Zetterqvist, Johan; Waernbaum, Ingeborg

Working Paper

Semi-parametric estimation of multi-valued treatment effects for the treated: Estimating equations and sandwich estimators

Working Paper, No. 2020:4

Provided in Cooperation with:
IFAU - Institute for Evaluation of Labour Market and Education Policy, Uppsala

Suggested Citation: Zetterqvist, Johan; Waernbaum, Ingeborg (2020) : Semi-parametric estimation of multi-valued treatment effects for the treated: Estimating equations and sandwich estimators, Working Paper, No. 2020:4, Institute for Evaluation of Labour Market and Education Policy (IFAU), Uppsala

This Version is available at:
https://hdl.handle.net/10419/227852

WORKING PAPER 2020:4

Semi-parametric estimation of multi-valued treatment effects for the treated: estimating equations and sandwich estimators

Johan Zetterqvist
Ingeborg Waernbaum

IFAU
INSTITUTE FOR EVALUATION OF LABOUR MARKET AND EDUCATION POLICY

The Institute for Evaluation of Labour Market and Education Policy (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala.

IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market and educational policies, studies of the functioning of the labour market and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

More information about IFAU and the institute's publications can be found on the website www.ifau.se

ISSN 1651-1166

Semi-parametric estimation of multi-valued treatment effects for the treated: estimating equations and sandwich estimators[a]

by

Johan Zetterqvist[b] and Ingeborg Waernbaum[c]

February 14, 2020

Abstract

An estimand of interest in empirical studies with observational data is the average treatment effect of a multi-valued treatment in the treated subpopulation. We demonstrate three estimation approaches: outcome regression, inverse probability weighting and inverse probability weighted regression, where the latter estimator holds a so called doubly robust property. Here, we define the estimators in the framework of partial M-estimation and derive corresponding sandwich estimators of their variances. The finite sample properties of the estimators and the proposed variance estimators are evaluated in simulations that reproduce designs from a previous simulation study in the literature of multi-valued treatment effects. The proposed variance estimators are investigated and compared to a bootstrap estimator.

Keywords: ATT, causal inference, inverse probability weighting, doubly robust, weighted ordinary least squares
JEL-codes: C14

[a] The authors are grateful to Derya Uysal for valuable comments.

[b] IFAU, johan.zetterqvist@ki.se

[c] IFAU, ingeborg.waernbaum@ifau.uu.se

Table of contents

1 Introduction

A natural extension when considering causal parameters for a binary treatment is to define treatments with more than two treatment levels, which we refer to as a multi-valued treatment in the sequel. In this setting when there are three, or more levels of the treatment, the causal effect parameters are usually defined as contrasts between two levels of the treatment, see e.g. the review by Imbens and Wooldridge (2009). Estimators of multi-valued treatment effects have been previously proposed and compared (Cattaneo, 2010; Linden et al., 2016; Yang et al., 2016). For estimators using the probability of treatment given the covariates, the propensity score is replaced with the extended generalized propensity score (GPS), a function of the covariates predicting the discrete levels of the treatment (Imbens, 2000; Imai and Van Dyk, 2004; Feng et al., 2012).

An estimator using a GPS weighted outcome regression model was proposed by Uysal (2015) for the average causal effect and the average causal effect of the treated (ATT). The estimator, henceforth referred to as the weighted ordinary least squares (WOLS) estimator, uses parametric model assumptions for the outcome regression and the GPS and is doubly robust, meaning that it is consistent for the true treatment effect if at least one of the two models is correctly specified, see e.g. Bang and Robins (2005); Seaman and Vansteelandt (2018); Tan (2010) for overviews on doubly robust estimators.

In this paper we demonstrate three estimators for the average treatment effect of the treated: 1) an outcome regression estimator, the ordinary least squares (OLS) estimator, 2) an inverse-probability weighting estimator (IPW), and 3) the GPS-weighted OLS estimator WOLS. We present the regression and IPW estimator(s) for the average causal effect of the treated in Uysal (2015) in the framework of partial M-estimators (Boos and Stefanski, 2013, Chapter 7). Under an assumption of multinomial logistic regression for the generalized propensity score and linear regression for the outcome we give estimating equations for the treatment effect estimators and derive their corresponding sandwich matrices. The M-estimation framework makes it straightforward to extend the estimators to also include estimation of several ATT parameters, thus making inference on functions of ATT parameters possible.

Using two simulation designs by Yang et al. (2016) we study the finite sample performance of the estimators and the sandwich estimators of their respective variance. We compare the bias and root mean squared error (RMSE) of OLS, IPW, and WOLS using combinations of correctly and incorrectly specified models for the outcome regression and GPS. Further, we compare the proposed sandwich estimator with a bootstrap estimator. As a reference, we compare these values with the empirical standard errors calculated over the simulations. Finally, we construct 95% Wald confidence intervals (CI) and evaluate the coverage probabilities.

2 Model and Theory – parameters of interest

We consider a sample of N individuals indexed by $i = 1, \ldots, N$ and a discrete treatment t with levels $t \in \mathscr{T} = \{0, 1, \ldots, K\}$. We let the random variable T_i represent the received treatment for individual i and use the treatment indicator $D_i(t) = I(T_i = t)$ which is equal to 1 if individual i received treatment t and 0 otherwise. To denote the number of individuals with factual treatment level t, we use $N_t = \sum_{i=1}^{N} D_i(t)$. The potential outcomes $Y_i(t)$ are defined for the corresponding treatment levels and we assume consistency, i.e., the observed outcome is the potential outcome under the level of treatment received $Y_i = Y_i(T_i)$. We use X_i to denote a vector of covariates, which is observed for each individual i. We also assume that every X_i are measured before treatment assigment, i.e. before T_i is realized. To simplify notation, we will let 1 be the first element of X_i. Finally we let O_i denote the observed vector $O_i = (Y_i, T_i, X_i)$. Henceforth, we will drop the index i when not needed. We follow the notation in Uysal (2015) and define the parameter of interest as the average causal effect of treatment l vs m for some $l, m \in \mathscr{T}$. Here, we use the notation $\mu_l = E[Y(l)]$ and $\mu_m = E[Y(m)]$, for the expected potential outcomes when the treatment level is set to l and m respectively. The average treatment effect of l versus m is denoted

$$\tau_{lm} = E\left[Y(l) - Y(m)\right] = \mu_l - \mu_m.$$

Similarly, the average potential outcomes when the treatment level is set to l and m among individuals with factual treatment level l are denoted $\mu_{l|l} = E[Y(l)|T = l]$ and $\mu_{m|l} =$

$E[Y(m)|T=l]$ respectively. The average treatment effect of treatment l versus m for the population receiving treatment $T=l$ can then be written as

$$\gamma_{lm|l} = \mu_{l|l} - \mu_{m|l} = E\left[Y(l) - Y(m) \mid T=l\right].$$

We will henceforth use l to denote the reference level of the treatment assumed in estimation of the ATT parameter. Throughout we use

Assumption 1 *[consistency rule]*
For each individual, $T=t \Longrightarrow Y=Y(t)$.

We will also use

Assumption 2 *[No unmeasured confounding]*
$Y(m) \perp\!\!\!\perp D(t)|X \qquad$ *for* $t=l,m$,

where m is the level of T considered as treated in the ATT parameter. We note that a similar unconfoundedness assumption was used by Słoczyński and Wooldridge (2018), but we made the assumption slightly weaker in that we only require the conditional independence to hold for the levels l and m involved in the target parameter $\gamma_{lm|l}$. We also assume overlapping distributions

Assumption 3 *[Overlap]*
$\eta < P(T=t|X)$, *for some* $\eta > 0$, *all* X *and each* $t \in \mathscr{T}$.

where $P(T=t|X)$ is the GPS, i.e. the probability of receiving treatment t conditional on the covariates. It follows from Assumptions 1 and 2 that

$$\mu_{t|l} = \mathrm{E}\left[Y(t)|T=l\right] = \mathrm{E}\left[\mathrm{E}\left(Y|T=t,X\right)|T=l\right].$$

When Assumption 3 holds, the ATT parameter $\gamma_{lm|l} = \mu_{l|l} - \mu_{m|l}$ is therefore nonparametrically identified from the data. Using a linear regression model

Assumption 4 *[Outcome regression model]*

$$\mu_t(X;\beta_t) = X'\beta_t \quad \textit{for} \quad t \in \mathscr{T}$$

for $\mathrm{E}(Y|X,T=t)$, we can also parametrically identify $\gamma_{lm|l}$, since

$$\gamma_{lm|l} = \mu_{l|l} - \mu_{m|l} = \mathrm{E}\left[\mu_l(X;\beta_l) - \mu_m(X;\beta_m) \mid T = l\right]$$

We will write $\hat{\mu}_t(X)$ to denote an estimator $\mu_t(X;\hat{\beta}_t)$ of $\mathrm{E}[Y|X,T=t]$, where $\hat{\beta}_t$ is an estimator of β_t. As shown in Appendix A, it follows from Assumptions 1 and 2 that

$$\mu_{t|l} = \frac{1}{\Pr(T=l)}\mathrm{E}\left[\frac{\Pr(T=l \mid X)}{\Pr(T=t \mid X)}D(t)Y\right],$$

for $t \in \{l,m\}$. With Assumptions 1 - 3 and a model for the GPS $\Pr(T=l \mid X)$, it is therefore possible to parametrically identify $\gamma_{lm|l} = \mu_{l|l} - \mu_{m|l}$ from the observed data, without model assumptions about $\mathrm{E}(Y \mid T=t,X)$ for $t = l,m$. Following the notation in Uysal (2015), we use $r(t,X)$ to denote a parametric model for the GPS, $\Pr(T=t|X;\delta)$, indexed by a parameter $\delta = (\delta_0,\delta_1,\ldots,\delta_p)$. In this paper, we assume that the GPS follows the multinomial logistic regression model:

Assumption 5 *[GPS model]*

$$r(t,X) = \Pr(T=t|X;\delta) = \frac{\sum_{s=1}^{K} I(t=s)\exp(X'\delta_s)}{\sum_{s=1}^{K}\exp(X'\delta_s)},$$

where $\delta_0 \equiv 0$. We will write $\hat{r}(t,X) = \Pr(T=t|X;\hat{\delta})$ for an estimator of $r(t,X)$ based on an estimator $\hat{\delta}$ of δ and observed covariates X. We note that the method does not require this particular model assumption for GPS and that an extension to other models, e.g. an ordered logit model, is straight-forward.

3 Estimators

In the following we study three estimators of $\gamma_{lm|l}$, the average treatment effect of treatment l versus treatment m for the population taking treatment l. Using the framework of M-estimators (Boos and Stefanski, 2013, Chapter 7), we use M-estimators for the parameters in the GPS-model in Assumption 5 and for the parameters in the outcome regression model in Assumption 4. By combining these estimators with partial estimators for the

parameter $\gamma_{lm|l}$, we derive an IPW-based estimator, an outcome regression estimator and a doubly robust estimator. Technically, this is achieved by stacking estimating equations for the components of $\hat{\gamma}_{lm|l}$ to construct full M-estimators for $\gamma_{lm|l}$. The doubly robust estimator combines the IPW-based estimator and the outcome regression estimator for $\gamma_{lm|l}$ in such a way that it is consistent for the true value of $\gamma_{lm|l}$ when at least one, not necessarily both, models are correctly specified and when the corresponding partial M-estimator is consistent for the true value of the parameter. The estimator thus provides some protection against bias due to model misspecification of either the GPS model or the outcome regression model.

We first define an outcome regression estimator of $\mu_{t|l}$, for $t, l \in \mathscr{T}$, as

$$\hat{\mu}^{ols}_{t|l} = \bar{X}_l' \hat{\beta}_t$$

where $\bar{X}_l = \sum_{i=1}^{N} D_i(l) X_i \Big/ N_l$ and where $\hat{\beta}_t$ is the OLS estimator of the outcome regression model parameter β_t in Assumption 4 using observations where $D(t) = 1$. We define an estimator $\hat{\gamma}^{ols}_{lm|l}$ for the parameter $\gamma_{lm|l}$ as

$$\hat{\gamma}^{ols}_{lm|l} = \hat{\mu}^{ols}_{l|l} - \hat{\mu}^{ols}_{m|l} = \bar{X}_l'(\hat{\beta}_l - \hat{\beta}_m).$$

The estimating function corresponding to $\hat{\gamma}^{ols}_{lm|l}$, $\hat{\beta}_l$, and $\hat{\beta}_m$, as defined above, can be written as

$$M^{ols}(O; \gamma_{lm|l}, \beta_l, \beta_m) = \begin{bmatrix} M^{ols-att}(O; \gamma_{lm|l}, \beta_l, \beta_m) \\ M^{ols-l}(O; \beta_l) \\ M^{ols-m}(O; \beta_m) \end{bmatrix},$$

where

$$M^{ols-att}(O; \gamma_{lm|l}, \beta_l, \beta_m) = D(l) \left[\gamma_{lm|l} - X'(\beta_l - \beta_m) \right]$$

and where

$$M^{ols-t}(O; \beta_t) = D(t) X \left[Y - X' \beta_t \right]$$

are the OLS estimating functions for β_t, for $t = l, m$. As shown in Section 4, the solution $(\hat{\gamma}^{ols}_{lm|l}, \hat{\beta}_l, \hat{\beta}_m)$ to the estimating equation

$$\sum_{i=1}^{N} M^{ols}(O_i; \gamma_{lm|l}, \beta_l, \beta_m) = 0 \tag{1}$$

is a consistent and asymptotically normally distributed estimator of the true parameter vector $(\gamma_{lm|l}, \beta_l, \beta_m)$ indexing the distribution for the data generating process when the outcome model in Assumption 4 is correctly specified.

Next, we define weights

$$w(t,l) = D(t)\frac{r(l,X)}{r(t,X)} \tag{2}$$

based on the working model in Assumption 5 for the GPS. Using (2), we can now define an IPW-based estimator

$$\hat{\mu}^{ipw}_{t|l} = \frac{1}{N_l}\sum_{i=1}^{N} \hat{w}_i(t,l)Y_i,$$

of $\mu_{t|l}$, for $t \in \{l, m\}$, where

$$\hat{w}_i(t,l) = D_i(t)\frac{\hat{r}(l,X_i)}{\hat{r}(t,X_i)}$$

are estimated weights based on maximum likelihood estimators $\hat{r}(l,X_i)$ and $\hat{r}(t,X_i)$ of the GPS $r(l,X_i)$ and $r(t,X_i)$ as defined in Assumption 5, for $t \in \{l, m\}$. We note that $w_i(l,l) = D_i(l)$. Given $\hat{\mu}^{ipw}_{m|l}$ and $\hat{\mu}^{ipw}_{l|l}$, an IPW estimator of the average treatment effect among the treated $\gamma_{lm|l}$ can be written as

$$\hat{\gamma}^{ipw}_{lm|l} = \hat{\mu}^{ipw}_{l|l} - \hat{\mu}^{ipw}_{m|l}.$$

The estimating function corresponding to $\hat{\gamma}^{ipw}_{lm|l}$ can be written as

$$M^{ipw}(O; \gamma_{lm|l}, \delta) = \begin{bmatrix} M^{ipw-att}(O; \gamma_{lm|l}, \delta) \\ M^{gps}(O; \delta) \end{bmatrix},$$

where

$$M^{ipw-att}(O;\gamma_{lm|l},\delta) = D(l)\gamma_{lm|l} - D(l)Y + w(m,l)Y, \tag{3}$$

where $w(m,l)$ is defined as in (2) with $t = m$, and where

$$M^{gps}(O;\delta) = \begin{bmatrix} X\left[D(1) - r(1,X)\right] \\ \vdots \\ X\left[D(K) - r(K,X)\right] \end{bmatrix} \tag{4}$$

is the maximum likelihood score function for δ in the GPS-model in Assumption 5. In Section 4, we show that the solution $(\hat{\gamma}^{ipw}_{lm|l}, \hat{\delta})$ to the estimating equation

$$\sum_{i=1}^{N} M^{ipw}(O_i;\gamma_{lm|l},\delta) = 0, \tag{5}$$

is a consistent estimator of the true parameters indexing the distribution for the data generating process when the GPS model in Assumption 5 is correctly specified.

Finally, we define the doubly robust estimator, WOLS, of $\gamma_{lm|l}$. The estimator is obtained using weighted least squares estimators $\hat{\beta}^*_l$ and $\hat{\beta}^*_m$ of β_l and β_m, respectively. See also Equation 23 in Uysal (2015). We define a doubly robust estimator of $\gamma_{lm|l}$ as

$$\hat{\gamma}^{wols}_{lm|l} = \bar{X}_l'(\hat{\beta}^*_l - \hat{\beta}^*_m).$$

The full estimating function corresponding to $\hat{\gamma}^{wols}_{lm|l}$ can be written as

$$M^{wols}(O;\gamma_{lm|l},\beta_l,\beta_m,\delta) = \begin{bmatrix} M^{wols-att}(O;\gamma_{lm|l},\beta_l,\beta_m) \\ M^{wols-l}(O;\beta_l,\delta) \\ M^{wols-m}(O;\beta_m,\delta) \\ M^{gps}(O;\delta) \end{bmatrix},$$

where

$$M^{wols-att}(O;\gamma_{lm|l},\beta_l,\beta_m) = D(l)\left[\gamma_{lm|l} - X'(\beta_l - \beta_m)\right]$$

is the estimating function for $\gamma_{lm|l}$,

$$M^{wols-t}(\beta_t, \delta) = Xw(t,l)\left(Y - X'\beta_t\right)$$

is the WOLS estimating function for β_t, $t = l, m$ and $M^{gps}(\delta)$ is the ML score function for δ as defined in (4).

As demonstrated in Uysal (2015), the estimator $\hat{\gamma}^{wols}_{lm|l}$, obtained as the first element of the solution $(\hat{\gamma}^{wols}_{lm|l}, \hat{\beta}^{*'}_l, \hat{\beta}^{*'}_m, \hat{\delta}')'$ to the estimating equation

$$\sum_{i=1}^{N} M^{wols}(O_i; \gamma_{lm|l}, \beta_l, \beta_m, \delta) = 0, \tag{6}$$

is a doubly robust estimator of $\mathrm{E}[Y(l) - Y(m)|T = l]$. This means that, given that assumptions 1, 2 and 3 hold $\hat{\gamma}^{wols}_{lm|l}$ is consistent for $\mathrm{E}[Y(l) - Y(m)|T = l]$ when one of the model assumptions 4 and 5 holds, not necessarily both.

4 Large sample properties of the estimators

In order to show that the estimators are consistent and asymptotically normal, we utilize theory of M-estimation (Boos and Stefanski, 2013).

According to standard theory of M-estimation, given an estimating function $M(O;\theta)$, observations $O_1, \ldots, O_N$, and a distribution P_O such that the solution θ_0 to the equation

$$\int M(o;\theta) dP_O(o) = 0 \tag{7}$$

exists and is unique, the solution $\hat{\theta}$ to the estimating equation

$$\sum_{i=1}^{N} M(O_i;\theta) = 0 \tag{8}$$

is a consistent estimator of θ_0 when P_O is the true distribution underlying the observed data. Further, $\sqrt{N}(\hat{\theta} - \theta_0)$ is asymptotically normal with variance Σ that can be consis-

tently estimated by

$$\widehat{\Sigma}_N = \left[A_N(\hat{\theta})\right]^{-1} B_N(\hat{\theta}) \left\{ \left[A_N(\hat{\theta})\right]^{-1} \right\}', \tag{9}$$

where

$$A_N(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \left. \frac{\partial M(O_i;\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} \tag{10}$$

and

$$B_N(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} M(O_i;\hat{\theta}) M(O_i;\hat{\theta})'. \tag{11}$$

The expression $\widehat{\Sigma}_N/N$ is usually referred to as the sandwich estimator of the variance of $\hat{\theta}$. We note that this variance estimator is correct for the true value of the variance of $\hat{\theta}$, assuming that there exists a unique solution to

$$\mathrm{E}\left[M(O;\theta)\right] = \int M(o;\theta) dP_O(o) = 0. \tag{12}$$

This is then true regardless of what model assumptions the estimating functions M are based on. When one or more of the model assumptions are incorrect, we may not be able to interpret $\hat{\theta}$ as an estimator of a parameter in our assumed model. However, the variance estimator in (9) is still consistent for Σ.

In the next section, we verify consistency for the estimators $\hat{\gamma}^{ols}_{lm|l}$, $\hat{\gamma}^{ipw}_{lm|l}$, and $\hat{\gamma}^{wols}_{lm|l}$, defined as first elements of the solutions to estimating equations of the form (8), by verifying that the corresponding equations of the form (7) hold when P_O, indexed by θ, is the distribution for the data generating process. Although consistency of $\hat{\gamma}^{ols}_{lm|l}$, $\hat{\gamma}^{ipw}_{lm|l}$, and $\hat{\gamma}^{wols}_{lm|l}$, have been shown elsewhere, no expression for the asymptotic variance of $\hat{\gamma}^{ipw}_{lm|l}$, and $\hat{\gamma}^{wols}_{lm|l}$ have previously been presented. By stacking estimating equations corresponding to partial M-estimators, we derive analytical expression for the asymptotic variance of the estimators.

4.1 Consistency of $\hat{\gamma}^{ols}_{lm|l}$, $\hat{\gamma}^{ipw}_{lm|l}$, and $\hat{\gamma}^{wols}_{lm|l}$

Throughout this section, we assume that Assumptions 2 and 3 hold. We first assume that the outcome regression model assumption in 4 is correctly specified with true parameters

β_l and β_m. It now follows from standard theory of least squares estimation that

$$\mathrm{E}\left[M^{ols-t}(O;\beta_t)\right] = \int M^{ols-t}(o;\beta_t)dP_O(o) = 0$$

for any distribution P_O such that Assumption 4 hold for $t = l, m$. It also follows from Assumption 4 that

$$\mathrm{E}\left[D(l)X'\beta_t\right] = \Pr(T=l)\,\mathrm{E}\left(X'\beta_t|T=l\right) = \Pr(T=l)\,\mathrm{E}\left[Y(t)|T=l\right]$$

for $t \in \mathscr{T}$. It follows that

$$\mathrm{E}\left[M^{ols-att}(O;\gamma_{lm|l},\beta_l,\beta_m)\right] = \Pr(T=l)\left\{\gamma_{lm|l} - \mathrm{E}\left[Y(l)-Y(m)|T=l\right]\right\} = 0,$$

when $\gamma_{lm|l}$ is the true parameter. We conclude that

$$\mathrm{E}[M^{ols}(O;\gamma_{lm|l},\beta_l,\beta_m) = 0.$$

Therefore, the solution $(\hat{\gamma}^{ols}_{lm|l},\hat{\beta}_l,\hat{\beta}_m)$ to the estimating equations

$$\sum_{i=1}^{N} M^{ols}(O_i;\gamma_{lm|l},\beta_l,\beta_m) = 0$$

is a consistent estimator for the true parameter vector $(\gamma_{lm|l},\beta_l,\beta_m)$ for the data generating process. Thus, the estimator $\hat{\gamma}^{ols}_{lm|l}$ is consistent for $\gamma_{lm|l}$.

Next, we assume that the distribution P_O is such that the GPS model in Assumption 5 is correctly specified, with δ being the true parameter vector for the GPS. It follows from standard theory of maximum likelihood estimation that

$$\mathrm{E}\left[M^{gps}(O;\delta)\right] = \int M^{gps}(o;\delta)dP_O(o) = 0.$$

In Appendix A, we further demonstrate that $\mathrm{E}\left[M^{ipw-att}(O;\gamma_{lm|l},\delta)\right]=0$ when $(\gamma_{lm|l},\delta)$ is the true parameters for the data generating process. We conclude that the solution to $(\hat{\gamma}^{ipw}_{lm|l},\hat{\delta})$ to the estimating equations

$$\sum_{i=1}^{N} M^{ipw}(O_i;\gamma_{lm|l},\delta)=0$$

is consistent for the true value of the parameter vector $(\gamma_{lm|l},\delta)$ for the data generating process.

As shown in Appendix A,

$$\mathrm{E}\left[M^{wols-att}\left(O;\gamma_{lm|l},\beta_l,\beta_m,\delta\right)\right]=0$$

when the GPS model in Assumption 5 is correct with true parameter value δ or when the outcome model in Assumption 4 is correctly specified with true parameter values β_l and β_m. Under mild conditions the equation

$$\mathrm{E}\left[M^{ols-t}(O;\beta_t)\right]=0$$

has a solution β_t^*, for $t=l,m$, regardless of whether the outcome model Assumption 4 is correct or not. Likewise, the equation

$$\mathrm{E}\left[M^{gps}(O;\delta)\right]=0$$

has a solution δ^*, regardless of whether the GPS model Assumption 5 is correct or not. It follows that the first component $\hat{\gamma}^{wols}_{lm|l}$ of the solution $(\hat{\gamma}^{wols}_{lm|l},\hat{\beta}^*_l,\hat{\beta}^*_m,\hat{\delta})$ to the estimating equations

$$\sum_{i=1}^{N} M(O_i;\gamma_{lm|l},\beta_l,\beta_m,\delta)=0$$

is consistent for the true value of the ATT parameter $\gamma_{lm|l}$ when at least one of the models in Assumptions 4 and 5 is true, not necessarily both. We conclude that $\hat{\gamma}^{wols}_{lm|l}$ is doubly

robust.

4.2 Asymptotic variance

In order to derive the asymptotic variances of the estimators $\hat{\gamma}^{ols}_{lm|l}$, $\hat{\gamma}^{ipw}_{lm|l}$, and $\hat{\gamma}^{wols}_{lm|l}$, we need the gradients of (1), (5), and (6), with respect to their parameters.

The asymptotic variance of $\hat{\gamma}^{ols}_{lm|l}$

The gradient of $M^{ols}(O_i;\gamma_{lm|l},\beta_l,\beta_m)$ with respect to the vector $(\gamma_{lm|l},\beta_l',\beta_m')$ can be written as

$$\frac{\partial M^{ols}(O_i;\gamma_{lm|l},\beta_l,\beta_m)}{\partial(\gamma_{lm|l},\beta_l',\beta_m')} = \begin{bmatrix} D_i(l) & -D_i(l)X_i' & D_i(l)X_i' \\ 0 & -D_i(l)X_iX_i' & 0 \\ 0 & 0 & -D_i(m)X_iX_i' \end{bmatrix}.$$

By substituting M^{ols} for M, $(\gamma_{lm|l},\beta_l',\beta_m')$ for θ', and $(\hat{\gamma}_{lm|l},\hat{\beta}_l',\hat{\beta}_m')$ for $\hat{\theta}$ in (10) and (11), an estimator of the variance of $\hat{\gamma}^{ols}_{lm|l}$ can be obtained as the first diagonal element of $\hat{\Sigma}_N/N$, where $\hat{\Sigma}_N$ is defined as in (9).

The asymptotic variance of $\hat{\gamma}^{ipw}_{lm|l}$

The gradient of $M^{ipw}(O_i;\gamma_{lm|l},\delta)$ with respect to $(\gamma_{lm|l},\delta')$ can be written as

$$\frac{\partial M^{ipw}(O_i;\gamma_{lm|l},\delta)}{\partial(\gamma_{lm|l},\delta')} = \begin{bmatrix} D_i(l) & \frac{\partial M^{ipw-att}(O_i;\gamma_{lm|l},\delta)}{\partial\delta'} \\ 0 & \frac{\partial M^{gps}(O_i;\delta)}{\partial\delta'} \end{bmatrix}, \tag{13}$$

where

$$\frac{\partial M^{ipw-att}(O_i;\gamma_{lm|l},\delta)}{\partial\delta'} = \begin{bmatrix} \frac{\partial M^{ipw-att}(O_i;\gamma_{lm|l},\delta)}{\partial\delta_1'} & \cdots & \frac{\partial M_i^{ipw-att}(O_i;\gamma_{lm|l},\delta)}{\partial\delta_K'} \end{bmatrix}$$

with

$$\frac{\partial M^{ipw-att}(O_i;\gamma_{lm|l},\delta)}{\partial\delta_s'} = \begin{cases} X_i'w_i(m,l)Y_i & \text{for } s=l \\ -X_i'w_i(m,l)Y_i & \text{for } s=m \\ 0 \text{ otherwise} \end{cases},$$

for $s\in\mathscr{T}$ and where $\frac{\partial M^{gps}(O_i;\delta)}{\partial\delta'}$ is the gradient of the estimating function corresponding to the GPS model with respect to the GPS parameter δ. In Appendix A, we derive an

expression for $\frac{\partial M^{gps}(O_i;\delta)}{\partial \hat{\delta}'}$ when M^{gps} is the maximum likelihood score function for the multinomial model defined in Assumption 5. By substituting M^{ipw} for M, $(\gamma_{lm|l}, \delta')$ for θ', and $(\hat{\gamma}_{lm|l}, \hat{\delta}')$ for $\hat{\theta}$ in (10) and (11), an estimator of the variance of $\hat{\gamma}^{ipw}_{lm|l}$ can be obtained as the first diagonal element of $\hat{\Sigma}_N/N$, where $\hat{\Sigma}_N$ is defined as in (9).

The asymptotic variance of $\hat{\gamma}^{wols}_{lm|l}$

The gradient of $M^{wols}(O_i;\gamma_{lm|l},\beta_l,\beta_m,\delta)$ with respect to $(\gamma_{lm|l},\beta_l',\beta_m',\delta')$ can be written as

$$\frac{\partial M^{wols}(O_i;\gamma_{lm|l},\beta_l,\beta_m,\delta)}{\partial(\gamma_{lm|l},\beta_l',\beta_m',\delta')} = \begin{bmatrix} D_i(l) & -D_i(l)X_i' & D_i(l)X_i' & 0 \\ 0 & -D_i(l)X_iX_i' & 0 & 0 \\ 0 & 0 & -w_i(m,l)X_iX_i' & \frac{\partial M^{wols-m}(O_i;\gamma_{lm|l},\beta_l,\beta_m,\delta)}{\partial \delta'} \\ 0 & 0 & 0 & \frac{\partial M^{gps}(O_i;\delta)}{\partial \delta'} \end{bmatrix} \quad (14)$$

where

$$\frac{\partial M^{wols-m}(O_i;\gamma_{lm|l},\beta_l,\beta_m,\delta)}{\partial \delta_s'} = \begin{cases} w_i(m,l)X_iX_i'(Y_i - X_i'\beta_m) & \text{when } s = l \\ -w_i(m,l)X_iX_i'(Y_i - X_i'\beta_m) & \text{when } s = m\,, \\ 0 & \text{otherwise} \end{cases}$$

for $s \in \{1,\ldots,K\}$. As for the IPW estimator $\hat{\gamma}^{ipw}_{lm|l}$, $M^{gps}(O_i;\delta)$ is the maximum likelihood score function based on the GPS model in Assumption 5. By substituting M^{wols} for M, $(\gamma_{lm|l},\beta_l',\beta_m',\delta')$ for θ', and $(\hat{\gamma}_{lm|l},\hat{\beta}_l',\hat{\beta}_m',\hat{\delta}')$ for $\hat{\theta}$ in (10) and (11), an estimator of the variance of $\hat{\gamma}^{wols}_{lm|l}$ can be obtained as the first diagonal element of $\hat{\Sigma}_N/N$, where $\hat{\Sigma}_N$ is defined as in (9).

5 Simulation studies

To assess the properties of the WOLS estimator and the sandwich estimator of its variance, we performed two simulation studies comparing the WOLS estimator with the IPW

and the OLS estimator. For the purpose of transparency, we used a study design from a previous study (Yang et al., 2016).

5.1 Model for the pretreatment variables

In both simulation studies, the pretreatment variables were simulated under the model

$$\begin{aligned}
(X_1, X_2, X_3) &\sim \text{MVN}(0, \Sigma) \\
\Sigma &= \begin{bmatrix} 2 & 1 & -1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{bmatrix} \\
X_4 &\sim \text{U}(-3, 3) \\
X_5 &\sim \chi^2(1) \\
X_6 &\sim \text{Be}(0.5) \\
X &= (1, X_1, X_2, X_3, X_4, X_5, X_6)
\end{aligned}$$

Treatment levels were then simulated according to a multinomial model

$$\begin{aligned}
T &\sim \text{Multinom}[p(t_0|X), p(t_2|X), \ldots, p(t_K|X)] \\
p(t|X) &= \exp(X'\delta_t) / \textstyle\sum_{s=0}^{K} \exp(X'\delta_s) \text{ for } t = 0, 1, \ldots, K
\end{aligned}$$

for coefficient vectors $\delta_0, \delta_1, \ldots, \delta_K$, with $\delta_0 \equiv 0$. For each treatment level t, we also simulated a potential outcome

$$\begin{aligned}
Y(t) &= X'\beta_t + \varepsilon \\
\varepsilon &\sim \text{N}(0, 1),
\end{aligned}$$

for some parameter vectors $\beta_0, \ldots, \beta_K$. We let $Y = \sum_{t=0}^{K} I\{T = t\} Y(t)$ represent the observed outcome. For each sample, the ATT parameter of interest, $\gamma_{lm|l}$ was calculated using OLS, IPW and WOLS. For IPW and WOLS, we either used the correct working models

$$\delta_{t0} + \delta_{t1} X_1 + \delta_{t2} X_2 + \delta_{t3} X_3 + \delta_{t4} X_4 + \delta_{t5} X_5 + \delta_{t6} X_6$$

or the incorrect working models

$$\delta_{t0} + \delta_{t1}X_1 + \delta_{t2}X_2 + \delta_{t3}X_3 + \delta_{t5}X_5$$

for the linear predictor in the multinomial distribution for each treatment level t. Similarly, for OLS and WOLS, we either used the correct working models

$$\beta_{t0} + \beta_{t1}X_1 + \beta_{t2}X_2 + \beta_{t3}X_3 + \beta_{t4}X_4 + \beta_{t5}X_5 + \beta_{t6}X_6$$

or the incorrect working models

$$\beta_{t0} + \beta_{t1}X_1 + \beta_{t2}X_2 + \beta_{t3}X_3 + \beta_{t5}X_5$$

for the linear predictor in the outcome regressions models for each treatment level t. In both misspecified models, we thus left out the uniformly distributed covariate X_4 and the Bernoulli distributed covariate X_6. Since both of these variables are uncorrelated with the other covariates, we expect this misspecification to cause biased estimates of $\gamma_{lm|l}$. Since we are replicating previous simulation studies containing only first order terms in the data generating process, we choosed to misspecifiy the working models by omitting confounding variables instead of misspecifying the functional forms of the models, thereby generating more substantial bias when the working models are misspecified. For each combination of working models and estimation methods the mean bias, the mean estimated standard error, and the root mean sqare error (RMSE) of $\hat{\gamma}_{lm|l}$ were calculated over the samples. For comparison, the nonparametric bootstrap standard error of of $\hat{\gamma}_{lm|l}$ was calculated for each sample based on 500 bootstrap replicates for each sample. We used the empirical standard errors calculated over the simulations as approximations of the true standard errors, and compared these values with the mean estimated and the mean bootstrap standard errors. For each sample and each model, a Wald-based 95% confidence interval was calculated using both the sandwich and the bootstrap estimates of the standard error. The coverage probabilities were calculated for the two kinds of confidence intervals. In the calculations of the RMSE and the coverage probabilities, the true value

of $\gamma_{lm|l}$ was approximated by using simulated values of $Y(l)$ and $Y(m)$ among those with treatment level $T = l$ in a sample with 10,000,000 observations.

5.2 Simulation study 1

In the first simulation study, we simulated three treatment levels according to the multinomial model

$$\begin{aligned} T &\sim \text{Multinom}[p(0|X), p(1|X), p(2|X)] \\ p(t|X) &= \exp(X'\delta_t)/\textstyle\sum_{s=0}^{2}\exp(X'\delta_s) \text{ for } t = 0,1,2 \\ \delta_0 &= (0,0,0,0,0,0,0) \\ \delta_1 &= 0.7 \times (0,1,1,1,-1,1,1) \\ \delta_2 &= 0.4 \times (0,1,1,1,1,1,1). \end{aligned}$$

In contrast to the simulation study by Yang et al. (2016), we did not fix the numbers of treated for each treatment level. Instead, the proportions were determined from the pretreatment variables X and the conditional distribution of T given X. When using the model in Subsection 5.1 for the pretreatment variables $X_1, \ldots, X_6$, we get the marginal probabilities $p(T = 0) \approx 0.19$, $p(T = 1) \approx 0.48$ and $p(T = 2) \approx 0.33$.
The potential outcomes were simulated using the linear model

$$\begin{aligned} Y(t) &= X'\beta_t + \varepsilon \\ \varepsilon &\sim \text{N}(0,1) \\ \beta_0 &= (-1.5,1,1,1,1,1,1) \\ \beta_1 &= (-3,2,3,1,2,2,2) \\ \beta_2 &= (1.5,3,1,2,-1,-1,-1). \end{aligned}$$

This gave us the theoretical values

$$\begin{aligned} \gamma_{10|1} = \text{E}[Y(1) - Y(0)|T = 1] &\approx -0.00970 \\ \gamma_{20|2} = \text{E}[Y(2) - Y(0)|T = 2] &\approx -2.35 \\ \gamma_{21|2} = \text{E}[Y(2) - Y(1)|T = 2] &\approx -3.55 \end{aligned}$$

of the ATT parameters. For each observation, we let $Y = \sum_{t=0}^{2} I\{T = t\}Y(t)$ represent the observed outcome. Using this design 1000 samples of 1000 observations were generated. For each sample we estimated $\gamma_{10|1}$, $\gamma_{20|2}$, and $\gamma_{21|2}$ using OLS, IPW and WOLS using both correctly and incorrectly specified models. The results are shown in Tables 1, 3, and 5. When the outcome model is correctly specified, the OLS estimators of the ATT parameters $\gamma_{10|1}$, $\gamma_{20|2}$, and $\gamma_{21|2}$ are unbiased, which we see in the first row in Tables 1, 3, and 5. The mean estimates of the standard errors obtained from the sandwich estimator are close to the mean bootstrap standard errors and the empirical standard errors. This results in coverage probabilities close to 0.95 and a low RMSE for all three ATT parameter estimates. When the working model for the outcome is misspecified, the parameter estimates from OLS are severely biased. The estimated standard errors are still close to the empirical and the bootstrap standard errors. However, a consequence of the biases is small coverage probabilities for the CI:s and high RMSE:s.

When both the exposure model and the outcome model are correctly specified, the IPW estimator of the ATT parameters is more biased compared to the OLS estimator, particularly for the smaller sample size. Further, the standard errors from the sandwich estimator is substantially smaller than both the bootstrap and the empirical standard error, indicating that the sandwich estimator might yield optimistic inference. This is most clearly seen in Tables 1 and 3, where the biases are larger for the IPW estimator compared to the OLS estimator. Since the estimated standard error is also underestimated, this results in confidence intervals with less than 95% coverage. Interestingly, the OLS and IPW estimates are more similar when their respective models are incorrectly specified.

When both the exposure and outcome models are correctly specified, the bias of the WOLS estimator is similar to the bias obtained from OLS. We see only a slight increase in RMSE. The mean standard error of $\hat{\gamma}_{10|1}$ is slightly overestimated by the sandwich estimator, but remains close to both the mean bootstrap and the empirical standard error for $\hat{\gamma}_{20|2}$ and $\hat{\gamma}_{21|2}$. Thus the coverage probabilities of the 95% CI:s is also close to the nominal values.

The same is true when only the outcome model is correct, with similar results as when both models are correct. Thus the misspecification of the exposure model only had a minor impact on the bias and the variance.

In contrast, when only the exposure model is correctly specified, we see an increase in bias of both the ATT estimates and their variances for WOLS. The results are slightly worse than the corresponding results using IPW with correct exposure model. However, since the standard error of the WOLS estimator is smaller than the IPW standard error, we see a decrease in RMSE for the WOLS estimator compared to the IPW estimator in this case.

When both working models are incorrectly specified, the bias of the WOLS estimator for the ATT parameters are similar to the corresponding ones for OLS and IPW.

To see in what extent the observed biases in the estimates of the ATT parameters and their standard errors decreased with larger sample size, we repeated simulation study 1 with $n = 5000$. The results are shown in Tables 2, 4, and 6. The larger sample sizes contributed to smaller mean biases and smaller empirical standard errors, resulting in smaller RMSE for all estimators.

Again, we see that the results for OLS and WOLS are very similar when the outcome working model is correctly specified. The RMSE:s for the WOLS ATT estimates are only slightly larger compared to the corresponding OLS ATT estimates. Both the sandwich and the bootstrap estimates of the standard errors are close to the empirical standard error. Thus the coverage probabilities are close to 95%.

Compared to the smaller sample size, the bias of the IPW estimator is reduced when the exposure model is correctly specified. The sandwich estimates of the standard errors remain underestimated, resulting in coverage probability lower than 95%. This is most

prominent in Table 2 for $\gamma_{10|1}$. However, the sandwich estimates of the standard error are similar to the bootstrap estimates.

When only the exposure model is correctly specified, also the WOLS estimator is less biased. Since this estimator has lower standard error compared to the IPW estimator, the RMSE is also smaller. However, the bias of the sandwich estimator remains. Therefore the coverage probability of a 95% CI is smaller than 95%.

When both working models are incorrect, the bias and RMSE is similar for all estimators of the ATT parameters.

The instability of IPW-based estimators have been noted by several authors (Little and Rubin, 2019; Lefebvre et al., 2008; Cole and Hernán, 2008). This is mainly caused by few observations in some strata, leading to small values of the propensity scores for some strata, which in turn leads to large IPW weights for these observations. These observations will therefore have a large impact on the estimate. The IPW and WOLS estimators of the ATT parameters and their standard errors depend on the weights

$$w(m,l) = D(m)\frac{r(l,X)}{r(m,X)}$$

for $m = 0,1,2$ and reference levels $l = 1,2$. Observations with $T = l$ will have weights $w(l,l) = 1$, regardless of whether the working exposure model is correct or not. Observations with $T = 0$ will have weights that depend on the working exposure model as well as the estimated values of $r(l,X)$ and $r(m,X)$. Observations where $r(l,X)$ is small and where $r(m,X)$ is large will have a small impact on the ATT estimates and their standard errors. Observations where $r(l,X)$ is large and where $r(m,X)$ is small will have a large impact on the ATT estimates and theirs standard errors. This has a direct analogue to propensity scores close to 0 when estimating the ATE and when the treatment is binary. In Figures 1, 2, and 3 we see the distribution of the logarithms of weights $\hat{w}(0,1)$, $\hat{w}(0,2)$, and $\hat{w}(1,2)$ used in the IPW and WOLS estimators of $\gamma_{10|1}$, $\gamma_{20|2}$, and $\gamma_{21|2}$, respectively,

when the exposure model is correctly specified. Since some weights had very large values, the logarithms of the weights were used instead. Only non-zero values of $\hat{w}(l,m)$ were exluded. We see a few influential values in the graphs, particularly in Figure 1. The largest value of $\hat{w}(1,0)$ is 137, corresponding to a value of 4.9 for $\log[\hat{w}(1,0)]$. We also observe some influential values for $\log[\hat{w}(2,0)]$ and $\log[\hat{w}(2,1)]$ in Figures 2 and 3, respectively. In the histograms, we see a lot of small values, particularly in Figures 2 and 3. These observations have very little impact on the IPW and WOLS estimates of the ATT parameters. It is also of interest how misspecification of the working exposure model influences the weights, since such weights are used for the estimation of some values in Tables 1-6. In particular, such weights will affect the properties of the WOLS estimator, which is consistent as long as the working outcome model is correct. Therefore, we also plotted the distribution of the logarithms of the weights under exposure model misspecification in Figures 4, 5, and 6. The results are similar to the ones in Figures 1, 2, and 3, where a number of large influential values are seen for $\log[\hat{w}(0,1)]$ and $\log[\hat{w}(0,2)]$.

The weights $w(m,l)$ depend on the conditional probabilities $\Pr(T = m|T \in \{m,l\},X)$. In particular, the IPW estimator is expected to be sensitive to values of $\Pr(T = m|T \in \{m,l\},X)$ close to 0. Therefore, in Appendix B, we have added plots of estimates $\hat{\Pr}(T = m|T \in \{l,m\},X)$ of these probabilities, along with their complements $\hat{\Pr}(T = l|T \in \{l,m\},X)$, for each ATT parameter $\gamma_{lm|l}$ using either the correct or the incorrect working model for the exposure, see Figures 11, 12, and 13. Since only individuals with $T = m$ contribute to the estimators, we only show these observations. In Figures 11 and 12, we observe a satisfying overlap, corresponding to the contrasts $\gamma_{10|1}$ and $\gamma_{20|2}$. In contrast, Figure 13 shows a substantial lack of overlap when the correct working model for the GPS is used. In particular, the probability of having treatment level $T = 2$ is low compared to the probability of having treatment level $T = 1$. A consequence of this is that the parameter estimates of β_2 is based on fewer observations. This is reflected in the relatively high variance for the OLS and WOLS estimates when the outcome regression model is correct. It is also reflected in the high variance of the IPW estimates when the exposure model is correct. For all estimates, it leads to more uncertainty in the estimation of $\mu_{2|2}$ and also to more

uncertainty in the estimation of $\gamma_{21|2}$. When the GPS model is misspecified, no lack of overlap can be seen in Figure 13. In this case, the misspecification leads to more stable estimation when using WOLS. The lack of overlap is a feature of the simulation design by Yang et al. (2016) and gives us a hint on what to expect when this lack of overlap is seen in applications.

5.3 Simulation study 2

In the second simulation study, again replicating Yang et al. (2016), we used the same distribution for the six pretreatment variables, but doubled the number of treatment levels to six according to the model

$$\begin{aligned}
T &\sim \text{Multinom}[p(0|X), p(1|X), p(2|X), p(3|X), p(4|X), p(5|X)] \\
p(t|X) &= \exp(X'\delta_t)/\textstyle\sum_{s=0}^{5}\exp(X'\delta_s) \quad \text{for} \quad t = 0,\ldots,5 \\
\delta_0 &= 0 \\
\delta_1 &= 0.4 \times (0,1,1,2,1,1,1) \\
\delta_2 &= 0.6 \times (0,1,1,1,1,1,-5) \\
\delta_3 &= 0.8 \times (0,1,1,1,1,1,5) \\
\delta_4 &= 1.0 \times (0,1,1,1,-2,1,1) \\
\delta_5 &= 1.2 \times (0,1,1,1,-2,-1,1).
\end{aligned}$$

As in Simulation study 1, we let the distribution of T be determined by the pretreatment variables X and the conditional distribution of T given X, thus again deviating slightly from the simulation design by Yang et al. (2016). When using the model in Subsection 5.1 for the pretreatment variables $X_1,\ldots,X_6$, we get the marginal probabilities $p(T = 0) \approx 0.048$, $p(T = 1) \approx 0.059$, $p(T = 2) \approx 0.067$, $p(T = 3) \approx 0.42$, $p(T = 4) \approx 0.25$, and $p(T = 5) \approx 0.15$.

Potential outcomes $Y(0),\ldots,Y(5)$ where generated according to the model

$$\begin{aligned}
Y(t) &= X'\beta_t + \varepsilon \quad \text{for} \quad t = 0,\ldots,5 \\
\varepsilon &\sim N(0,1) \\
\beta_0 &= (-1.5,1,1,1,1,1,1) \\
\beta_1 &= (-3,2,3,1,2,2,2) \\
\beta_2 &= (3,3,1,2,-1,-1,-4) \\
\beta_3 &= (2.5,4,1,2,-1,-1,-3) \\
\beta_4 &= (2,5,1,2,-1,-1,-2) \\
\beta_5 &= (1.5,6,1,2,-1,-1,-1).
\end{aligned}$$

We focused on the two ATT parameters $\gamma_{10|1}$ and $\gamma_{54|5}$, where the overlap were best and where there were least influential observations as displayed in the plots of log weights $\log[\hat{w}(m,l)]$. The theoretical values of these two parameters are

$$\begin{aligned}
\gamma_{10|1} = \mathrm{E}[Y(1)-Y(0)|T=1] &\approx -1.40 \\
\gamma_{54|5} = \mathrm{E}[Y(5)-Y(4)|T=5] &\approx 0.279.
\end{aligned}$$

Observed outcomes were defined as $Y = \sum_{s=0}^{5} I\{T=s\}Y(s)$. Using this design, 1000 samples of 1000 observations were generated. For each sample, these parameters were estimated using OLS, IPW and WOLS with both correctly and incorrectly specified models. The results are shown in Tables 7 and 9. When both working models are correct, the OLS and WOLS estimators have similar bias and RMSE. The smallest RMSE:s are observed for the OLS estimator. The IPW estimator displays the greatest bias and variability, although the performance is not too far from the results from OLS and WOLS estimation. The sandwich estimates of the standard error are close to the empirical standard error. In this scenario, the sandwich estimator performs slightly better than the bootstrap estimator.

When only the exposure working model is misspecified, the IPW estimators are biased. However, the standard errors are close to the empirical standard error. In contrast, the WOLS estimators is largely unaffected by this misspecification, although we see some

bias in the sandwich estimate of the standard error. Thus the coverage probability differ from the nominal value of 95%.

Misspecification of the outcome working model leads to bias for the OLS estimator of the ATT parameters. In contrast, the sandwich estimates of the standard error is close to the empirical standard errors. The WOLS estimators, which are assumed to be consistent for the ATT estimators in this scenario, displays only slightly more bias than when the outcome model is correctly specified. The sandwich estimate of the standard error of $\hat{\gamma}_{10|1}$ is slightly above the empirical standard error.

When both working models are incorrect, the WOLS estimators have more bias, although these biases are smaller in magnitude than the bias for the IPW estimators. This is particularly apparent for $\gamma_{54|5}$ in Table 9. We also see some bias for the sandwich estimators of the standard errors.

Since we allow more levels for the treatment in simulation study 2, three times more parameters are estimated for the same sample size compared to simulation study 1. This likely affects the precision of the estimators. To assess whether the observed bias can be attributed to small sample bias, we increased the sample size to 5000 observations in each sample. The results are shown in Tables 8 and 10. As expected, the biases are also reduced for all estimators when both working models are correctly specified. The least reduction in bias can be seen for the WOLS estimates of $\gamma_{10|1}$, where the biases with the larger sample sizes are about one fourth of the corresponding ones for the smaller sample sizes. decreasing from 0.086 with the smaller sample size to 0.020 with the larger sample size. Since the estimators are consistent, the empirical standard errors were reduced. Compared to the estimates based on the smaller sample size the empirical standard errors and the RMSE:s were more than halved. The observed biases of the sandwich estimators of the standard errors of $\hat{\gamma}_{10|1}$ seen with $n = 1000$, were greatly reduced with $n = 5000$. However, no improvements were observed for the estimators of the standard errors of $\hat{\gamma}_{54|5}$. Likewise, no substantial differences for either ATT parameter were seen regarding

the coverage probabilities for the 95% CI when comparing the results with the two sample sizes.

When only the working outcome model was misspecified, we observed similar bias of the ATT parameters and their estimated standard errors when comparing the results for the two different sample sizes. The coverage probabilities of the 95% CI:s for the OLS estimates were farther from the nominal value of 0.95 with the larger sample size, which is a consequence of the lower standard errors with the larger sample size.

When only the working exposure model was misspecified, the sandwich estimates of the standard errors of the ATT parameter were close to the empirical standard error. Due to the misspecification of the working exposure model, the estimators obtained through IPW were biased. Since the standard errors were smaller for the larger sample size, we also see a smaller coverage of the 95% CI for the IPW estimates. In contrast, the bias of the ATT estimates obtained with WOLS decreased with the larger sample size, thus demonstrating the doubly robustness property of the WOLS estimator. The standard error of $\hat{\gamma}_{10|1}$ obtained by the sandwich estimator was improved by the larger sample size, leading to coverage probabilities being closer to 0.95. In contrast, the sandwich estimates of the standard error of $\hat{\gamma}_{54|5}$ are still below the empirical standard error. Therefore, the coverage probabilities are also lower than 0.95.

When both working models are incorrect, we observe similar biases for both samples sizes, both in terms of absolute bias for the ATT parameters and for the relative bias of the estimators standard errors. Due to the smaller standard errors in the larger sample, we also see lower coverage probabilities.

In order to visualize influential observations, we use graphs of the logarithms of the weights $\hat{w}(0,1)$, and $\hat{w}(4,5)$ used for the IPW and the WOLS estimators for $n = 5000$. We also do this using both a correct and an incorrect working model for the exposure. The histograms of the log weights based on a correct working model are shown in Fig-

ures 7 and 8. In Figure 7, we see some influential observations. However, the magnitude of these influential weights are not as extreme as the ones we saw in simulation study 1. In Figure 8, we saw few influential weights, but some weights are instead very small and will therefore contribute very little in the estimation of $\gamma_{54|5}$ using IPW and WOLS. In Figures 9 and 10, we see the histogram of the logarithm of the weights $\hat{w}(0,1)$ and $\hat{w}(4,5)$ where the working model for the exposure is misspecified by the omission of two covariates. As in simulation study 1, we calculated the estimated conditional probabilities $\hat{\Pr}(T=m|T\in\{m,l\},X)$ based both on a correct and an incorrect model for the exposure. These probabilities are plotted, along with their complements $\hat{\Pr}(T=m|T\in\{l,m\},X)$, for each ATT parameter $\gamma_{lm|l}$. We only show observations with $T=m$. The results are shown in Appendix C (Figures 14 and 15). We observe that there is a substantial overlap in Figure, 14and that there are few values of $\hat{\Pr}(T=0|X,T\in\{0,1\})$ close to 0. In Figure 14, we see overlap. However, Figure 15, the probabilities of having treatment level $T=5$ are close to 0. This leads to more variability when estimating $\mu_{5|5}$, as it is based on fewer observations. No apparent difference where observed when comparing graphs based on correct and incorrect working models for the exposure.

6 Discussion

We have derived the asymptotic variance for the IPW estimator and the WOLS estimator of the ATT, proposed by Uysal (2015). In the framework of M-estimation, we have presented estimators of both the ATT parameters and the sandwich estimators for their variances for the case when the GPS can be described by a multinomial logistic regression model and when the outcome model is linear. Using the presented formulas, it is straightforward to implement the estimators in standard statistical software such as R, Stata, or SAS.

Using two simulation designs previously used by Yang et al. (2016), we compared the estimators with OLS estimation of the same ATT parameters. In the first design, the treatment had three levels. In the second design, the number of levels increased to six. In both designs, different contrasts of the treatment levels were compared. To assess the robust-

ness properties and the variance estimators under model misspecification, we estimated the parameters under both correctly and incorrectly specified working models for the exposure or for the outcome. The sandwich estimators of the variance, were evaluated by comparing the mean estimated standard errors of the ATT estimators with the mean standard errors obtained by bootstrapping and with empirical standard errors calculated over the simulations. To assess the consequence of bias on inference, coverage probabilities we calculated for 95% CI:s, where the true value of the standard error is approximated by the empirical standard error. When both working models were misspecified, the estimators were consistent for values that are different from the true value of the target parameter, i.e. the ATT. The sandwich estimator as well as the bootstrap estimator of the standard error are then still consistent estimators of the true standard error of this other parameter. As we want to make inference about the true ATT parameter, underestimation of the standard errors in this situation is more likely lead to incorrect conclusions than when the standard errors are overestimated. As model misspecification of the working models, both in terms of the functional form and in terms of omitted confounding variables, is likely to be present at some degree in most real applications, underestimation of the standard errors are undesirable.

In both simulation studies, the proposed sandwich estimator of the standard errors of the IPW estimators gives optimistic, i.e. smaller, standard errors compared to the empirical standard errors calculated over the simulations. This is particularly apparent for estimates of the parameter $\gamma_{10|1}$ in both simulations studies. The underestimated standard errors also led to coverage probabilities of the Wald-based 95% CI:s to be below 0.95. The relative difference in standard error when comparing the empirical standard error with the mean standard error obtained from the sandwich estimator does not improve with bigger sample size. In contrast, the bootstrap estimates of the standard errors of the IPW-estimator were generally larger than the sandwich estimates and were therefore closer to the empirical standard error. In some scenarios, the bootstrap estimates of the standard errors were even overly pessimistic, leading conservative, i.e. too wide, 95% CI:s with coverage probabilities above 0.95.

In both simulation studies, the double robustness property of the WOLS estimator was

clearly seen, although the small sample biases were larger when the outcome model was misspecified compared to when the GPS model was misspecified. In most simulated scenarios when using WOLS, the sandwich estimator gave standard errors which were lower than the empirical standard errors. In contrast, the bootstrap estimates of the standard errors were closer to the empirical standard error. As a consequence, the coverage probability of the 95 % CI:s based the bootstrap estimate of the standard error were closer to 0.95.

Although the results of these simulations showed that most cases were not in favour of our proposed sandwich estimators, there were scenarios in which they are similar or better than the bootstrap estimator. As there might be other scenarios where our sandwich estimator works better, we do not want to make a general recommendation. An obvious advantage of the sandwich estimator is computational speed.

Estimating an ATT parameter $\gamma_{lm|l}$ using the WOLS estimator, as formulated by Uysal (2015), involves estimating the parameters for the full GPS distribution. However, when the treatment variable T follows a multinomial distribution as defined in Assumption 5, we have seen in Appendix A that to calculate thee weights, we only need to estimate the difference $\delta^* = \delta_m - \delta_l$ thus reducing the dimension of the parameter space. Further, only observations with factual treatment level $T = m$ or $T = l$ are required in the estimation of the ATT parameter $\gamma_{lm|l}$. When only one ATT parameter $\gamma_{lm|l}$ is of interest, the ordinary IPW and WOLS estimators of the ATT using only the subset of the observation where $T = m$ and $T = l$ may be more efficient. In the form that we have presented the estimators, the full GPS distribution is estimated, involving estimation of all GPS parameters $\delta_0, \ldots, \delta_K$. This is useful when assessing several ATT parameters involving all levels of the treatment variable T.

Table 1: Comparing mean estimates of $\gamma_{10|1}$ in simulation study 1 over 1000 simulations with $n = 1000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	-0.005	0.234	0.237	0.236	0.947	0.948	0.236
OLS	-	Incorrect	-1.376	0.318	0.324	0.339	0.019	0.020	1.417
IPW	Correct	-	0.099	0.587	0.705	1.125	0.831	0.850	1.128
IPW	Incorrect	-	-1.259	0.622	0.725	1.233	0.415	0.466	1.762
WOLS	Correct	Correct	-0.007	0.412	0.276	0.294	0.955	0.927	0.294
WOLS	Incorrect	Correct	-0.007	0.337	0.262	0.272	0.958	0.939	0.272
WOLS	Correct	Incorrect	-0.190	0.426	0.371	0.428	0.861	0.841	0.468
WOLS	Incorrect	Incorrect	-1.388	0.262	0.349	0.379	0.032	0.049	1.438

Table 2: Comparing mean estimates of $\gamma_{10|1}$ in simulation study 1 over 1000 simulations with $n = 5000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	-0.00	0.10	0.10	0.10	0.95	0.95	0.10
OLS	-	Incorrect	-1.38	0.14	0.14	0.14	0.00	0.00	1.39
IPW	Correct	-	0.04	0.36	0.38	0.77	0.78	0.79	0.77
IPW	Incorrect	-	-1.34	0.37	0.39	0.68	0.02	0.03	1.51
WOLS	Correct	Correct	0.00	0.15	0.13	0.13	0.95	0.94	0.13
WOLS	Incorrect	Correct	0.00	0.14	0.12	0.12	0.94	0.95	0.12
WOLS	Correct	Incorrect	-0.08	0.17	0.18	0.20	0.86	0.88	0.10
WOLS	Incorrect	Incorrect	-1.40	0.11	0.16	0.16	0.00	0.00	1.41

Table 3: Comparing mean estimates of $\gamma_{20|2}$ in simulation study 1 over 1000 simulations with $n = 1000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	0.002	0.268	0.269	0.274	0.942	0.942	0.274
OLS	-	Incorrect	0.933	0.277	0.281	0.281	0.089	0.095	0.975
IPW	Correct	-	0.061	0.456	0.516	0.633	0.894	0.902	0.636
IPW	Incorrect	-	0.970	0.372	0.410	0.482	0.227	0.252	1.083
WOLS	Correct	Correct	-0.001	0.284	0.274	0.279	0.943	0.945	0.279
WOLS	Incorrect	Correct	0.001	0.292	0.273	0.278	0.951	0.948	0.278
WOLS	Correct	Incorrect	0.051	0.245	0.296	0.297	0.888	0.945	0.301
WOLS	Incorrect	Incorrect	0.926	0.221	0.286	0.291	0.060	0.113	0.971

Table 4: Comparing mean estimates of $\gamma_{20|2}$ in simulation study 1 over 1000 simulations with $n = 5000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	0.00	0.12	0.12	0.12	0.95	0.95	0.12
OLS	-	Incorrect	0.93	0.12	0.12	0.12	0.00	0.00	0.94
IPW	Correct	-	-0.00	0.24	0.24	0.27	0.92	0.92	0.27
IPW	Incorrect	-	0.93	0.19	0.19	0.21	0.05	0.05	0.96
WOLS	Correct	Correct	0.00	0.12	0.12	0.13	0.95	0.95	0.13
WOLS	Incorrect	Correct	0.00	0.12	0.12	0.13	0.96	0.96	0.13
WOLS	Correct	Incorrect	0.01	0.11	0.13	0.13	0.88	0.95	0.13
WOLS	Incorrect	Incorrect	0.92	0.10	0.13	0.13	0.00	0.00	0.93

Table 5: Comparing mean estimates of $\gamma_{21|2}$ in simulation study 1 over 1000 simulations with $n = 1000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	0.005	0.346	0.346	0.344	0.947	0.946	0.343
OLS	-	Incorrect	4.593	0.298	0.299	0.301	0.000	0.000	4.603
IPW	Correct	-	-0.017	0.652	0.683	0.695	0.946	0.953	0.695
IPW	Incorrect	-	4.600	0.311	0.314	0.318	0.000	0.000	4.611
WOLS	Correct	Correct	0.008	0.357	0.356	0.356	0.950	0.949	0.356
WOLS	Incorrect	Correct	0.005	0.349	0.347	0.345	0.954	0.948	0.345
WOLS	Correct	Incorrect	0.129	0.326	0.447	0.463	0.809	0.913	0.480
WOLS	Incorrect	Incorrect	4.596	0.244	0.299	0.301	0.000	0.000	4.606

Table 6: Comparing mean estimates of $\gamma_{21|2}$ in simulation study 1 over 1000 simulations with $n = 5000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	-0.00	0.15	0.16	0.16	0.95	0.94	0.15
OLS	-	Incorrect	4.60	0.13	0.13	0.13	0.00	0.00	4.60
IPW	Correct	-	-0.00	0.29	0.29	0.31	0.94	0.94	0.31
IPW	Incorrect	-	4.61	0.14	0.14	0.14	0.00	0.00	4.61
WOLS	Correct	Correct	0.00	0.16	0.16	0.16	0.95	0.95	0.16
WOLS	Incorrect	Correct	-0.00	0.16	0.16	0.16	0.94	0.94	0.16
WOLS	Correct	Incorrect	0.03	0.14	0.20	0.21	0.83	0.94	0.21
WOLS	Incorrect	Incorrect	4.60	0.11	0.13	0.13	0.00	0.00	4.60

Table 7: Comparing mean estimates of $\gamma_{10|1}$ in simulation study 2 over 1000 simulations with $n = 1000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	-0.035	0.486	0.500	0.480	0.949	0.960	0.481
OLS	-	Incorrect	0.526	0.569	0.609	0.577	0.852	0.875	0.781
IPW	Correct	-	0.032	0.639	0.866	0.735	0.940	0.962	0.735
IPW	Incorrect	-	0.628	0.634	0.756	0.676	0.786	0.821	0.923
WOLS	Correct	Correct	-0.021	0.769	0.532	0.518	0.967	0.950	0.518
WOLS	Incorrect	Correct	-0.029	0.822	0.522	0.493	0.981	0.956	0.493
WOLS	Correct	Incorrect	0.086	0.627	0.599	0.548	0.950	0.961	0.555
WOLS	Incorrect	Incorrect	0.531	0.550	0.623	0.594	0.803	0.873	0.796

Table 8: Comparing mean estimates of $\gamma_{10|1}$ in simulation study 2 over 1000 simulations with $n = 5000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	-0.003	0.217	0.218	0.222	0.942	0.949	0.222
OLS	-	Incorrect	0.532	0.258	0.259	0.255	0.457	0.463	0.590
IPW	Correct	-	0.000	0.298	0.315	0.343	0.936	0.943	0.343
IPW	Incorrect	-	0.608	0.290	0.298	0.295	0.421	0.434	0.676
WOLS	Correct	Correct	-0.005	0.227	0.221	0.226	0.944	0.945	0.226
WOLS	Incorrect	Correct	-0.005	0.245	0.221	0.226	0.961	0.946	0.226
WOLS	Correct	Incorrect	0.020	0.223	0.238	0.241	0.919	0.939	0.242
WOLS	Incorrect	Incorrect	0.527	0.205	0.261	0.258	0.299	0.479	0.587

Table 9: Comparing mean estimates of $\gamma_{54|5}$ in simulation study 2 over 1000 simulations with $n = 1000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	0.014	0.165	0.165	0.169	0.942	0.944	0.170
OLS	-	Incorrect	0.115	0.199	0.204	0.208	0.907	0.913	0.237
IPW	Correct	-	0.031	0.568	0.649	0.588	0.945	0.968	0.589
IPW	Incorrect	-	0.895	0.504	0.547	0.523	0.584	0.621	1.036
WOLS	Correct	Correct	0.018	0.163	0.193	0.184	0.914	0.947	0.185
WOLS	Incorrect	Correct	0.016	0.126	0.183	0.182	0.825	0.950	0.183
WOLS	Correct	Incorrect	0.020	0.206	0.201	0.196	0.954	0.951	0.197
WOLS	Incorrect	Incorrect	0.186	0.142	0.230	0.235	0.636	0.870	0.299

Table 10: Comparing mean estimates of $\gamma_{54|5}$ in simulation study 2 over 1000 simulations with $n = 5000$.

Method	Exposure model	Outcome model	Bias	sandwich s.e.	bootstrap s.e.	empirical s.e.	Coverage 95% CI	Coverage 95% bootstrap CI	RMSE
OLS	-	Correct	0.001	0.073	0.074	0.075	0.940	0.940	0.075
OLS	-	Incorrect	0.098	0.090	0.090	0.092	0.805	0.804	0.135
IPW	Correct	-	0.003	0.239	0.245	0.239	0.949	0.958	0.239
IPW	Incorrect	-	0.867	0.220	0.223	0.220	0.021	0.023	0.895
WOLS	Correct	Correct	0.001	0.070	0.080	0.081	0.895	0.939	0.081
WOLS	Incorrect	Correct	0.001	0.053	0.080	0.081	0.807	0.940	0.081
WOLS	Correct	Incorrect	0.002	0.091	0.086	0.087	0.951	0.939	0.087
WOLS	Incorrect	Incorrect	0.166	0.062	0.100	0.100	0.302	0.630	0.194

References

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Boos, D. D. and L. A. Stefanski (2013). *Essential statistical inference: theory and methods*, Volume 120. Springer Science & Business Media.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics 155*(2), 138–154.

Cole, S. R. and M. A. Hernán (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology 168*(6), 656–664.

Feng, P., X.-H. Zhou, Q.-M. Zou, M.-Y. Fan, and X.-S. Li (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine 31*(7), 681–697.

Imai, K. and D. A. Van Dyk (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association 99*(467), 854–866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Lefebvre, G., J. A. Delaney, and R. W. Platt (2008). Impact of misspecification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine 27*(18), 3629–3642.

Linden, A., S. D. Uysal, A. Ryan, and J. L. Adams (2016). Estimating causal effects for multivalued treatments: a comparison of approaches. *Statistics in Medicine 35*(4), 534–552.

Little, R. J. and D. B. Rubin (2019). *Statistical analysis with missing data*, Volume 793. John Wiley & Sons.

Seaman, S. R. and S. Vansteelandt (2018, 05). Introduction to double robust methods for incomplete data. *Statistical Science 33*(2), 184–197.

Słoczyński, T. and J. M. Wooldridge (2018). A general double robustness result for estimating average treatment effects. *Econometric Theory 34*(1), 112–133.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika 97*(3), 661–682.

Uysal, S. D. (2015). Doubly robust estimation of causal effects with multivalued treatments: an application to the returns to schooling. *Journal of Applied Econometrics 30*(5), 763–786.

Yang, S., G. W. Imbens, Z. Cui, D. E. Faries, and Z. Kadziola (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics 72*(4), 1055–1065.

Appendix A

The expected value of $D(t)U|V$

A fact that will be used repeatedly in this appendix is that, for all $t \in \mathscr{T}$ and any random variables U and V,

$$\begin{aligned}
&\mathrm{E}\left[D(t)U|V\right] \\
=&\Pr\left[D(t)=1|V\right]\mathrm{E}\left[D(t)U|D(t)=1,V\right] \\
&\quad+\Pr\left[D(t)=0|V\right]\underbrace{\mathrm{E}\left[D(t)U|D(t)=0,V\right]}_{=0} \qquad \text{(LIE)} \\
=&\Pr\left[D(t)=1|V\right]\mathrm{E}\left[U|D(t)=1,V\right], \qquad (15)
\end{aligned}$$

where the law of iterated expectation (LIE) was used in the first equality. Since $D(t) = 1 \Leftrightarrow T = t$ we will also use the formulation

$$\mathrm{E}[D(t)U|V] = \Pr[T = t|V]\,\mathrm{E}[U|T = t, V] \tag{16}$$

The expected value of $D(l)Y$ used in IPW and WOLS estimation

Using the facts stated in the previous subsection, we have that,

$$\begin{aligned}
&\mathrm{E}[D(l)Y] \\
&= \Pr[T = l]\,\mathrm{E}[Y|T = l] && \text{by (16)} \\
&= \Pr[T = l]\,\mathrm{E}[Y(l)|T = l] && \text{by Assumption 1} \\
&= \Pr[T = l]\,\mu_{l|l} && (17)
\end{aligned}$$

The expected value of $w(m,l)Y$ used in IPW and WOLS estimation

By the law of iterated expectations and Assumptions 1 and 2,

$$
\begin{aligned}
&\mathrm{E}\left[D(m)\frac{\Pr(T=l\mid X)}{\Pr(T=m\mid X)}Y\right] \\
&=\mathrm{E}\left\{\mathrm{E}\left[D(m)\frac{\Pr(T=l\mid X)}{\Pr(T=m\mid X)}Y\middle|X\right]\right\} && \text{(LIE)}\\
&=\mathrm{E}\left\{\frac{\Pr(T=l\mid X)}{\Pr(T=m\mid X)}\mathrm{E}\left[D(m)Y\middle|X\right]\right\} \\
&=\mathrm{E}\left\{\frac{\Pr(T=l\mid X)}{\Pr(T=m\mid X)}\Pr\left[T=m\middle|X\right]\mathrm{E}\left[Y\middle|T=m,X\right]\right\} && \text{by (16)}\\
&=\mathrm{E}\left\{\Pr\left[T=l\middle|X\right]\mathrm{E}\left[Y\middle|T=m,X\right]\right\} \\
&=\mathrm{E}\left\{\Pr\left[T=l\middle|X\right]\mathrm{E}\left[Y(m)\middle|T=m,X\right]\right\} && \text{by Assumption 1}\\
&=\mathrm{E}\left\{\mathrm{E}\left[D(l)\middle|X\right]\mathrm{E}\left[Y(m)\middle|D(m)=1,X\right]\right\} && \text{since } T=t\Leftrightarrow D(t)=1\\
&=\mathrm{E}\left\{\mathrm{E}\left[D(l)\middle|X\right]\mathrm{E}\left[Y(m)\middle|X\right]\right\} && \text{by Assumption 2}\\
&=\mathrm{E}\left\{\mathrm{E}\left[D(l)Y(m)\middle|X\right]\right\} && \text{by Assumption 2}\\
&=\mathrm{E}\left[D(l)Y(m)\right] && \text{by (LIE)}\\
&=\Pr\left[T=l\right]\mathrm{E}\left[Y(m)\mid T=l\right] && \text{by (16)}\\
&=\Pr\left[T=l\right]\mu_{m|l}. && (18)
\end{aligned}
$$

Now, assume that the GPS model in Assumption 5 is correctly specified. Then

$$r(t,X)=\Pr(T=t|X) \text{ for } t\in\mathscr{T}$$

and

$$w(m,l)=D(m)\frac{r(l,X)}{r(m,X)}=D(m)\frac{\Pr(T=l|X)}{\Pr(T=m|X)}.$$

It then immediately follows that

$$\mathrm{E}\left[w(m,l)Y\right]=\Pr(T=l)\mu_{m|l} \tag{19}$$

The expected value of the IPW estimating function

Assuming that the GPS model in Assumption 5 is correct, it follows from (17) and (19) that

$$\begin{aligned}\mathrm{E}\left[M^{ipw-att}(\gamma_{lm|l},\delta)\right] &= \mathrm{E}\left\{D(l)\gamma_{lm|l} - [D(l)Y - w(m,l)Y]\right\} \\ &= \Pr(T=l)\left\{\gamma_{lm|l} - \left[\mu_{l|l} - \mu_{m|l}\right]\right\} = 0\end{aligned}$$

when $\gamma_{lm|l}$ and δ in Assumption 5 are the true parameters of the data generating process.

The solution to the estimating equations for β_l and β_m

For $t \in \{l,m\}$, let $\tilde{\beta}_t$ be the unique solution to the equation

$$\mathrm{E}\left[M^{wols-t}(\beta_t,\delta)\right] = \mathrm{E}\left[Xw(t,l)(Y - X'\beta_t)\right] = 0\,.$$

We note that such solutions $\tilde{\beta}_l$ and $\tilde{\beta}_m$ exist under fairly mild regularity conditions (Boos and Stefanski, 2013), regardless of whether the outcome model in Assumption 4 is correctly specified or not. Then, since 1 is the first element of X, it follows that

$$\mathrm{E}\left[w(t,l)Y\right] = \mathrm{E}\left[w(t,l)X'\tilde{\beta}_t\right] \text{ for } t \in \{l,m\} \tag{20}$$

Estimating $\Pr(T=l)\mu_{l|l}$

Since $w(l,l) = D(l)$, it follows from (20) and (17) that

$$\mathrm{E}\left[D(l)X'\tilde{\beta}_l\right] = \mathrm{E}\left[D(l)Y\right] = \Pr(T=l)\mu_{l|l}, \tag{21}$$

regardless of whether the outcome model in Assumption 4 is correctly specified or not.

Estimating $\Pr(T=l)\mu_{m|l}$

Now, assume that the GPS model in Assumption 5 is correct with true parameter δ. Then it follows that

$$
\begin{aligned}
&\mathrm{E}\left[D(l)X'\tilde{\beta}_m\right] \\
=&\mathrm{E}\left\{\mathrm{E}\left[D(l)X'\tilde{\beta}_m|X\right]\right\} && \text{(LIE)} \\
=&\mathrm{E}\left\{\mathrm{E}\left[D(l)|X\right]X'\tilde{\beta}_m\right\} \\
=&\mathrm{E}\left[\Pr(T=l|X)X'\tilde{\beta}_m\right] && \text{since } D(l)=1 \Leftrightarrow T=l \\
=&\mathrm{E}\left[\frac{\Pr(T=l|X)}{\Pr(T=m|X)}\Pr(T=m|X)X'\tilde{\beta}_m\right] \\
=&\mathrm{E}\left\{\frac{\Pr(T=l|X)}{\Pr(T=m|X)}\mathrm{E}[D(m)|X]X'\tilde{\beta}_m\right\} && \text{since } D(m)=1 \Leftrightarrow T=m \\
=&\mathrm{E}\left\{\mathrm{E}\left[D(m)\frac{\Pr(T=l|X)}{\Pr(T=m|X)}X'\tilde{\beta}_m\middle|X\right]\right\} \\
=&\mathrm{E}\left\{\mathrm{E}\left[w(m,l)X'\tilde{\beta}_m\middle|X\right]\right\} && \text{by Assumption 5} \\
=&\mathrm{E}\left[w(m,l)X'\tilde{\beta}_m\right] && \text{(LIE)} \\
=&\mathrm{E}\left[w(m,l)Y\right] && \text{by (20)} \\
=&\Pr(T=l)\mu_{m|l} && \text{by (19) and Assumption 5.}
\end{aligned}
$$

The expected value of the WOLS estimating function when the GPS model is correctly specified

When the GPS model in Assumption 5 is correct, it follows from the two previous sections that

$$
\begin{aligned}
&\mathrm{E}[M^{wols-att}(\gamma_{lm|l},\tilde{\beta}_l,\tilde{\beta}_m,\delta)] \\
=&\mathrm{E}\left\{D(l)\left[\gamma_{lm|l}-\left(X'\tilde{\beta}_l-X'\tilde{\beta}_m\right)\right]\right\} \\
=&\Pr(T=l)\left[\gamma_{lm|l}-\left(\mu_{l|l}-\mu_{m|l}\right)\right] \\
=&0
\end{aligned}
$$

when $\gamma_{lm|l}$ is the true value of the ATT.

The expected value of the WOLS estimating function when the outcome model is correctly specified

We note that, under mild regularity conditions (Boos and Stefanski, 2013), the equation

$$\mathrm{E}\left[M^{gps}(O;\delta)\right]=0$$

has a unique solution, $\tilde{\delta}$. The working models $r(t,X)$ for $\Pr(T=t|X)$ in Assumption 5 are therefore defined, regardless of whether the GPS model is correctly specified or not. Now, assume that the outcome model in Assumption 4 is correct, such that, for $t=l,m$, the solution $\tilde{\beta}_t$ to the equation

$$\mathrm{E}\left[M^{wols-t}(O;\beta_t,\delta)\right]=0$$

is also the true parameter β_t in the model in Assumption 4. In other words, for $t\in\mathscr{T}$,

$$\mathrm{E}(Y|T=t,X)=X'\beta_t=X'\tilde{\beta}_t.$$

Then

$$\begin{aligned}
&\mathrm{E}\left[Xw(t,l)\left(Y-X'\beta_t\right)\right]\\
=&\mathrm{E}\left\{\mathrm{E}\left[Xw(t,l)\left(Y-X'\beta_t\right)|X\right]\right\} && \text{(LIE)}\\
=&\mathrm{E}\left\{X\frac{r(l,X)}{r(t,X)}\mathrm{E}\left[D(t)\left(Y-X'\beta_t\right)|X\right]\right\} && \text{by the definition of } w(t,l)\\
=&\mathrm{E}\left(X\frac{r(l,X)}{r(t,X)}\left\{\Pr[D(t)=1|X]\,\mathrm{E}\left[D(t)\left(Y-X'\beta_t\right)|D(t)=1,X\right]\right.\right.\\
&\qquad\left.\left.+\quad\Pr[D(t)=0|X]\underbrace{\mathrm{E}\left[D(t)\left(Y-X'\beta_t\right)|D(t)=0,X\right]}_{=0}\right\}\right) && \text{(LIE)}\\
=&\mathrm{E}\left\{X\frac{r(l,X)}{r(t,X)}\Pr(T=t|X)\,\mathrm{E}\left[\left(Y-X'\beta_t\right)|T=t,X\right]\right\} && \text{since } D(t)=1\Leftrightarrow T=t\\
=&\mathrm{E}\left[X\frac{r(l,X)}{r(t,X)}\Pr(T=t|X)\times 0\right] && \text{by Assumption 4}\\
=&0\,,
\end{aligned}$$

showing that β_t solves the equation $\mathrm{E}[M^{wols-t}(\beta_t,\delta)]=0$ for $t=l,m$ regardless of whether the GPS model is correct or not. Further, we have that

$$
\begin{aligned}
&\mathrm{E}[D(l)X'\beta_m] \\
=&\Pr(T=l)\,\mathrm{E}\left(X'\beta_m|T=l\right) && \text{by (16)}\\
=&\Pr(T=l)\,\mathrm{E}\left[\mathrm{E}\left(Y|X,T=m\right)|T=l\right] && \text{by Assumption 4}\\
=&\Pr(T=l)\,\mathrm{E}\left\{\mathrm{E}\left[Y(m)|X,T=m\right]|T=l\right\} && \text{by Assumption 1}\\
=&\Pr(T=l)\,\mathrm{E}\left\{\mathrm{E}\left[Y(m)|X,D(m)=1\right]|T=l\right\} && \text{since } D(m)=1\Leftrightarrow T=m\\
=&\Pr(T=l)\,\mathrm{E}\left\{\mathrm{E}\left[Y(m)|X\right]|T=l\right\} && \text{by Assumption 2}\\
=&\Pr(T=l)\,\mathrm{E}\left\{\mathrm{E}\left[Y(m)|X,D(l)\right]|T=l\right\} && \text{by Assumption 2}\\
=&\Pr(T=l)\,\mathrm{E}\left\{\mathrm{E}\left[Y(m)|X,T=l\right]|T=l\right\} && \text{since } D(l)=1\Leftrightarrow T=l\\
=&\Pr(T=l)\,\mathrm{E}\left[Y(m)|T=l\right] && \text{(LIE)}\\
=&\Pr(T=l)\mu_{m|l}.
\end{aligned}
$$

Since $\tilde{\beta}_l=\beta_l$ in (21), it now follows that

$$
\begin{aligned}
&\mathrm{E}[M^{wols-att}(\gamma_{lm|l},\beta_l,\beta_m,\delta)]\\
=&\mathrm{E}\left\{D(l)\left[\gamma_{lm|l}-\left(X'\beta_l-X'\beta_m\right)\right]\right\}\\
=&\Pr(T=l)\left[\gamma_{lm|l}-\left(\mu_{l|l}-\mu_{m|l}\right)\right]\\
=&0
\end{aligned}
$$

when $\gamma_{lm|l}$ is the true value of the ATT.

The gradient of M^{gps} with respect to δ

Assuming that model in Assumption 5 is multinomial, we have, for $t \in \mathscr{T}$ and $s \in \mathscr{T} \setminus \{t, K\}$ that

$$\begin{aligned}
\frac{\partial r(t,X)}{\partial \delta_s} =& \frac{\partial}{\partial \delta_s} \frac{\exp(X'\delta_t)}{\sum_{k=1}^{K} \exp(X'\delta_k)} \\
=& \frac{\partial}{\partial \delta_s} \frac{1}{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)]} \\
=& \frac{-\frac{\partial}{\partial \delta_s} \exp[X'(\delta_s - \delta_t)]}{\left\{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)]\right\}^2} \\
=& -X \frac{\exp[X'(\delta_s - \delta_t)]}{\left\{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)]\right\}^2} \\
=& -X \frac{\exp(X'\delta_s)\exp(X'\delta_t)}{\left\{\sum_{k=1}^{K} \exp(X'\delta_k)\right\}^2} \\
=& -X r(t,X) r(s,X)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial r(t,X)}{\partial \delta_t} =& \frac{\partial}{\partial \delta_t} \frac{\exp(X'\delta_t)}{\sum_{k=1}^{K} \exp(X'\delta_k)} \\
=& \frac{\partial}{\partial \delta_t} \frac{1}{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)]} \\
=& \frac{-\sum_{k=1}^{K} \frac{\partial}{\partial \delta_t} \exp[X'(\delta_k - \delta_t)]}{\left\{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)]\right\}^2} \\
=& X \frac{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)] - 1}{\left\{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)]\right\}^2} \\
=& X \frac{1}{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)]} \frac{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)] - 1}{\sum_{k=1}^{K} \exp[X'(\delta_k - \delta_t)]} \\
=& X r(t,X) \left[1 - r(t,X)\right] \\
=& -X r(t,X) \left[r(t,X) - 1\right],
\end{aligned}$$

for $t \in \mathscr{T} \setminus \{K\}$. More compactly, we can thus write

$$\frac{\partial r(t,X)}{\partial \delta_s} = -Xr(t,X)\left[r(s,X) - I(t=s)\right],$$

for $t \in \mathscr{T}$ and $k \in \mathscr{T} \setminus \{K\}$. Therefore

$$\begin{aligned}\frac{\partial \log r(t,X)}{\partial \delta_s} =& \frac{1}{r(t,X)} \times \frac{\partial r(t,X)}{\partial \delta_s} \\ =& \frac{1}{r(t,X)} \times -Xr(t,X)\left[r(s,X) - I(t=s)\right] \\ =& -X\left[r(s,X) - I(t=s)\right],\end{aligned}$$

The contribution to the likelihood for the GPS parameters $\delta_1, \ldots, \delta_{K-1}$ from an observation is

$$\log\left[\sum_{t=1}^{K-1} D(t) r(t,X)\right] = \sum_{t=1}^{K-1} D(t) \log r(t,X).$$

Thus the ML score function for each GPS parameter δ_s, $s \in \mathscr{T} \setminus \{K\}$, is

$$M^{gps-s}(\delta) = -X\left[r(s,X) - I(T=s)\right] = -X\left[r(s,X) - D(s)\right]$$

Taking the gradient of M^{gps-s} with respect to δ_t, we get

$$\frac{\partial M^{gps-s}(\delta)}{\partial \delta_t'} = -X\frac{\partial r(s,X)}{\partial \delta_t'} = XX'r(s,X)\left[r(t,X) - I(s=t)\right]$$

for $s,t \in \mathscr{T} \setminus \{K\}$.

The gradient of $w_i(t,l)$ with respect to δ_s

Next, we calculate the gradient of the weights $w_i(t,l)$ with respect to δ_s for $t \in \mathscr{T}$:

$$\frac{\partial w_i(t,l)}{\partial \delta_s'} = \frac{\partial}{\partial \delta_s'} D_i(t) e^{(\delta_l' - \delta_t')X_i} = \begin{cases} X_i' w_i(t,l) & \text{when } s = l \text{ and } t \neq l \\ -X_i' w_i(t,l) & \text{when } s = t \text{ and } t \neq l \\ 0 & \text{otherwise} \end{cases}$$

The gradient of $M^{ipw-att}$ with respect to δ_s

Since

$$\frac{\partial M_i^{ipw-att}(\theta^{ipw})}{\partial \delta_s} = \frac{\partial w_i(m)}{\partial \delta_s} Y_i,$$

we have, for $s \in \mathscr{T} \setminus \{0\}$, that

$$\frac{\partial M_i^{ipw-att}(\theta^{ipw})}{\partial \delta_s} = \begin{cases} w_i(m,l) X_i' Y_i & \text{when } s = l \\ -w_i(m,l) X_i' Y_i & \text{when } s = m \\ 0 & \text{otherwise} \end{cases}$$

assuming that $m \neq l$.

The gradient of M^{wols-m} with respect to δ_s

Since the estimating function for the parameter β_t is

$$M_i^{wols-t} = w_i(t,l) X_i (Y_i - X_i' \beta_t),$$

for $t \in \{l, m\}$, it follows that

$$\frac{\partial M_i^{wols-t}}{\partial \delta_s'} = X_i (Y_i - X_i' \beta_t) \frac{\partial w_i(t,l)}{\partial \delta_s'},$$

so

$$\frac{\partial M_i^{wols-t}}{\partial \delta_s'} = \begin{cases} X_i X_i' w_i(t,l)(Y_i - X_i' \beta_t) & \text{when } s = l \text{ and } t \neq l \\ -X_i X_i' w_i(t,l)(Y_i - X_i' \beta_t) & \text{when } s = m \text{ and } t \neq l \cdot \\ 0 & \text{otherwise} \end{cases}$$

assuming that $m \neq l$.

Appendix B

The GPS conditional on two levels

When the distribution of the treatment level T follows a multinomial model as defined in Assumption 5, conditioning on the treatment being in one of two levels t and l can be written as

$$\begin{aligned}
\Pr(T=t|T\in\{t,l\},X) =& \frac{\Pr(T=t|X)}{\Pr(T\in\{t,l\},X)} \\
=& \frac{\Pr(T=t|X)}{\Pr(T=t|X)+\Pr(T=l|X)} \\
=& \frac{\exp(X'\delta_t)}{\sum_{s=1}^{K}\exp(X'\delta_s)} \Bigg/ \frac{\exp(X'\delta_t)+\exp(X'\delta_l)}{\sum_{s=1}^{K}\exp(X'\delta_s)} \\
=& \frac{\exp(X'\delta_t)}{\exp(X'\delta_t)+\exp(X'\delta_l)} \\
=& \frac{\exp[X'(\delta_t-\delta_l)]}{1+\exp[X'(\delta_t-\delta_l)]},
\end{aligned}$$

showing that, the distribution follows a logistic regression model. Further, in order to estimate the conditional probabilities, we only need to estimate the parameter $\delta^* = \delta_t - \delta_l$, thus reducing the dimensionality of the estimation problem.

The weights $w(m,l)$ depends on δ_m and δ_l only

The weights $w(m,l)$ can be rewritten as

$$\begin{aligned}
w(m,l) =& D(m)\frac{\Pr(T=l|X)}{\Pr(T=m|X)} \\
=& D(m)\frac{\Pr(T=l|X)/\Pr(T\in\{m,l\}|X)}{\Pr(T=m|X)/\Pr(T\in\{m,l\}|X)} \\
=& D(m)\frac{\Pr(T=l|T\in\{m,l\},X)}{\Pr(T=m|T\in\{m,l\},X)} \\
=& D(m)\frac{\Pr(T=l|T\in\{m,l\},X)}{1-\Pr(T=l|T\in\{m,l\},X)} \\
=& D(m)\frac{1-\Pr(T=m|T\in\{m,l\},X)}{\Pr(T=m|T\in\{m,l\},X)} \\
=& D(m)\exp[X'(\delta_m-\delta_l)].
\end{aligned}$$

This shows that IPW and WOLS estimators only depends on the conditional probabilities $\Pr(T = m|T \in \{m,l\},X)$ and not the full GPS distribution.

The GPS conditional on more than two levels

Assuming that the distribution of the treatment level T follows a multinomial model as defined in Assumption 5, conditioning on the treatment being in a strict subset $\mathscr{R}$ of $\mathscr{T} = \{1,\ldots,K\}$ yields another multinomial distribution. Given a reference level $l \in \mathscr{R}$, the probability of this distribution can be written as

$$\begin{aligned}
\Pr(T = t|T \in \mathscr{R},X) =& \frac{\Pr(T = t|X)}{\Pr(T \in \mathscr{R},X)} \\
=& \frac{\Pr(T = t|X)}{\sum_{r\in\mathscr{R}} \Pr(T = r|X)} \\
=& \frac{\exp(X'\delta_t)}{\sum_{s=1}^{K} \exp(X'\delta_s)} \Big/ \frac{\sum_{r\in\mathscr{R}} \exp(X'\delta_r)}{\sum_{s=1}^{K} \exp(X'\delta_s)} \\
=& \frac{\exp(X'\delta_t)}{\sum_{r\in\mathscr{R}} \exp(X'\delta_r)} \\
=& \frac{\exp[X'(\delta_t - \delta_l)]}{1 + \sum_{r\in\mathscr{R}\setminus\{l\}} \exp[X'(\delta_r - \delta_l)]},
\end{aligned}$$

for $t \in \mathscr{R}$, showing that, the restricting the exposure to a subset of the original levels also yields a multinomial distribution, which can be parameterized by parameters $\delta^*_{r,l}$, $r \in \mathscr{R} \setminus \{l\}$, thereby restricting the dimensionality of the estimation problem.

Joint inference for multiple ATT parameters

When several ATT parameters $\gamma_{l_1,m_1|l_1},\ldots,\gamma_{l_R,m_R|l_R}$ for $(m_1,l_1),\ldots,(m_R,l_R) \in \mathscr{T} \times \mathscr{T}$ are of interest, it is straightforward to extend the estimating function (3) for the IPW estimator to include components $M^{ipw-att}(O;\gamma_{lm|l},\delta)$ for for each ATT contrast $\gamma_{lm|l}$ of interest. Similarly, by extending the estimating function (3) to include components $M^{wols-l}(O;\beta_l,\delta)$, $M^{wols-m}(O;\beta_m,\delta)$ and $M^{wols-att}(O;\gamma_{lm|l},\beta_l,\beta_m)$ for each contrast $\gamma_{lm|l}$, we can obtain an estimator of the full vector $(\gamma_{l_1,m_1|l_1},\ldots,\gamma_{l_R,m_R|l_R})$. Further, the variance of this estimator can then be obtained by extending the gradients (13) and (14) to include components for each ATT parameter $\gamma_{lm|l}$ of interest along with the corresponding components for β_l and β_m. Using the delta

method, comparisons of different ATT parameters can then be performed by considering linear combinations of the parameter vector $(\gamma_{l_1,m_1|l_1},\ldots,\gamma_{l_R,m_R|l_R})$. In this way, we take the uncertainty from the estimation of the parameters δ_t and β_t, for $t \in \mathscr{T}$, into account, while allowing us to test several parameters at once.

Appendix C

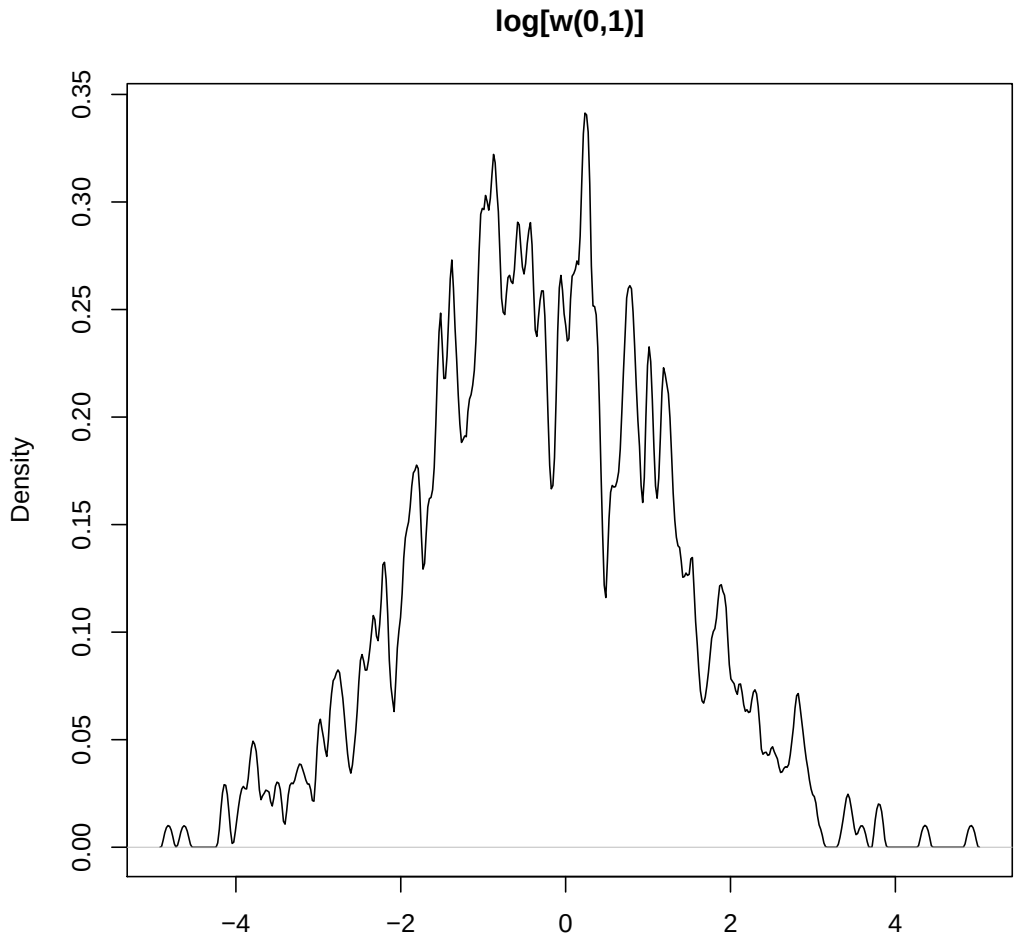

Figure 1: Distributions of $\log[\hat{w}(0,1)] = \log[\hat{\Pr}(T=1|X)] - \log[\hat{\Pr}(T=0|X)]$ over 5000 observations when the correct working exposure model in simulation study 1 is used. Only non-zero values of $\hat{w}(0,1)$ showed.

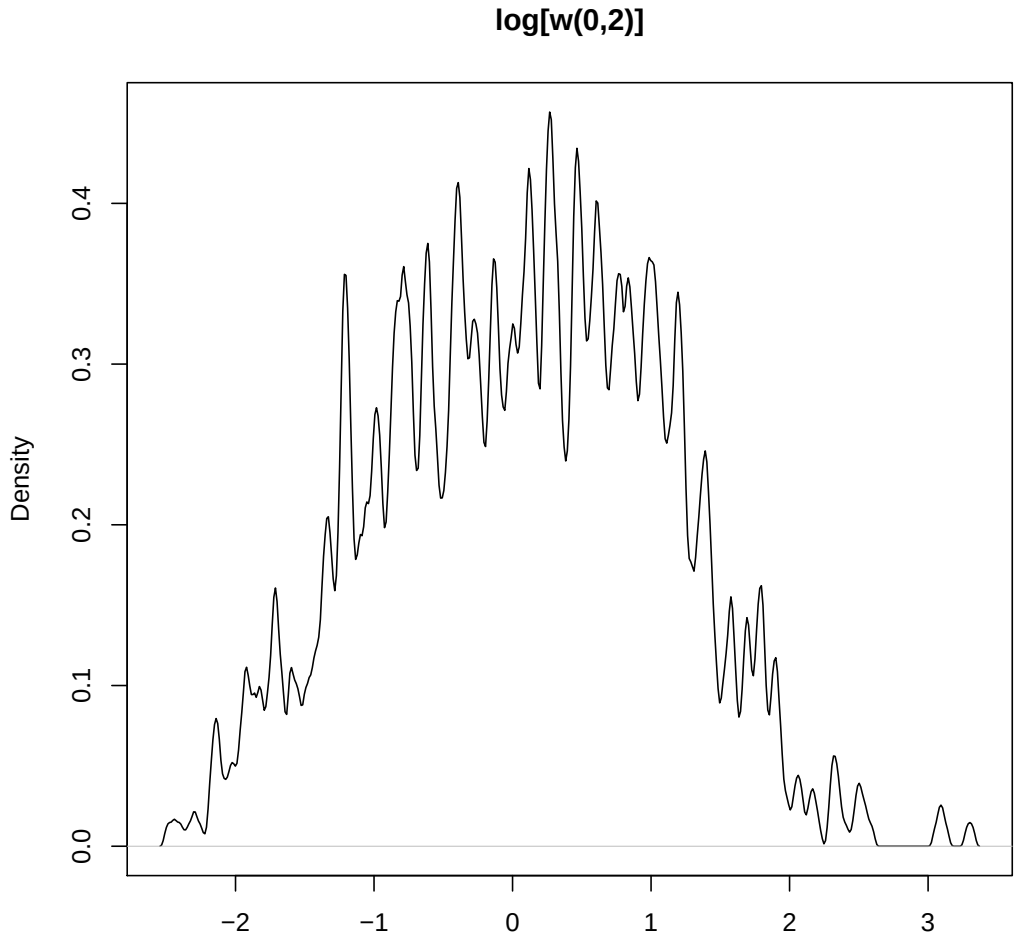

Figure 2: Distributions of $\log[\hat{w}(0,2)] = \log[\hat{\Pr}(T=2|X)] - \log[\hat{\Pr}(T=0|X)]$ over 5000 observations when the correct working exposure model in simulation study 1 is used. Only non-zero values of $\hat{w}(0,2)$ showed.

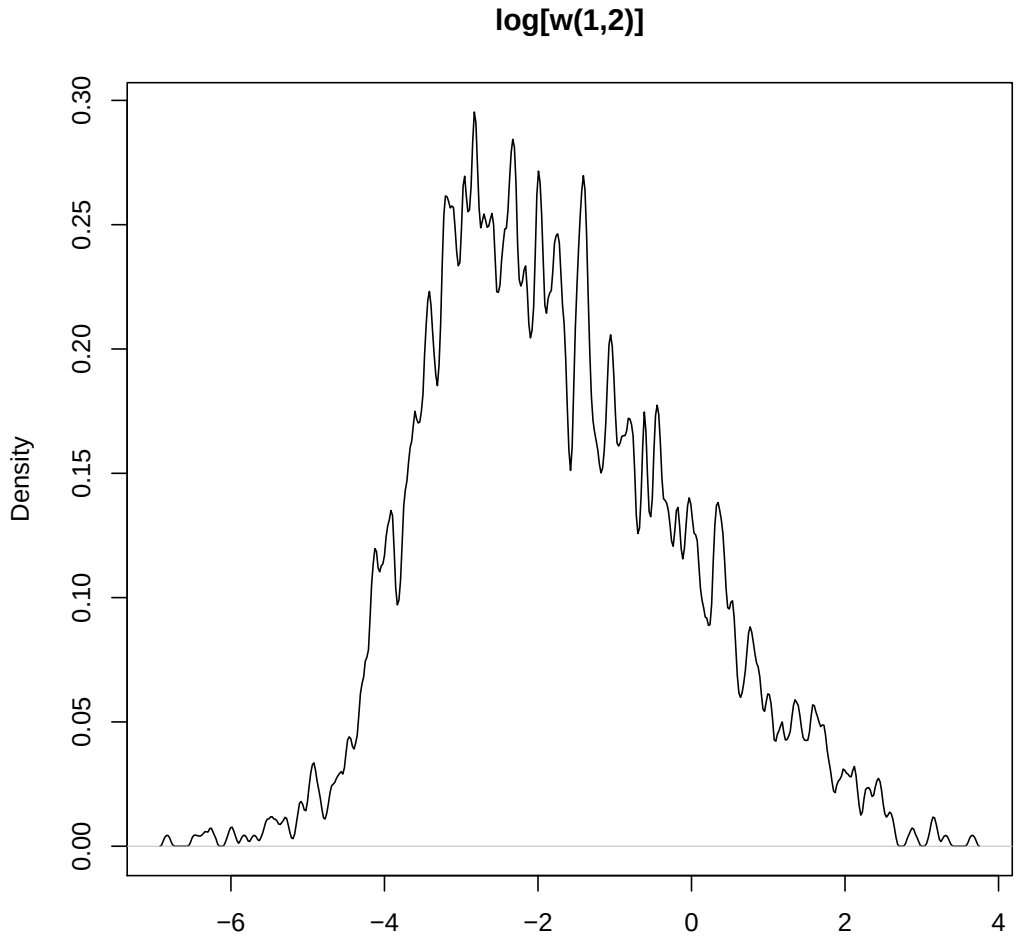

Figure 3: Distributions of $\log[\hat{w}(1,2)] = \log[\hat{\Pr}(T=2|X)] - \log[\hat{\Pr}(T=1|X)]$ over 5000 observations when the correct working exposure model in simulation study 1 is used. Only non-zero values of $\hat{w}(1,2)$ showed.

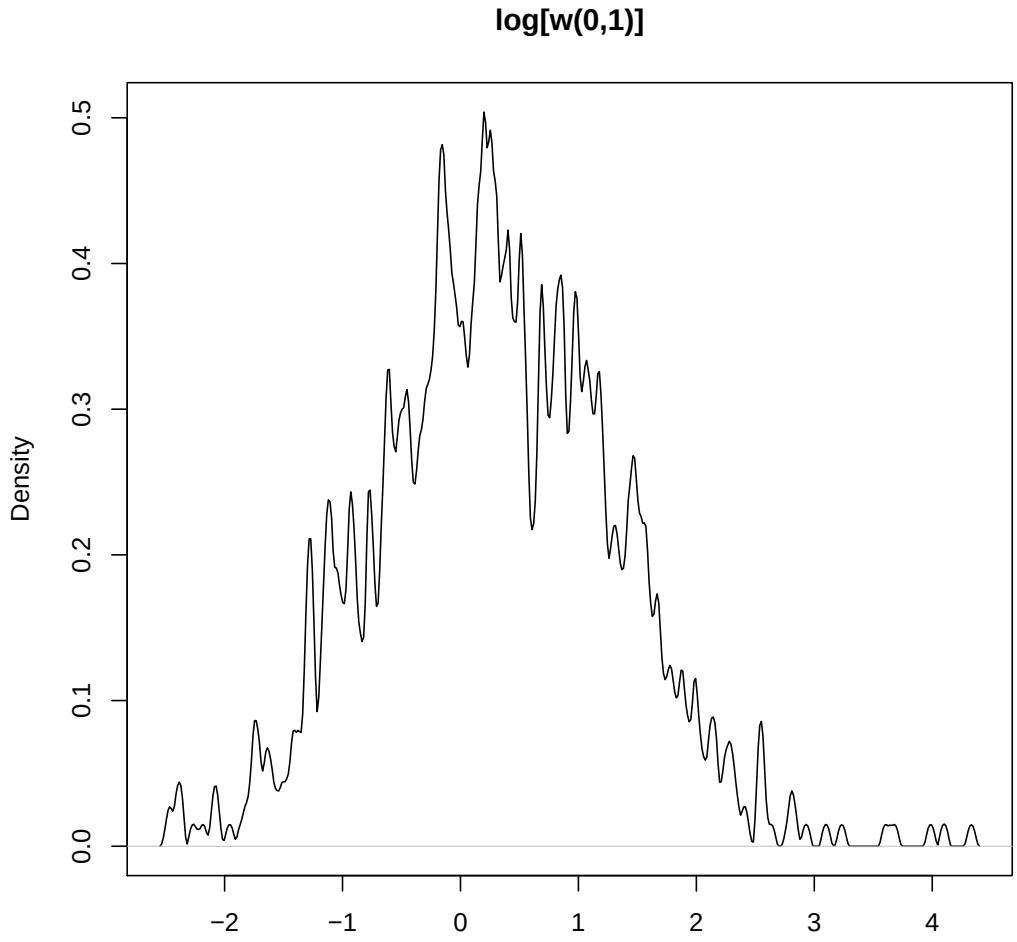

Figure 4: Distributions of $\log[\hat{w}(0,1)] = \log[\hat{\Pr}(T=1|X)] - \log[\hat{\Pr}(T=0|X)]$ over 5000 observations when the incorrect working exposure model in simulation study 1 is used. Only non-zero values of $\hat{w}(0,1)$ showed.

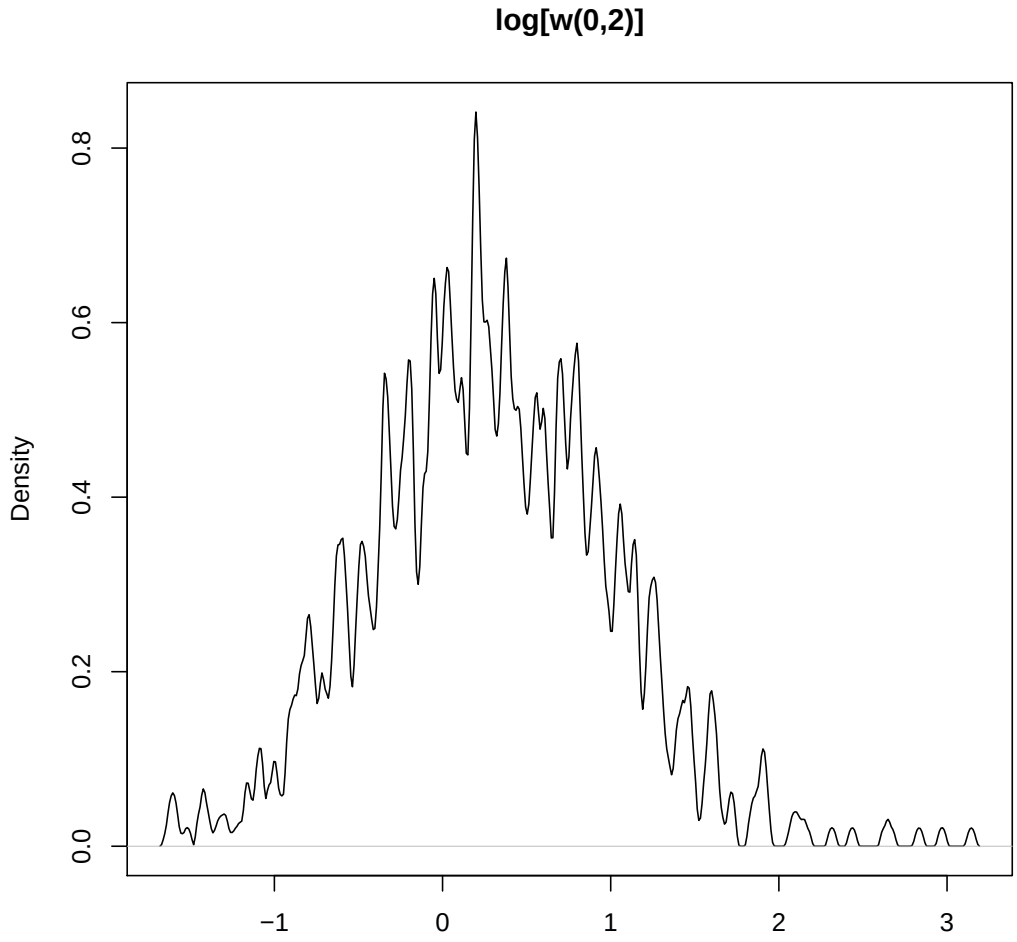

Figure 5: Distributions of $\log[\hat{w}(0,2)] = \log[\hat{\Pr}(T=2|X)] - \log[\hat{\Pr}(T=0|X)]$ over 5000 observations when the incorrect working exposure model in simulation study 1 is used. Only non-zero values of $\hat{w}(0,2)$ showed.

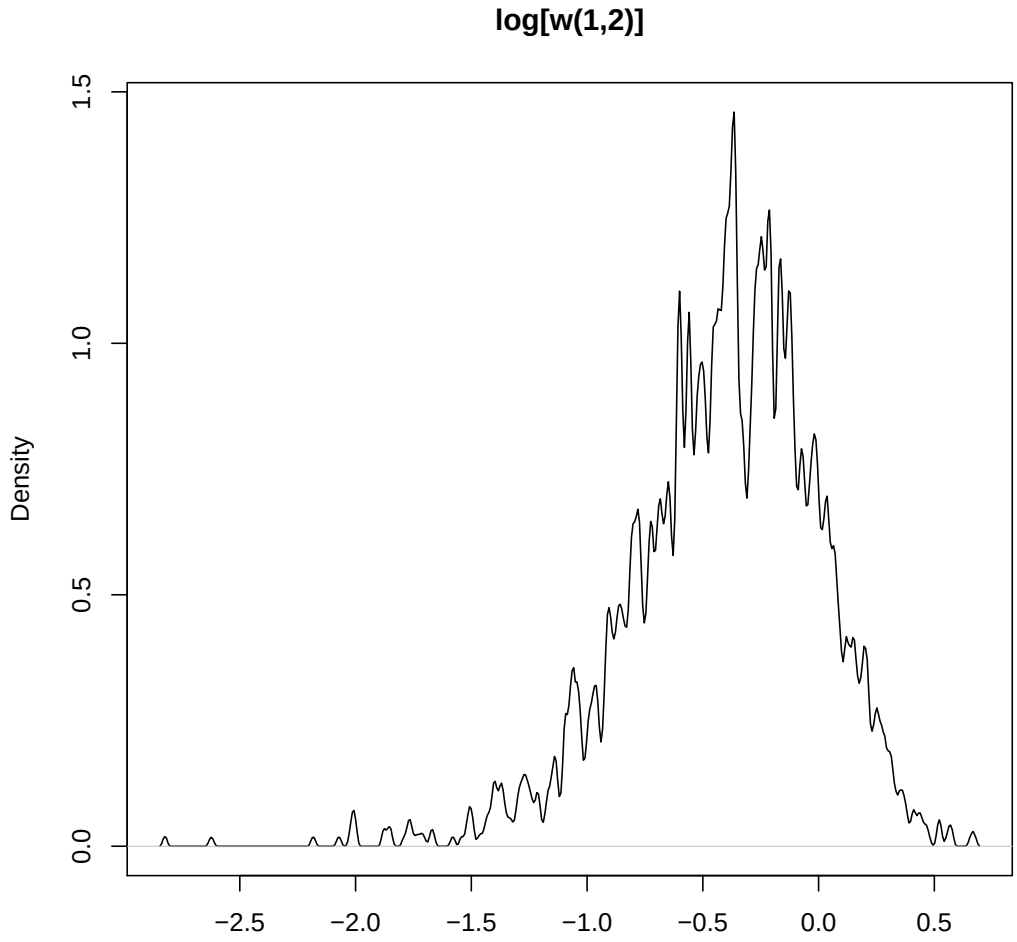

Figure 6: Distributions of $\log[\hat{w}(1,2)] = \log[\hat{\Pr}(T=2|X)] - \log[\hat{\Pr}(T=1|X)]$ over 5000 observations when the incorrect working exposure model in simulation study 1 is used. Only non-zero values of $\hat{w}(1,2)$ showed.

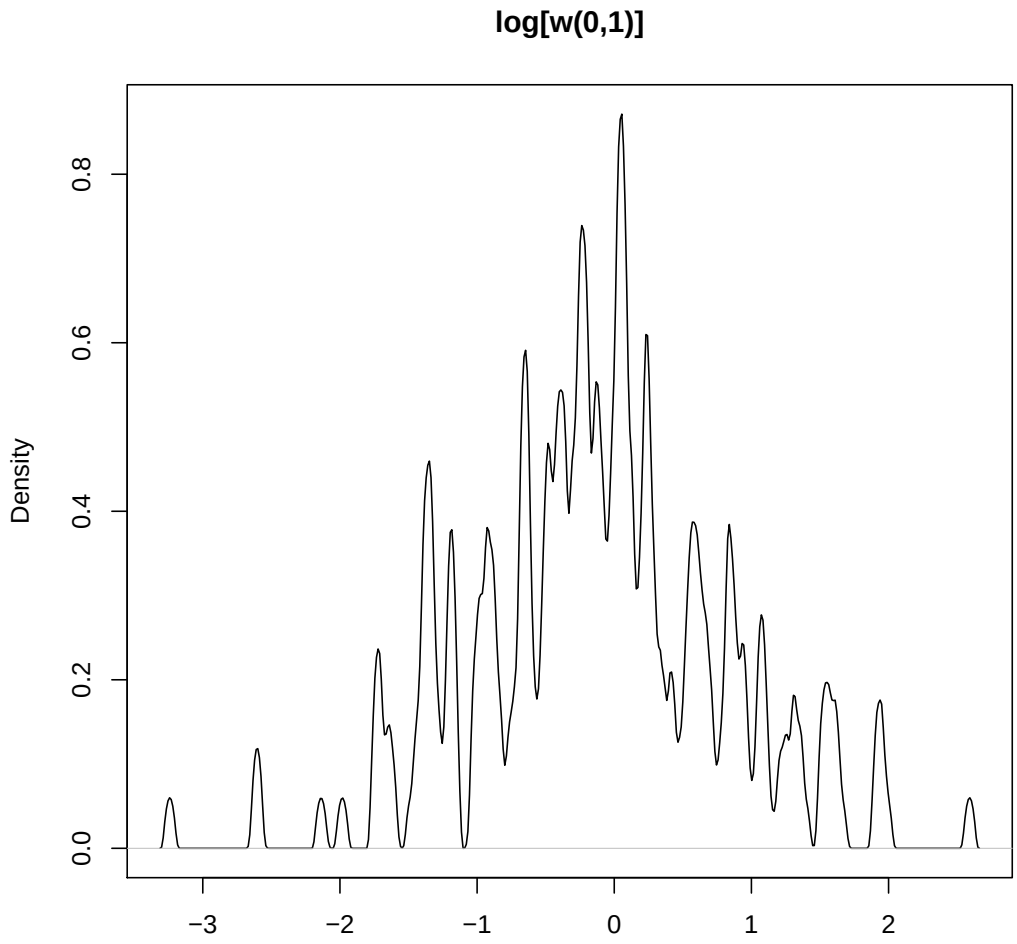

Figure 7: Distributions of $\log[\hat{w}(0,1)] = \log[\hat{\Pr}(T=1|X)] - \log[\hat{\Pr}(T=0|X)]$ over 5000 observations using a correct model for the exposure in simulation study 2.

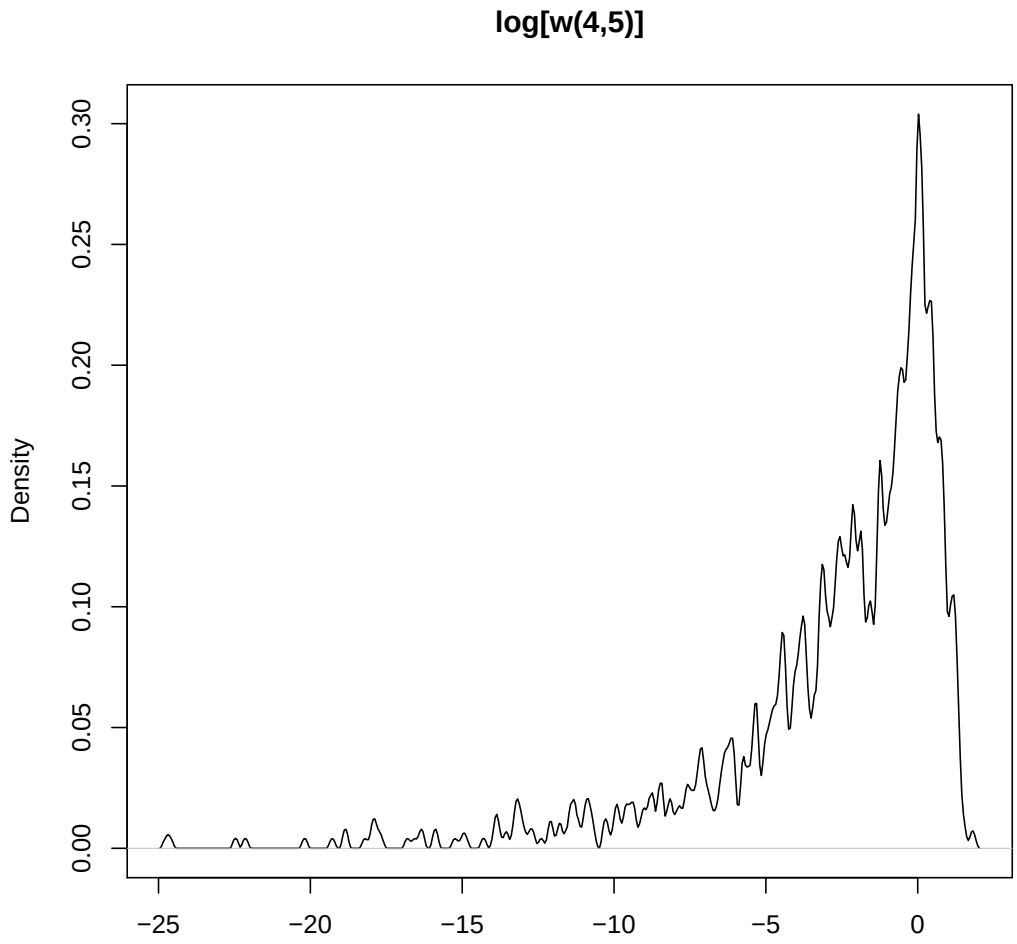

Figure 8: Distributions of $\log[\hat{w}(4,5)] = \log[\hat{\Pr}(T=5|X)] - \log[\hat{\Pr}(T=4|X)]$ over 5000 observations using a correct model for the exposure in simulation study 2.

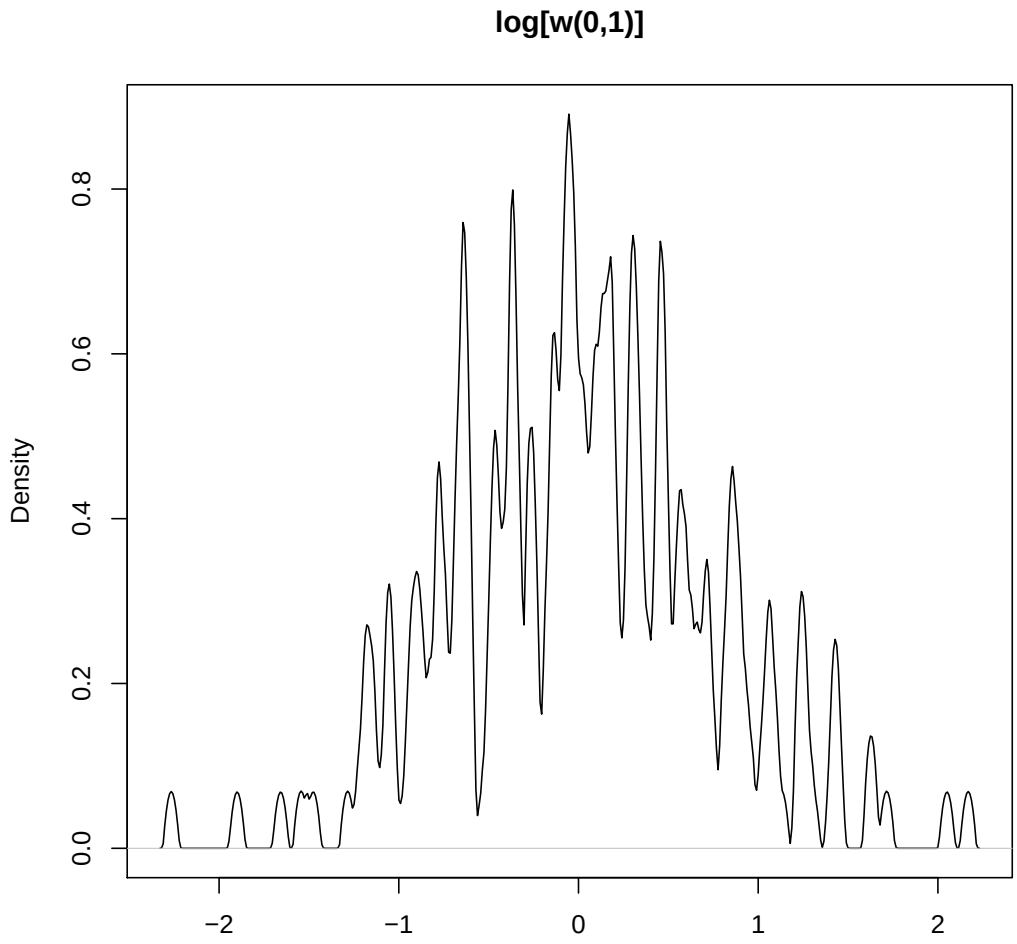

Figure 9: Distributions of $\log[\hat{w}(0,1)] = \log[\hat{\Pr}(T=1|X)] - \log[\hat{\Pr}(T=0|X)]$ over 5000 observations using an incorrect model for the exposure in simulation study 2.

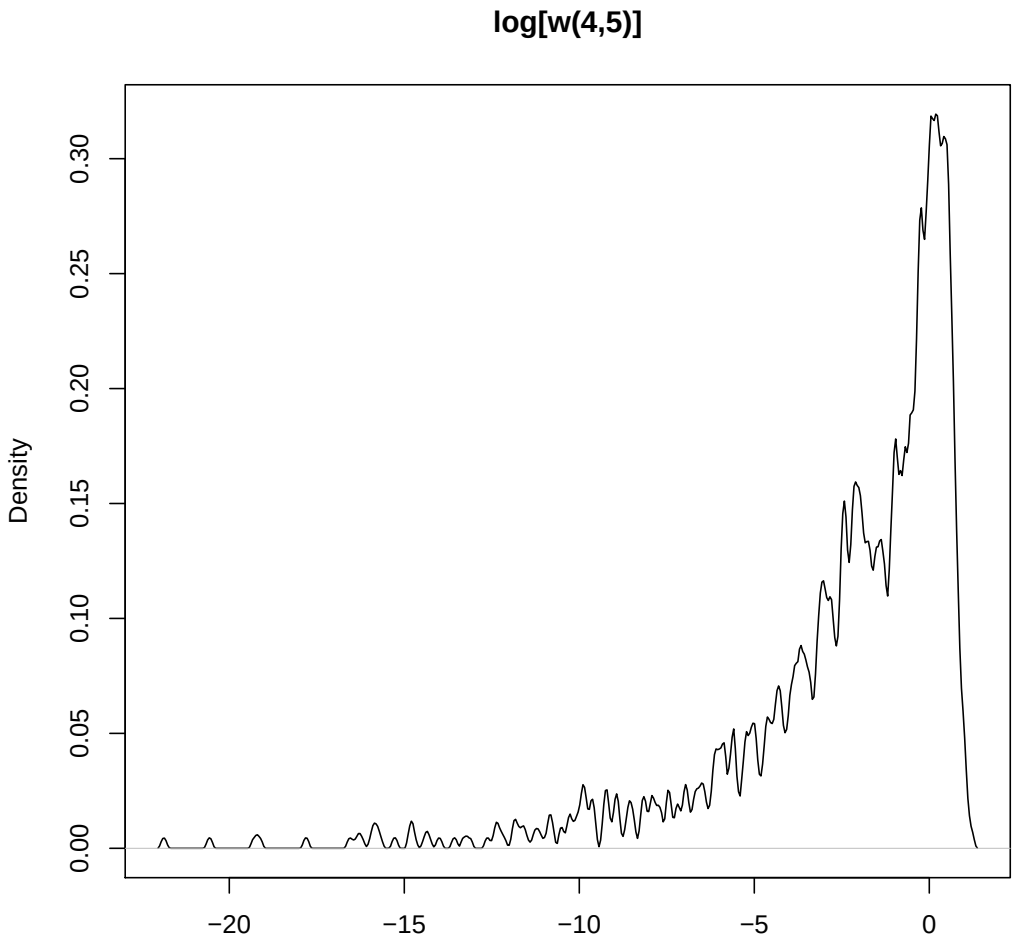

Figure 10: Distributions of $\log(\hat{w}(4,5)) = \log[\hat{\Pr}(T=5|X)] - \log[\hat{\Pr}(T=4|X)]$ over 5000 observations using an incorrect model for the exposure in simulation study 2.

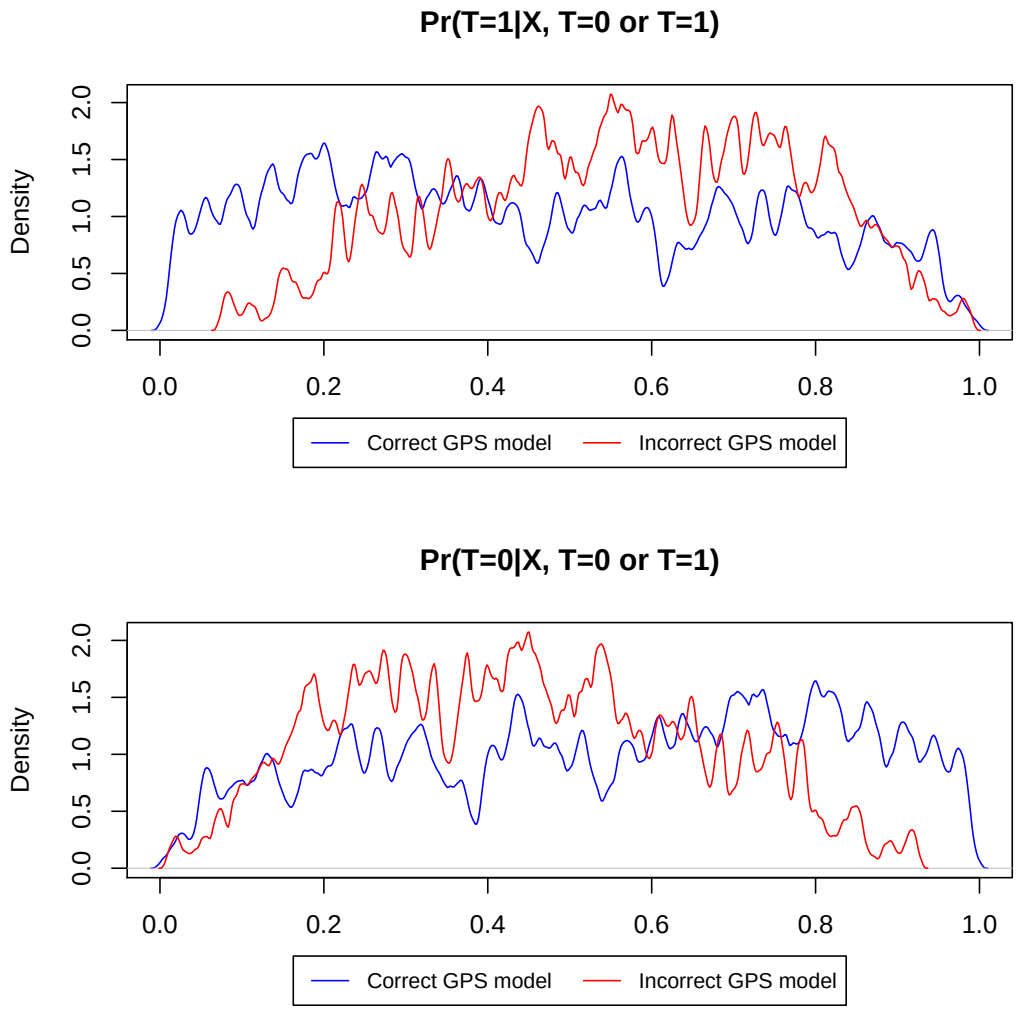

Figure 11: Distributions of $\hat{\Pr}(T=1|X,T\in\{0,1\})$ and $\hat{\Pr}(T=0|X,T\in\{0,1\})$ over 5000 observations in simulation study 1.

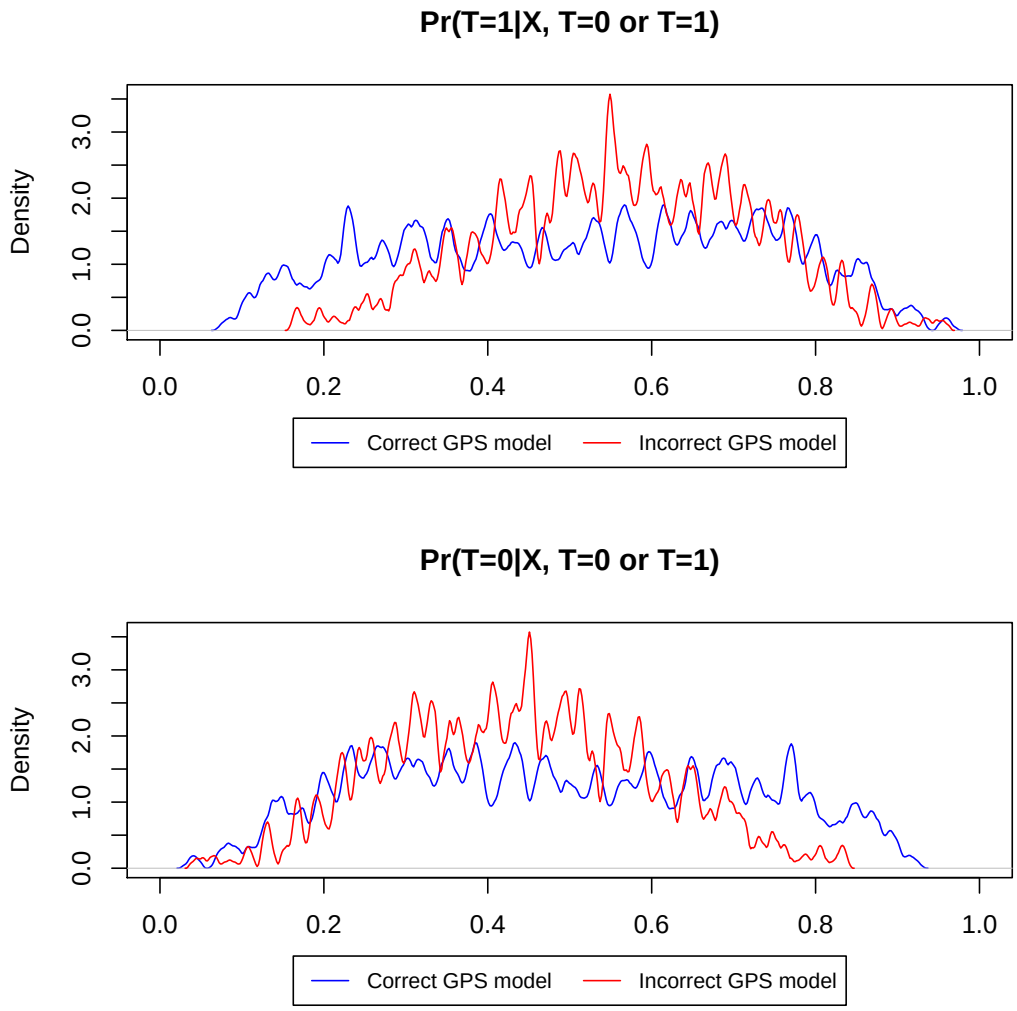

Figure 12: Distributions of $\hat{\Pr}(T=2|X,T\in\{0,2\})$ and $\hat{\Pr}(T=0|X,T\in\{0,2\})$ over 5000 observations in simulation study 1.

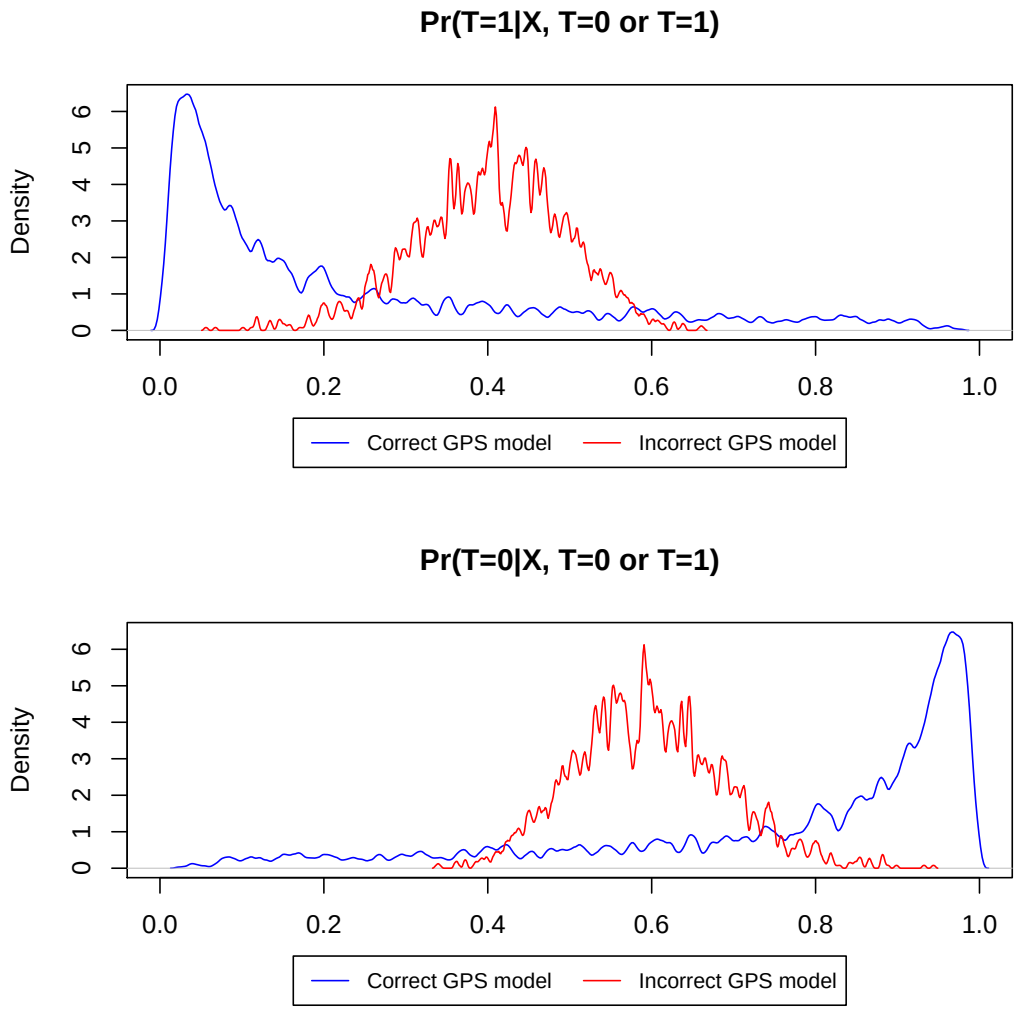

Figure 13: Distributions of $\hat{\Pr}(T=2|X,T\in\{1,2\})$ and $\hat{\Pr}(T=1|X,T\in\{1,2\})$ over 5000 observations in simulation study 1.

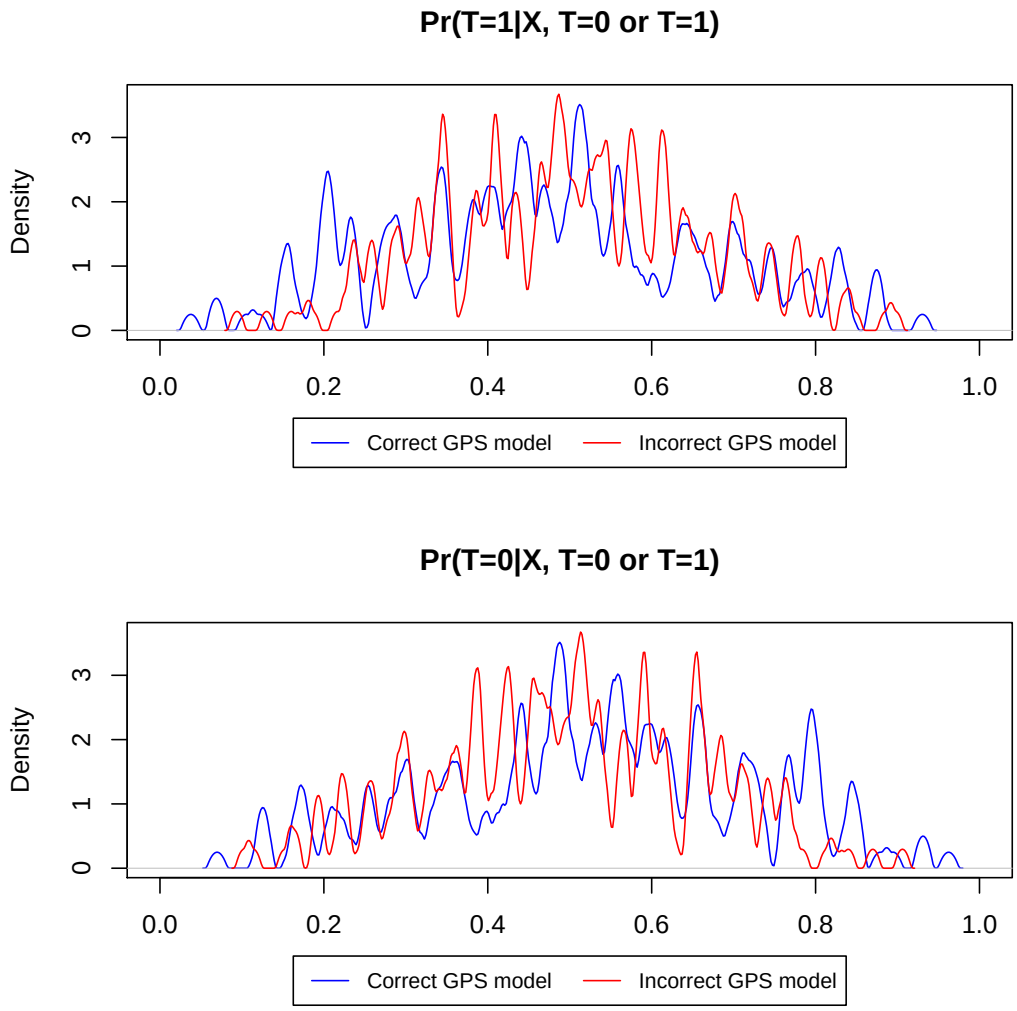

Figure 14: Distributions of $\hat{\Pr}(T=1|X,T\in\{0,1\})$ and $\hat{\Pr}(T=0|X,T\in\{0,1\})$ over 5000 observations in simulations study 2.

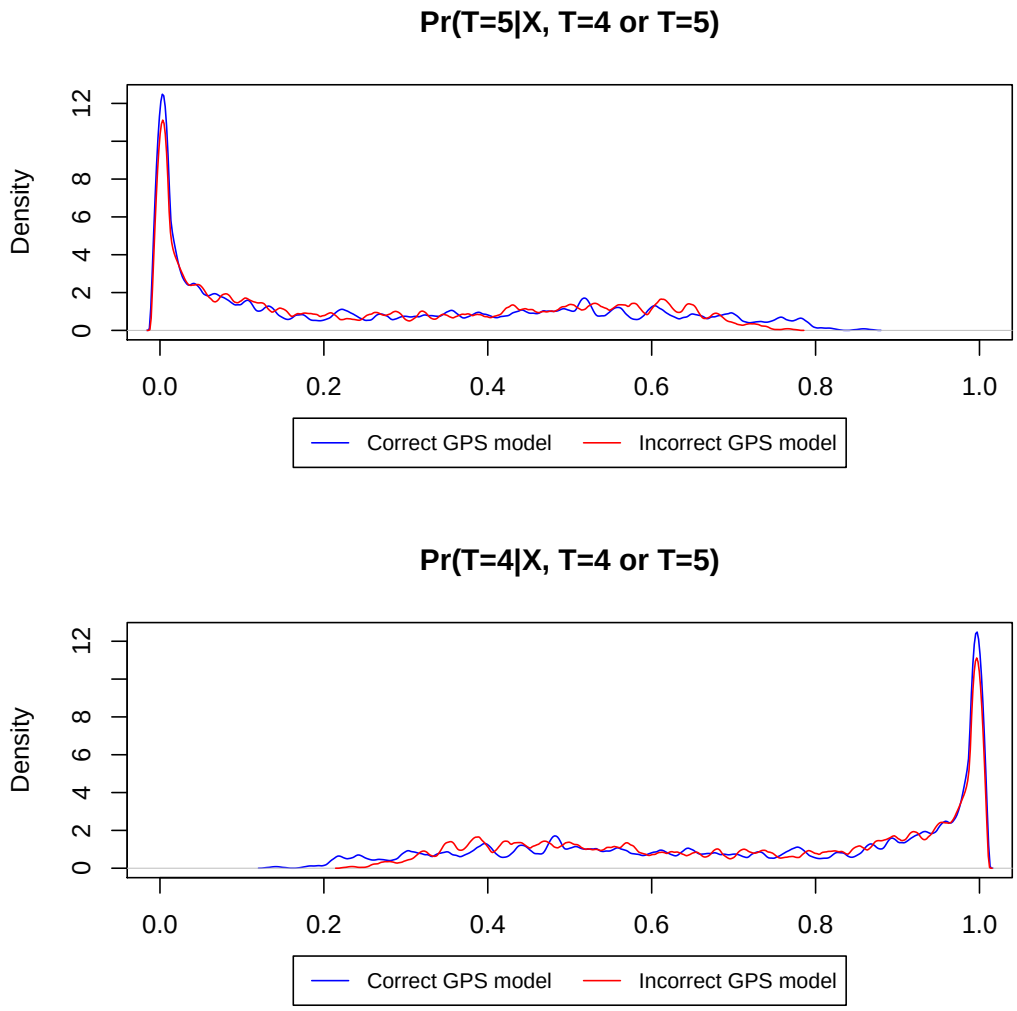

Figure 15: Distributions of $\hat{\Pr}(T = 5|X, T \in \{4,5\})$ and $\hat{\Pr}(T = 4|X, T \in \{4,5\})$ over 5000 observations in simulations study 2.