}) = \psi^{a_1,\mathbf{v}} - \psi^{a_2,\mathbf{v}}.$$

This is naturally estimated by $\widehat{\mathrm{ATE}}_n(a_1, a_2 \mid \mathbf{v}) = \hat{\psi}_n^{a_1,\mathbf{v}} - \hat{\psi}_n^{a_2,\mathbf{v}}$.

Corollary 8 *Let a_1, a_2 be two interventions such that $\mu(\boldsymbol{\beta};\mathbf{x})$ is dominated square integrable and that $\frac{\partial}{\partial\boldsymbol{\beta}}\mu(\boldsymbol{\beta};\mathbf{x})$ is dominated integrable with respect to $\mathbf{X}^{a_i}$ around $\boldsymbol{\beta}_0$ for $i = 1, 2$. Then $\widehat{\mathrm{ATE}}_n(a_1, a_2 \mid \mathbf{v}) \xrightarrow{\mathrm{P}} \mathrm{ATE}(a_1, a_2 \mid \mathbf{v})$ and*

$$\sqrt{n}\left(\widehat{\mathrm{ATE}}_n(a_1, a_2 \mid \mathbf{v}) - \mathrm{ATE}(a_1, a_2 \mid \mathbf{v})\right) \xrightarrow{d} N(0, \Lambda^{a_1,a_2,\mathbf{v}}) \tag{19}$$

for $n \to \infty$ with

$$\Lambda^{a_1,a_2,\mathbf{v}} = \mathrm{Var}\left(\left(\mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_1}) - \mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_2}) + \psi^{a_2,\mathbf{v}} - \psi^{a_1,\mathbf{v}}\right)\frac{\mathbf{1}(\mathbf{V} = \mathbf{v})}{P(\mathbf{V} = \mathbf{v})} + \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_1}) - \frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_2}) \mid \mathbf{V} = \mathbf{v}\right)\dot{\boldsymbol{\beta}}(\mathbf{Z})\right)$$

The asymptotic variance may be consistently estimated by

$$\hat{\Lambda}^{a_1,a_2,\mathbf{v}} = \frac{1}{n}\sum_{i=1}^{n}\left(\left(\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^{a_1}) - \mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^{a_2}) + \hat{\psi}_n^{a_2,\mathbf{v}} - \hat{\psi}_n^{a_1,\mathbf{v}}\right)\frac{\mathbf{1}(\mathbf{V}_i = \mathbf{v})}{m/n} + \left(\frac{1}{m}\sum_{j:\mathbf{V}_j=\mathbf{v}}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_j^{a_1}) - \frac{\partial}{\partial \boldsymbol{\beta}}\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_j^{a_2})\right)\right)\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i)\right)^2$$

under assumption (9).

Most commonly this is used when estimating the average treatment effect among the treated or untreated. If A is binary and $a = 1$ corresponds to being treated and $a = 0$ corresponds to being untreated, then by taking $\mathbf{V} = A$ we note that $\psi^{1,1} - \psi^{0,1}$ is the average treatment effect among the treated while $\psi^{1,0} - \psi^{0,0}$ is that of the untreated.

3.2 Relative treatment effects

The average treatment effect above is measured as an absolute difference between potential outcome means. Sometimes one might be interested in modeling the treatment effect as a relative difference instead, i.e. as $\mathrm{RTE}(a_1, a_2) = \theta^{a_1}/\theta^{a_2}$ (Rubin 2010; VanderWeele 2015). When the outcome is binary this corresponds to a causal risk ratio. The asymptotic distribution of the corresponding estimator may be derived in similar fashion but rather than working on the original scale it is often more desirable to work on a logarithmic scale. To wit, consider the logarithm of the relative treatment effect, $\log(\mathrm{RTE}(a_1, a_2)) = \log(\theta^{a_1}) - \log(\theta^{a_2})$, as our target parameter for which it is assumed that $\theta^{a_1}, \theta^{a_2} > 0$. This parameter is naturally estimated by $\log(\widehat{\mathrm{RTE}}_n(a_1, a_2)) = \log(\hat{\theta}_n^{a_1}) - \log(\hat{\theta}_n^{a_2})$. We then obtain the following result.

Corollary 9 *Let a_1, a_2 be two interventions such that $\mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated square integrable and that $\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}; \mathbf{x})$ is dominated integrable with respect to $\mathbf{X}^{a_i}$ around $\boldsymbol{\beta}_0$ for $i = 1, 2$. Assume also that θ^{a_1} and θ^{a_2} are positive. Then $\log(\widehat{\mathrm{RTE}}_n(a_1, a_2)) \xrightarrow{\mathrm{P}} \log(\mathrm{RTE}(a_1, a_2))$ and*

$$\sqrt{n}\left(\log(\widehat{\mathrm{RTE}}_n(a_1, a_2)) - \log(\mathrm{RTE}(a_1, a_2))\right) \xrightarrow{d} N(0, \Pi^{a_1,a_2})$$

for $n \to \infty$ with

$$\Pi^{a_1,a_2} = \mathrm{Var}\left(\frac{1}{\theta^{a_1}}\left(\mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_1}) + \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_1})\right)\dot{\boldsymbol{\beta}}(\mathbf{Z})\right) - \frac{1}{\theta^{a_2}}\left(\mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_2}) + \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}^{a_2})\right)\dot{\boldsymbol{\beta}}(\mathbf{Z})\right)\right)$$

The asymptotic variance may be consistently estimated by

$$\hat{\Pi}_n^{a_1,a_2} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{\hat{\theta}_n^{a_1}}\left(\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^{a_1}) + \left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_j^{a_1})\right)\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i)\right) - \frac{1}{\hat{\theta}_n^{a_2}}\left(\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^{a_2}) + \left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_j^{a_2})\right)\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i)\right)\right\}^2$$

under assumption (9).

4 Estimation of the conditional mean model

So far we have considered given an asymptotically linear estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}_0$. We will now describe some situations where such an estimator arises. One very general situation is when $\hat{\boldsymbol{\beta}}_n$ is obtained as a solution to the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} H(\boldsymbol{\beta}; \mathbf{X}_i)(Y_i - \mu(\boldsymbol{\beta}; \mathbf{X}_i)) = 0 \tag{20}$$

for some function H. This is an example of an M-estimator. In this case, $\boldsymbol{\beta}_0$ is given as the solution to the limiting estimating equation $\mathrm{E}(H(\boldsymbol{\beta}; \mathbf{X})(Y - \mu(\boldsymbol{\beta}; \mathbf{X}))) = 0$. When the conditional mean model is correctly specified, then $\boldsymbol{\beta}_0$ describes the conditional distribution of Y given $\mathbf{X}$ but otherwise $\boldsymbol{\beta}_0$ corresponds to a best fit of $\mu(\boldsymbol{\beta}; \mathbf{X})$ to $\mathrm{E}(Y \mid \mathbf{X})$ in the sense of this limiting estimating equation.

One can show that, under certain regularity conditions, such estimators are indeed asymptotically linear.

Lemma 10 *Make the following regularity conditions:*

1. $\mathbf{x} \mapsto \mu(\boldsymbol{\beta}; \mathbf{x})$ *and* $\mathbf{x} \mapsto H(\boldsymbol{\beta}; \mathbf{x})$ *are measurable for all* $\boldsymbol{\beta}$ *in a neighborhood of* $\boldsymbol{\beta}_0$.
2. $\boldsymbol{\beta} \mapsto \mu(\boldsymbol{\beta}; \mathbf{x})$ *and* $\boldsymbol{\beta} \mapsto H(\boldsymbol{\beta}; \mathbf{x})$ *are twice continuously differentiable for all* $\mathbf{x}$.
3. $H(\boldsymbol{\beta}_0; \mathbf{X})(Y - \mu(\boldsymbol{\beta}_0; \mathbf{X}))$ *is square integrable.*
4. *The matrix*

$$\mathbf{M} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}}H(\boldsymbol{\beta}_0; \mathbf{X})(Y - \mu(\boldsymbol{\beta}_0; \mathbf{X})) - H(\boldsymbol{\beta}_0; \mathbf{X})\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X})\right) \tag{21}$$

exists and is invertible.

5. *The second-order partial derivatives,* $\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} H(\boldsymbol{\beta}; \mathbf{x})(y - \mu(\boldsymbol{\beta}; \mathbf{x}))$*, are dominated integrable with respect to* $(\mathbf{X}, Y)$ *in a neighborhood of* $\boldsymbol{\beta}_0$*.*

Under assumptions 1–5 we have, with a probability tending to 1 *as* $n \to \infty$*, an estimator* $(\hat{\boldsymbol{\beta}}_n)$ *solving* (20) *which is asymptotically linear with influence function*

$$\dot{\boldsymbol{\beta}}(\mathbf{z}) = -\mathbf{M}^{-1} H(\boldsymbol{\beta}_0; \mathbf{x})(y - \mu(\boldsymbol{\beta}_0; \mathbf{x})), \tag{22}$$

with $\mathbf{z} = (y, \mathbf{x})$*. That is*

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + \frac{1}{n} \sum_{i=1}^{n} \dot{\boldsymbol{\beta}}(\mathbf{Z}_i) + o_{\mathrm{P}}(n^{-1/2}) \tag{23}$$

and $\dot{\boldsymbol{\beta}}(\mathbf{Z})$ *is square integrable with mean zero.*

In particular,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{M}^{-1}\boldsymbol{\Sigma}(\mathbf{M}^{-1})^T) \tag{24}$$

as $n \to \infty$*, with* $\boldsymbol{\Sigma} = \mathrm{Var}\left(H(\boldsymbol{\beta}_0; \mathbf{X})(Y - \mu(\boldsymbol{\beta}_0; \mathbf{X})) \right)$.

Proof The is a direct consequence of Theorem 5.41 and 5.42 of van der Vaart (2000). □

If the conditional mean model is correctly specified, then the matrix $\mathbf{M}$ reduces to $\mathrm{E}(-H(\boldsymbol{\beta}_0; \mathbf{X}) \frac{\partial}{\partial\boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0; \mathbf{X}))$. Our estimate of $\mathbf{M}$ is, however, based on (21) as it is robust against model misspecification. Thus, we estimate $\mathbf{M}$ by

$$\hat{\mathbf{M}}_n = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial}{\partial\boldsymbol{\beta}} H(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i)(Y_i - \mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i)) - H(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i) \frac{\partial}{\partial\boldsymbol{\beta}} \mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i) \right).$$

and the influence function of (22) by

$$\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{z}) = -\hat{\mathbf{M}}_n^{-1} H(\hat{\boldsymbol{\beta}}_n; \mathbf{x})(y - \mu(\hat{\boldsymbol{\beta}}_n; \mathbf{x})). \tag{25}$$

This estimator of the influence function can be seen to satisfy the condition in (9) under the assumptions of Lemma 10, and hence we conclude that the results of this paper apply to the situation where this type of M-estimator is used to estimate $\boldsymbol{\beta}_0$ in the initial step.

The generalized linear model framework offers a way of fitting some conditional mean models by M-estimation. This means that existing generalized linear model software can be used to obtain the relevant regression parameter estimates even if it is not desirable to work under the distributional assumptions of generalized linear models. The distribution of the outcome, Y, given covariates $\mathbf{X}$, sometimes

referred to as the family, determines a function $v : \mathbf{R} \to \mathbf{R}$ usually called the variance function. A link function $g : \mathbf{R} \to \mathbf{R}$ determines the conditional mean model $\mu(\boldsymbol{\beta}; \mathbf{X}) = g^{-1}(\boldsymbol{\beta}^T f(\mathbf{X}))$ or $\mu(\boldsymbol{\beta}; \mathbf{X}) = m(\boldsymbol{\beta}^T f(\mathbf{X}))$ if m denotes the inverse of the link function. Here, $\mathbf{x} \mapsto f(\mathbf{x})$ is a vector function that determines how $\mathbf{X}$ enters the model technically, e.g., potentially with various interaction terms or with quadratic, cubic, or spline terms. The estimating equation, stemming from the corresponding maximum likelihood estimation, is then

$$\sum_{i=1}^{n} f(\mathbf{X}_i) \frac{m'(\boldsymbol{\beta}^T f(\mathbf{X}_i))}{v(m(\boldsymbol{\beta}^T f(\mathbf{X}_i)))} (Y_i - m(\boldsymbol{\beta}^T f(\mathbf{X}_i))) = 0.$$

In particular, this is an example of M-estimation with $H(\boldsymbol{\beta}; \mathbf{x}) = f(\mathbf{x}) m'(\boldsymbol{\beta}^T f(\mathbf{x}))/v(m(\boldsymbol{\beta}^T f(\mathbf{x})))$. If the link function of interest is canonical for the family specified then $m'(\boldsymbol{\beta}^T f(\mathbf{x})) = v(m(\boldsymbol{\beta}^T f(\mathbf{x})))$, leaving $H(\boldsymbol{\beta}; \mathbf{x}) = f(\mathbf{x})$.

All standard choices of m and v in the generalized linear model framework are seen to satisfy the first two assumptions of Lemma 10 and the remaining three assumptions are some that we are usually willing to assume. Hence we see that generalized linear models, which include important regression models such as linear and logistic regression, are all covered by our setup.

5 Simulations

In this section we evaluate the properties of the variance estimator proposed in the paper. First, we compare the performance of the asymptotic variance estimator to that of bootstrapping. Next, we compare the asymptotic variance estimator to the estimator proposed by Terza (2016) and Bartlett (2018) and investigate the impact of model misspecification. The Stata code used to carry out the simulations is available at https://github.com/snhansen/variance-estimation-gcomp-stata.

5.1 Comparison with bootstrapping

We consider here a scenario where the average treatment effect, $\mathrm{ATE}(1, 0) = \theta^1 - \theta^0$, is the parameter of interest. We estimate this by $\widehat{\mathrm{ATE}}_n(1, 0) = \hat{\theta}_n^1 - \hat{\theta}_n^0$ and as estimators of its standard error we consider (17) as well as bootstrapping. For the bootstrapping procedure, we calculate confidence intervals in two ways: by a Wald confidence interval and by the empirical 2.5- and 97.5-percentile.

As the data generating process, we consider a binary treatment variable $A \sim \mathrm{bin}(0.5)$, a continuous covariate $B \sim N(\mu_B, 1)$ and an outcome Y whose conditional distribution is given by $Y \mid A = a, B = b \sim N(\beta_0 + \beta_1 a + \beta_2 b + \beta_3 ab, 1)$. In this case, the average treatment effect is $\mathrm{ATE}(1, 0) = \beta_1 + \beta_3 \cdot \mu_B$. Simulations were run with $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 1, -1, 3)$ and $\mu_B = 2$, which implies that $\mathrm{ATE}(1, 0) = 7$.

A linear regression model with interaction between A and B was used as the conditional mean model which is then correctly specified. Sample sizes of 100 and 200 were

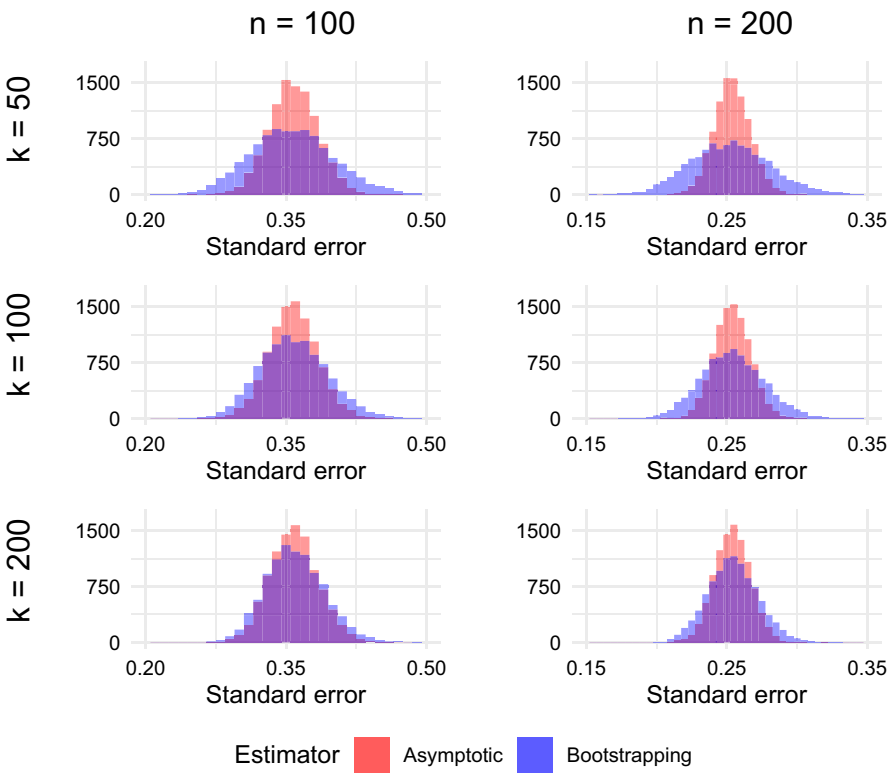

Fig. 1 Empirical distribution of standard error estimates using the asymptotic variance estimator and bootstrapping. Scenarios considered are sample sizes of $n = 100, 200$ and number of bootstrap samples of $k = 50, 100, 200$

considered and for the bootstrapping procedure we used 50, 100 and 200 bootstrap samples. For each configuration, 10,000 simulations were done.

In Fig. 1 we see the distribution of the 10,000 estimates of the standard error using the asymptotic estimator and bootstrapping, respectively. We note that the variation of the estimates based on bootstrapping is generally much larger compared to estimates based on the asymptotic approach. We do however see that, as the number of bootstrap samples increases, the variation of the bootstrap estimator decreases and it seems to converge towards that of the asymptotic estimator.

In Table 1 we calculated the coverage probabilities based on the 10,000 simulations. For the bootstrap approach we considered a Wald confidence interval as well as the interval given by the 2.5- and 97.5-percentile of the 10,000 estimates. The pattern here is the same. The asymptotic approach has coverage probabilities closer to 95 % in all scenarios, however, as the number of bootstrap samples increases, the normal-based confidence intervals do equally well. The confidence interval based on percentiles seem to be inferior to the two other approaches in most of the scenarios considered here.

Table 1 Coverage probabilities (average running time in seconds) over 10,000 simulations for the asymptotic variance estimator and bootstrap estimators

Estimator	$k = 50$	$k = 100$	$k = 200$
$n = 100$			
Asymptotic	94.7 (0.12 s)	94.1 (0.22 s)	94.6 (0.19 s)
Bootstrap (Wald)	94.1 (30.6 s)	94.0 (117 s)	94.5 (204 s)
Bootstrap (percentile)	91.9 (30.6 s)	93.3 (117 s)	94.2 (204 s)
$n = 200$			
Asymptotic	94.7 (0.11 s)	95.0 (0.20 s)	94.9 (0.12 s)
Bootstrap (Wald)	94.2 (31.5 s)	94.8 (113 s)	94.8 (239 s)
Bootstrap (percentile)	91.6 (31.5 s)	94.2 (113 s)	94.3 (239 s)

The bootstrap procedure is based on 50, 100 and 200 bootstrap samples (k), and we consider here a scenario with either 100 or 200 observations in total (n)

In Table 1 the average computational time (over the 10,000 simulations) is also given. Here we see that the time needed for the bootstrap approach quickly increases with the sample size and number of bootstrap samples as expected.

5.2 Model misspecification

Consider now an estimator $\hat{\boldsymbol{\beta}}_n$ based on a misspecified conditional mean model. This means that there is no $\boldsymbol{\beta}_0$ such that $\mathrm{E}(Y \mid \mathbf{X}) = \mu(\boldsymbol{\beta}_0; \mathbf{X})$ for the given mean function μ. As discussed in Sect. 3, the simplification used to arrive at (18) is not warranted in this case and the estimate based on this simplified expression may potentially be biased.

To investigate the magnitude of this bias, we compare coverage probabilities using the variance estimators based on (16) and (18). To do so, we assume that $Y \mid A = a, B = b, C = c \sim N(\alpha_1 a + \alpha_2 ab + \alpha_3 bc, 1)$ and that $B \sim N(0, 1)$ and $C \sim N(0, 1)$ are independent of each other. We also assume that $A \mid C = c \sim \mathrm{bin}(1, p_c)$ where $\mathrm{logit}(p_c) = \gamma c$. Under this data generating process, the average treatment effect is $\mathrm{ATE}(1, 0) = \alpha_1$.

We now consider using g-computation with the conditional mean model

$$\mu(\boldsymbol{\beta}; (A, B, C)) = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB + \beta_4 C.$$

This is a misspecified model as it fails to capture the interaction between B and C. The parameter targeted under this model is β_1 and one may show that $\beta_1 = \alpha_1$ owing to the independence between B and (A, C). This means that, despite using a misspecified model, the parameter targeted by this incorrect model, β_1, does in fact correspond to the average treatment effect, α_1, which is what we are interested in.

Simulations were done in two scenarios: $(\alpha_1, \alpha_2, \alpha_3) = (3, 4, 3)$ and $(\alpha_1, \alpha_2, \alpha_3) = (-1, -5, 2)$ and with γ values of 0, 1 and 3 and sample sizes 100, 500 and 1000. The results are shown in Table 2.

Table 2 Coverage probabilities over 10,000 simulations using the complete and simplified expressions for the asymptotic variance under a misspecified conditional mean model

	$n = 100$			$n = 500$			$n = 1000$		
	$\gamma = 0$	$\gamma = 1$	$\gamma = 3$	$\gamma = 0$	$\gamma = 1$	$\gamma = 3$	$\gamma = 0$	$\gamma = 1$	$\gamma = 3$
Scenario 1									
Complete	94.1	94.5	93.9	94.8	95.2	94.9	94.6	95.1	95.1
Simplified	94.3	97.9	99.1	94.8	98.2	99.3	94.6	98.6	99.4
Scenario 2									
Complete	94.1	93.0	92.7	94.6	94.7	94.7	94.4	94.7	94.8
Simplified	94.0	89.2	88.6	94.6	91.2	91.1	94.4	91.3	90.1

We consider here two scenarios: (1) $(\alpha_1, \alpha_2, \alpha_3) = (3, 4, 3)$ and (2) $(\alpha_1, \alpha_2, \alpha_3) = (-1, -5, 2)$ and varying degrees of dependency between A and C given by γ being 0, 1 or 3. We consider sample sizes of 100, 500 and 1000

We see that when the correlation between A and C is zero ($\gamma = 0$), then the two estimators result in similar coverage probabilities. However, when there is a substantial correlation between A and C, the simplified variance estimator will yield coverage probabilities far from 95 %. This is because the term which is omitted in the simplified expression is non-zero and its magnitude increases with the correlation between A and C. We also see that the simplified estimator can yield coverage probabilities well above and well below 95 % depending on the scenario, so that it can not be used as a conservative estimator either.

The above example corresponds to a situation where one is interested in the effect of some exposure A where, prior knowledge indicates an interaction between A and B, and where we wish to adjust for a confounder C. Most commonly the variable C is adjusted for by including their main effect and no interactions. In this specific situation, the model was misspecified in such a way that no bias was created, that is, the target parameter β_1 and the parameter of interest α_1 were identical. Although this is a realistic example, we will expect misspecification to induce bias more often than not. In this case, the target parameter will be $\mathrm{E}(\mu(\boldsymbol{\beta}_0; (1, B, C)) - \mathrm{E}(\mu(\boldsymbol{\beta}_0; (0, B, C))$ for some $\boldsymbol{\beta}_0$ which will not have an immediate causal interpretation. We will, however, still expect the simplified estimator to have poorer coverage compared to the complete estimator for this quantity.

6 Data example

Consider the publicly available data set from the observational study by Conors Jr et al. (1996) on the effect of right heart catherization (RHC) on the length of stay for patients admitted to an intensive care unit (ICU).[1] The data set includes information on $n = 5735$ patients admitted to an ICU for 1 of 9 prespecified disease categories. Information included length of stay in the ICU, disease category, RHC use within the first 24 h, sex, age, income category as well as a number of clinical measurements.

[1] Data available at https://hbiostat.org/data/.

The objective of the study was to investigate the relationship between the use of RHC during the first 24 h of care in the ICU and subsequent survival, length of stay, intensity of care and cost of care. Here we consider only the relationship between RHC use and length of stay and we will quantify this by the average treatment effect. That is, length of stay serves as the outcome, Y, and RHC use serves as the primary covariate, A. We assume conditional exchangeablity given disease category, sex, age and income, so these covariates collectively is what we denote by $\mathbf{B}$. A linear regression model will be used as the conditional mean model meaning that $\mu(\boldsymbol{\beta}; \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ with $\mathbf{X} = (A, \mathbf{B})$.

We start by showing step-by-step how the estimating procedure works in the scenario where we consider main effects of the categorical variables RHC (yes/no), disease category (1–9), sex (male/female) and income category (1–4) and where age enters linearly. This means that $\boldsymbol{\beta}$ has dimension $k = 15$ if an intercept is included. Solving the estimating equation of (20) with $H(\boldsymbol{\beta}; \mathbf{x}) = \mathbf{x}$ is, in this case, the same as the least squares approach and maximum likelihood estimation if the outcome follows a normal distribution. With an estimate, $\hat{\boldsymbol{\beta}}_n$, at hand, we calculate the two potential outcome means based on (4) for $a = 0, 1$ and subtract the two. If Y_i and $\mathbf{X}_i = (A_i, \mathbf{B}_i)$ denotes the observed outcome and observed covariates for the ith individual, respectively, and $\mathbf{X}_i^a = (a, \mathbf{B}_i)$ for $a = 0, 1$, then these are obtained by

$$\hat{\theta}_n^a = \frac{1}{n}\sum_{i=1}^{n} \mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^a) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^a \cdot \hat{\boldsymbol{\beta}}_n$$

We thus obtain an estimate of the average treatment effect of $\widehat{\mathrm{ATE}}_n(1, 0) = \hat{\theta}_n^1 - \hat{\theta}_n^0 = 3.82$.

Next, we turn to estimating the variance by (17). If $\mathbf{X}_d$ denotes the design matrix, that is, the $n \times k$ matrix whose ith row is $\mathbf{X}_i$, then $\hat{\mathbf{M}}_n = -\frac{1}{n}\mathbf{X}_d^T\mathbf{X}_d$ and the ith row of $\hat{\hat{\boldsymbol{\beta}}}$ is given by $\hat{\hat{\boldsymbol{\beta}}}_i = -(\hat{\mathbf{M}}_n)^{-1}\mathbf{X}_i(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_n)$ so that $\hat{\hat{\boldsymbol{\beta}}}$ is an $n \times k$ matrix. Finally, we note that $\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^a) = \mathbf{X}_i^a$ and so we arrive at an estimate of the asymptotic variance of $\hat{\Gamma}^{1,0} = 0.548$. Using this we obtain a Wald confidence interval of [2.38, 5.28].

In Table 3, the average treatment effect, standard error and Wald confidence interval is given for different specifications of the conditional mean model. We consider the following four specifications:

1. Main effects of RHC, disease category, sex, income category and age (linear)
2. Main effects of RHC, disease category, sex, income category and age (linear) with interaction between RHC and disease category.
3. Main effects of RHC, disease category, sex, income category and age (restricted cubic spline) with interaction between RHC and disease category.
4. Main effects of RHC, disease category, sex, income category and age (restricted cubic spline) with interaction between RHC and disease category and interaction between RHC and sex.

The results of Table 3 show that across the four models, estimates of the effect of RHC use on length of stay range between 3.59 and 3.89 days. The corresponding standard errors across the four models are very similar and hence the choice of model

Table 3 Estimate, standard error based on (17) and 95 % Wald confidence interval for the average treatment effect of RHC on length of hospital stay using four different conditional mean models

Model	Estimate	Standard error	95% CI
1	3.82	0.741	[2.38, 5.28]
2	3.89	0.742	[2.44, 5.35]
3	3.59	0.742	[2.14, 5.05]
4	3.61	0.742	[2.16, 5.06]

does not appear to have a large impact on the conclusion of the analysis. The computational time for calculating the estimate and standard error was around 100 ms for all four models.

The results were obtained using the Stata function `gcomp_effects` available at https://github.com/snhansen/variance-estimation-gcomp-stata. The code used for the example is available in the repository as well.

7 Discussion

We have derived the asymptotic distribution for estimators using g-computation of various causal parameters for a time-fixed exposure. This includes potential outcome means and average treatment effects, in the entire population but also in subgroups. The results are derived in a very general setup that requires only the existence of an asymptotically linear estimator of the parameters in the conditional mean model and a well-behaved estimator of its influence function. We saw that this setup includes when M-estimation is used as the estimation technique which the generalized linear models framework is an example of. In particular, we saw that important regression models such as linear and logistic regression are covered by this. As the results in the paper concerns the asymptotic behaviour of these estimators, we can only guarantee their stated properties in large samples. In small samples, the confidence intervals produced by the methods in this paper can, for example, have coverage far from 95 % even though models are correctly specified.

The existence of an asymptotically linear estimator of the parameters, $\boldsymbol{\beta}_0$, in the conditional mean model is guaranteed by the regularity conditions of Lemma 10. In particular, the assumption that $\mathbf{M}$ is invertible is crucial for the identification of $\boldsymbol{\beta}_0$ and this may be an issue in the setting of sparse data, positivity violations and overparameterized models. In the causal inference literature, it is common practice to also invoke a positivity assumption for identification of potential outcome means. The positivity assumption states that $P(A = a \mid B) > 0$ almost surely for all a considered (Petersen et al. 2012). This is typically necessary to ensure the identification of $\boldsymbol{\beta}_0$. However, positivity violations can be addressed by introducing additional parametric assumptions, although these assumptions cannot be tested from the data. Consequently, violations of positivity or near-positivity generally pose a threat to the invertibility of $\mathbf{M}$ and, consequently, the identification of $\boldsymbol{\beta}_0$. Therefore, caution should be exercised in such cases.

Based on the asymptotic distribution, a plug-in estimator was proposed as an estimator of its variance. We showed that, under mild regularity assumptions, this is indeed a consistent estimator. We evaluated its properties, such as coverage and variability, in a simulation study and showed how it compared to bootstrapping. The conclusion was that, compared to bootstrapping with too few bootstrap samples, it had better coverage and less variability around the asymptotic variance. When the number of bootstrap samples increased, this difference appeared to diminish however. The other obvious advantage of the plug-in estimator is the time needed to compute it. Bootstrapping is arguably the most dominant method for obtaining variance estimates of complex estimators but even with a simple conditional mean model, a small sample size and a modest number of bootstrap samples, the procedure is relatively slow. This obviously scales badly with increasing complexity of the conditional mean model, increasing sample size and number of bootstrap samples.

Because bootstrapping is a slow procedure, the analyst might be tempted or perhaps even forced to use relatively few bootstrap samples. This results, however, in a variance estimator that is more variable compared to the plug-in estimator. When there is a choice between two estimators of the asymptotic variance, both being unbiased, we argue that the less variable estimator will be preferred. Although both estimators will yield confidence intervals with the correct coverage probability, at least assuming a large enough sample, the power of the corresponding statistical tests will generally be larger using the less variable estimator. This seems like an appealing property.

In the simulations, we had to keep the number of bootstrap samples below 200 to ensure that simulations finished within a reasonable amount of time. Of course it is preferable with more bootstrap samples. Practical examples using more than 200 bootstrap samples do exist (Snowden et al. 2011; Keil et al. 2014; Wang et al. 2017; Westreich et al. 2012). Still, examples with 200 bootstrap samples are plenty (Wang and Arah 2015; Breskin et al. 2020). Moreso, a recent `R` package has implemented the g-computation method with 200 bootstrap samples as the default because, as the authors note, using more bootstrap samples "can result in potentially long runtimes, depending on the computing power of the user's computer" (Grembi and McQuade 2022).

In most practical examples, we will expect the conditional mean model to be misspecified to some degree. In this case, we have argued that the target parameter is $\mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^a))$ where $\boldsymbol{\beta}_0$ is a sort of best fit of $\mu(\boldsymbol{\beta}; \mathbf{X})$ to $\mathrm{E}(Y \mid \mathbf{X})$ and hence we will expect a discrepancy between the target parameter and the parameter of interest, $\mathrm{E}(Y(a))$. Our hope is that, by picking a sufficiently flexible model, the discrepancy between the two will be small however. In this situation, we would want to estimate our target, $\mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}))$, unbiasedly but also to estimate its standard error unbiasedly. Being able to do so is what enables us to calculate confidence intervals with the correct coverage. Unbiased estimation of the standard error is ensured by using the variance estimator in (17) but not by using the simplified estimator in (18). Using the simplified estimator can, as we have seen in simulations, yield coverage probabilities much below or above the wanted 95%.

As an alternative to the two-step approach of this paper, one could consider formulating the two steps as two estimating equations and use (stacked) M-estimation. The sandwich variance estimator from the stacked M-estimation approach corresponds to

the variance estimator of this paper. This M-estimation approach has been implemented in the Python library `delicatessen` as pointed out by a reviewer.

The results of the paper only apply to time-fixed exposures. With time-varying exposures, g-computation is much more complex and determining the asymptotic distribution becomes much more involved (Robins and Hernán 2009). The main result of Theorem 2 can, however, in combination with the delta method, be used to find the asymptotic distribution of any causal parameter which can be expressed as a smooth function of non-nested potential outcome means. For means of nested potential outcomes, which are used to define key parameters in e.g. mediation analysis and scenarios where exposures are measured at multiple time points, the results here cannot immediately be extended as it involves more than one conditional mean model.

Funding Open access funding provided by Aarhus Universitet.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

Appendix A: Theoretical results

Lemma 11 *Let* $\mathbf{X}$ *be a random variable with values in* $\mathbf{X}$ *and let* $\Theta \subseteq \mathbb{R}^p$. *Let also* $g : \mathbf{X} \times \Theta \to \mathbb{R}$ *be a function such that* $\theta \mapsto g(\theta; \mathbf{x})$ *is continuous at* θ_0 *for all* $\mathbf{x}$ *and such that* $g(\theta_0; \mathbf{x})$ *is dominated integrable with respect to* $\mathbf{X}$ *in a neighborhood around* θ_0. *Consider an i.i.d. sample* $\mathbf{X}_1, \mathbf{X}_2, \ldots$ *of* $\mathbf{X}$ *and assume that an estimator* $\hat{\theta}_n$ *exists with the property that* $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$. *Then*

$$\frac{1}{n}\sum_{i=1}^{n} g(\hat{\theta}_n; \mathbf{X}_i) \xrightarrow{\mathrm{P}} \mathrm{E}(g(\theta_0; \mathbf{X}))$$

as $n \to \infty$.

Proof A uniform law of large numbers ensures that

$$\sup_{\theta \in K} \left| \frac{1}{n}\sum_{i=1}^{n} g(\theta; \mathbf{X}_i) - \mathrm{E}(g(\theta_0; \mathbf{X})) \right| \xrightarrow{\mathrm{P}} 0$$

for all compact sets K including θ_0 small enough. Since $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, the probabilities $P(\|\hat{\theta}_n - \theta_0\| \leq \varepsilon)$ can be made arbitrarily close to 1 by choosing n large enough, which ensures the desired convergence. □

Appendix B: Proof of Proposition 1

Consider, for fixed $\mathbf{x}$, the mapping $[0, 1] \ni t \mapsto \tilde{\mu}_n(t; \mathbf{x}) = \mu(\boldsymbol{\beta}_0 + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0); \mathbf{x})$ which is differentiable with derivative

$$\tilde{\mu}_n'(t; \mathbf{x}) = \frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0 + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0); \mathbf{x})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0).$$

Now, the identity

$$\tilde{\mu}(1; \mathbf{x}) - \tilde{\mu}(0; \mathbf{x}) = \tilde{\mu}'(0; \mathbf{x}) + \int_0^1 (\tilde{\mu}_n'(t; \mathbf{x}) - \tilde{\mu}_n'(0; \mathbf{x}))\, \mathrm{d}t$$

translates into

$$\mu(\hat{\beta}_n; \mathbf{x}) - \mu(\boldsymbol{\beta}_0; \mathbf{x}) = \frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0; \mathbf{x})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + R_{1,n}(\mathbf{x}), \tag{B1}$$

where

$$R_{1,n}(\mathbf{x}) = \int_0^1 \left(\frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0 + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0); \mathbf{x}) - \frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0; \mathbf{x}) \right) (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\, \mathrm{d}t.$$

Inserting (3) into (B1) yields

$$\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{x}) = \mu(\boldsymbol{\beta}_0; \mathbf{x}) + \frac{1}{n} \sum_{i=1}^n \dot{\boldsymbol{\beta}}(\mathbf{Z}_i) \frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0; \mathbf{x}) + R_{1,n}(\mathbf{x}) + R_{2,n}(\mathbf{x}) \tag{B2}$$

with $R_{2,n}(\mathbf{x}) = n^{-1/2} \frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0; \mathbf{x}) U_n$ for some sequence (U_n) of random variables with $U_n \xrightarrow{\mathrm{P}} 0$ whose existence is guaranteed by (3). Thus,

$$\hat{\theta}_n^a = \theta^a + \frac{1}{n} \sum_{i=1}^n \dot{\theta}^a(\mathbf{Z}_i) + \frac{1}{n} \sum_{i=1}^n R_{1,n}(\mathbf{X}_i^a) + \frac{1}{n} \sum_{i=1}^n R_{2,n}(\mathbf{X}_i^a) + \frac{1}{n} \sum_{i=1}^n R_{3,n}(\mathbf{Z}_i),$$

where for $\mathbf{z} = (y, \mathbf{x})$

$$R_{3,n}(\mathbf{z}) = \left(\frac{1}{n} \sum_{i=1}^n \dot{\boldsymbol{\beta}}(\mathbf{z}) \right) \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0; \mathbf{x}^a) - \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} \mu(\boldsymbol{\beta}_0; \mathbf{X}^a) \right) \right).$$

What is left to show is that all three remainder terms are $o_{\mathrm{P}}(n^{-1/2})$.

For the second term, we have that

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} R_{2,n}(\mathbf{X}_i^a) = U_n \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}_i^a)$$

in which U_n is $o_P(1)$ and the remaining part is $O_P(1)$ by the law of large numbers and hence their product is $o_P(1)$.

For the third term, we have that

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} R_{3,n}(\mathbf{Z}_i) = \left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \dot{\boldsymbol{\beta}}(\mathbf{Z}_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}_i^a) - \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}^a)\right)\right)$$

in which the first factor is $O_P(1)$ by the central limit theorem and the latter factor is $o_P(1)$ by the law of large numbers and hence their product is $o_P(1)$.

For the first term, we define

$$g(\boldsymbol{\beta}; \mathbf{x}) = \int_0^1 \left| \frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0 + t(\beta - \boldsymbol{\beta}_0); \mathbf{x}) - \frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{x}) \right| \mathrm{d}t$$

and since $\boldsymbol{\beta} \mapsto g(\boldsymbol{\beta}; \mathbf{x})$ is seen to be continuous at $\boldsymbol{\beta}_0$ and dominated integrable with respect to $\mathbf{X}^a$ around $\boldsymbol{\beta}_0$ we have from Lemma 11 that

$$\frac{1}{n}\sum_{i=1}^{n} g(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^a) = o_{\mathrm{P}}(1).$$

Now we note that $|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0| = O_{\mathrm{P}}(n^{-1/2})$ as a consequence of (3), and hence the inequality

$$\left|\frac{1}{n}\sum_{i=1}^{n} R_{1,n}(\mathbf{X}_i^a)\right| \leq |\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0| \frac{1}{n}\sum_{i=1}^{n} g(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^a)$$

implies that the remainder term is $O_{\mathrm{P}}(n^{-1/2})o_{\mathrm{P}}(1) = o_{\mathrm{P}}(n^{-1/2})$ as wanted.

The square integrability of $\dot{\theta}^a(\mathbf{Z})$ follows from $\mu(\boldsymbol{\beta}_0; \mathbf{X}^a)$ and $\dot{\boldsymbol{\beta}}(\mathbf{Z})$ both being assumed square integrable and its mean is given by

$$\mathrm{E}(\dot{\theta}^a(\mathbf{Z})) = \mathrm{E}(\mu(\boldsymbol{\beta}_0; \mathbf{X}^a)) - \theta^a + \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}^a)\right)\mathrm{E}(\dot{\boldsymbol{\beta}}(\mathbf{Z})) = 0$$

since $\mathrm{E}(\dot{\boldsymbol{\beta}}(\mathbf{Z})) = 0$.

Appendix C: Proof of Theorem 3

If we let $b_{n,i} = \mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_i^a) - \hat{\theta}_n^a$ and $c_n = \frac{1}{n}\sum_{j=1}^n \frac{\partial}{\partial \boldsymbol{\beta}}\mu(\hat{\boldsymbol{\beta}}_n; \mathbf{X}_j^a)$, then we may write

$$\hat{\boldsymbol{\Gamma}}_n^{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^n \big(b_{n,i} + c_n \hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i)\big)^{\otimes 2}.$$

Now we define

$$\tilde{\boldsymbol{\Gamma}}_n^{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^n \big(b_{n,i} + c_n \dot{\boldsymbol{\beta}}(\mathbf{Z}_i)\big)^{\otimes 2}$$

and note that $\tilde{\boldsymbol{\Gamma}}_n^{\mathbf{a}} \to P\hat{\boldsymbol{\Gamma}}_n^{\mathbf{a}}$ as a consequence of Lemma 11 of the Appendix. Moreover, we have the decomposition

$$\hat{\boldsymbol{\Gamma}}_n^{\mathbf{a}} = \tilde{\boldsymbol{\Gamma}}_n^{\mathbf{a}} + R_{1,n}^a + R_{2,n}^a + (R_{2,n}^a)^T$$

where

$$R_{1,n}^a = \frac{1}{n}\sum_{i=1}^n \big(c_n(\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i) - \dot{\boldsymbol{\beta}}(\mathbf{Z}_i))\big)^{\otimes 2}$$

and

$$R_{2,n}^a = \frac{1}{n}\sum_{i=1}^n \big(b_{n,i} + c_n \dot{\boldsymbol{\beta}}(\mathbf{Z}_i)\big)\big(c_n(\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i) - \dot{\boldsymbol{\beta}}(\mathbf{Z}_i))\big)^T.$$

If $|\cdot|$ denotes the max-norm, then we have

$$\begin{aligned}
|R_{1,n}^a| &\le \frac{1}{n}\sum_{i=1}^n |c_n(\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i) - \dot{\boldsymbol{\beta}}(\mathbf{Z}_i))|^2 \\
&\le k^2|c_n|^2 \frac{1}{n}\sum_{i=1}^n |\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i) - \dot{\boldsymbol{\beta}}(\mathbf{Z}_i)|^2 \\
&\le k^2|c_n|^2 g_n^2 \frac{1}{n}\sum_{i=1}^n f(\mathbf{Z}_i)^2
\end{aligned}$$

which is seen to converge in probability to 0 as $|c_n|^2 \xrightarrow{\mathrm{P}} |\mathrm{E}(\frac{\partial}{\partial \boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0; \mathbf{X}^a))|^2$ by Lemma 11, $\frac{1}{n}\sum_{i=1}^n f(\mathbf{Z}_i)^2 \to P\mathrm{E}(f(\mathbf{Z})^2)$ and $g_n \xrightarrow{\mathrm{P}} 0$. For the second remainder term we have

$$|R^a_{2,n}|^2 \leq \frac{1}{n^2}\sum_{i=1}^{n}|(b_{n,i} + c_n\dot{\boldsymbol{\beta}}(\mathbf{Z}_i))|^2|(c_n(\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i) - \dot{\boldsymbol{\beta}}(\mathbf{Z}_i))^T|^2$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}|(b_{n,i} + c_n\dot{\boldsymbol{\beta}}(\mathbf{Z}_i))|^2\frac{1}{n}\sum_{i=1}^{n}|(c_n(\hat{\dot{\boldsymbol{\beta}}}_n(\mathbf{Z}_i) - \dot{\boldsymbol{\beta}}(\mathbf{Z}_i))^T|^2$$

by the Cauchy-Schwarz inequality. Now, the first factor can be bounded by

$$\frac{1}{n}\sum_{i=1}^{n}|b_{n,i}|^2 + k^2|c_n|^2\frac{1}{n}\sum_{i=1}^{n}|\dot{\boldsymbol{\beta}}(\mathbf{Z}_i)|^2$$

which is seen to converge in probability to $k^2|\mathrm{E}(\frac{\partial}{\partial\boldsymbol{\beta}}\mu(\boldsymbol{\beta}_0;\mathbf{X}^a))|^2\mathrm{E}(|\dot{\boldsymbol{\beta}}(\mathbf{Z})|^2)$. Now the result follows as the second factor converges to 0 in probability.

References

Arel-Bundock V (2022) The marginaleffects package for R. https://vincentarelbundock.github.io/marginaleffects/articles/sandwich.html. Accessed 11 Jan 2023

Bartlett JW (2018) Covariate adjustment and estimation of mean response in randomised trials. Pharm Stat 17(5):648–666

Breskin A, Edmonds A, Cole SR et al (2020) G-computation for policy-relevant effects of interventions on time-to-event outcomes. Int J Epidemiol 49(6):2021–2029

Chatton A, Le Borgne F, Leyrat C et al (2020) G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. Sci Rep 10(1):9219

Conors AF Jr, Speroff T, Dawson NV et al (1996) The effectiveness of right heart catheterization in the initial care of critically ill patients. Support investigators. JAMA 276(1):889–975

Dowd BE, Greene WH, Norton EC (2014) Computation of standard errors. Health Serv Res 49(2):731–750

Graubard BI, Korn EL (1999) Predictive margins with survey data. Biometrics 55(2):652–659

Grembi JA, McQuade ETR (2022) Introducing riskCommunicator: an R package to obtain interpretable effect estimates for public health. PLoS ONE 17(7):e02,65368

Hernán MA, Robins JM (2020) Causal inference: what if. Chapman & Hall/CRC, Boca Raton

Keil AP, Edwards JK, Richardson DB et al (2014) The parametric g-formula for time-to-event data: intuition and a worked example. Epidemiology 25(6):889–897

Newey WK, McFadden D (1994) Large sample estimation and hypothesis testing. In: Engle RF, McFadden D (eds) Handbook of Econometrics. North Holland

Nianogo RA, Wang MC, Wang A et al (2017) Projecting the impact of hypothetical early life interventions on adiposity in children living in low-income households. Pediatric Obes 12(5):398–405

Petersen ML, Porter KE, Gruber S et al (2012) Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res 21(1):31–54

Qu Y, Luo J (2015) Estimation of group means when adjusting for covariates in generalized linear models. Pharm Stat 14(1):56–62

Robins JM, Hernán MA (2009) Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, et al (eds), Longitudinal data analysis. CRC Press, chap 23

Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Modell 7(9–12):1393–1512

Rosenbaum PR, Rubin DB (1983) The central role for the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Rubin DB (2010) Causal inference. In: Peterson P, Baker E, McGaw B (eds) International encyclopedia of education, 3rd edn. Elsevier, Oxford, pp 66–71. https://doi.org/10.1016/B978-0-08-044894-7.01313-0

Snowden JM, Rose S, Mortimer KM (2011) Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol 1(173):731–738

Terza JV (2016) Inference using sample means of parametric nonlinear data transformations. Health Serv Res 51(3):1109–1113

van der Vaart AW (2000) Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge

VanderWeele TJ (2015) Explanation in causal inference: methods for mediation and interaction. Oxford University Press, Oxford

Wang A, Arah OA (2015) G-computation demonstration in causal mediation analysis. Eur J Epidemiol 30(10):1119–1127

Wang A, Nianogo RA, Arah OA (2017) G-computation of average treatment effects on the treated and the untreated. BMC Med Res Methodol 17(3):1–5

Westreich D, Cole SR, Young JG et al (2012) The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident aids or death. Stat Med 31(18):2000–2009