Censored count data regression with missing censoring information

Bilel Bousselmi, Jean-François Dupuy, Abderrazek Karoui

To cite this version:


HAL Id: hal-03049769
https://hal.science/hal-03049769
Submitted on 10 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Censored count data regression with missing censoring information

Bilel BOUSSELMI\textsuperscript{a}, Jean-François DUPUY\textsuperscript{a}, Abderrazek KAROUI\textsuperscript{b}

\textsuperscript{a}Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France
\textsuperscript{b}University of Carthage, Department of Mathematics, Faculty of Sciences of Bizerte, Tunisia

Abstract

We investigate estimation in the Poisson regression model when the count response is right-censored and the censoring indicators are missing at random. We propose several estimators based on the regression calibration, multiple imputation and augmented inverse probability weighting methods. Under appropriate regularity conditions, we prove the consistency of our estimators and we derive their asymptotic distributions. Simulation experiments are carried out to investigate the finite sample behaviour and relative performance of the proposed estimates.

Keywords: Poisson regression, Asymptotic properties, Missing data, Regression calibration, Multiple imputation, Augmented inverse probability weighting, Simulations

1. Introduction

Poisson regression is a popular tool for modeling the relationship between a count response (such as the number of cases of a specific disease in epidemiology, or the number of insurance claims within a given period of time) and a set of predictors or covariates. Over the past years, Poisson regression has been extended to accommodate censored count data. Although censoring is usually associated to lifetime data analysis, count data can also be censored, the most common type being right-censoring, which occurs when it is only known that the true count is higher than the observed one. For example, consider a study investigating the smoking habits of some population, where people report their number of cigarettes smoked per day. If one possible answer is "20 cigarettes or more", all cigarettes counts greater than 20 are right-censored at 20. Ignoring censoring is known to yield biased estimates and thus, incorrect inferences. Statistical inference in the censored Poisson regression model and its extensions was therefore addressed by several authors; see, for example, Terza (1985), Caudill and Mixon (1995), Famoye and Wang (2004), Xie and Wei (2007), Mahmoud and Alderiny (2010) for censored generalized Poisson regression, Karlis et al. (2016) for finite mixtures of censored Poisson regressions and Saffari and Adnan (2011) and Nguyen and Dupuy (2019, 2020) for zero-inflated censored Poisson regression.

Censored models for count data can be conveniently specified by introducing a censoring indicator which is set to 1 if the observed count is not censored and 0 otherwise. In this paper, we consider the situation where the censoring indicator is missing for some sample individuals. In the context
of survival analysis, this issue has been considered by several authors. For example, [Van Der Laan and McKeague (1998) and Subramanian (2006, 2011)] address estimation of the survival function of a random survival time with missing censoring indicators. [McKeague and Subramanian (1998)] consider estimation in Cox proportional hazards regression model with missing completely at random censoring indicators. [Wang et al. (2012) and Brunel et al. (2013)] propose various nonparametric estimates of the hazard and conditional hazard functions with censoring indicators missing at random. [Wang and Dinse (2011)] investigate estimation in the linear regression model for survival data with missing censoring indicators. Estimation in censored Poisson regression with missing censoring information is still an open problem. Our aim in this paper is to provide and compare several estimates adapted to this setting.

Missing data problems have given rise to a rich literature and several estimation methods have been proposed for this setting. A common and simple approach is to exclude individuals with missing data. This is the so-called complete-case analysis. This method can induce bias and substantial variance increase. Two alternatives are regression calibration and multiple imputation. In regression calibration, missing data are replaced by their conditional expectation given the observed data. In multiple imputation, missing data are replaced by data generated from an imputation model. This imputation is repeated $M$ times, generating $M$ completed data sets. Each of them is analysed and an overall estimator is obtained by combining the estimates of the $M$ completed samples. Both methods require a model for the missing data given the observed data. Inverse probability weighting constitutes another alternative method for dealing with missing data (see for example Seaman and White (2013) for a review of this method). Similarly to the complete-case analysis, inverse probability weighting only uses complete cases, but weights are used to rebalance the set of complete cases. Calculating these weights requires a model for the probability that an individual has complete data. Augmented inverse probability weighting was then proposed to ensure robustness against misspecification of the missingness model (see, for example, Tsiatis (2007) for a detailed account on the method).

In this paper, we investigate, both theoretically and numerically, the regression calibration, multiple imputation and augmented inverse probability weighting estimators of the regression parameter in the censored Poisson regression model with missing censoring indicators. Our analysis of these estimates will be based on parametric assumptions for the conditional models for missing data and the missingness mechanism. The plan of the paper is as follows. In Section 2, we describe the model setup and we introduce the notations that will be used throughout the paper. In Section 3, we introduce our regression calibration estimator and we establish its consistency and asymptotic normality. In Sections 4 and 5, we propose our multiple imputation and augmented inverse probability weighted estimators, and we derive their asymptotic properties. All our theoretical derivations are based on an incomplete gamma function formulation of the distribution function of the Poisson regression model. Consistent asymptotic variance estimates are also proposed for the regression calibration, multiple imputation and augmented inverse probability weighted estimators. In Section 6, we conduct a simulation study to assess the finite sample performance and robustness to parametric assumptions of the proposed estimates. Discussion and perspectives are given in Section 7. All proofs are deferred to appendices.

2. Model, data, notations

Let $Y$ denote the count of interest and $X = (1, X_2, \ldots, X_p)^\top$ be a $p$-vector of covariates ($\top$ denotes the transpose operator). We assume that the conditional distribution of $Y$ given $X$ is given
by a Poisson regression model with parameter \( \lambda = \exp(\beta^\top X) \), where \( \beta \in \mathbb{R}^p \) is a vector of unknown parameters.

We consider the situation where \( Y \) can be right-censored, that is, instead of the true \( Y \), we eventually observe a value which is smaller than \( Y \). This can be formalised by introducing a finite random variable \( C \) such that we observe either \( Y \) if \( Y < C \) or \( C \) if \( Y \geq C \), and an indicator \( \delta \) (called censoring indicator thereafter) which is equal to 1 if \( Y < C \) and 0 if \( Y \geq C \). Finally, we denote by \( Y^* \) the observed count value (that is, \( Y^* = \min(Y, C) \)).

Assume that \( n \) independent individuals are available and that for each of them, we observe the triplet \((Y_i^*, X_i, \delta_i)\) (with \( i \in \{1, \ldots, n\} \)). Based on these observations, the likelihood of \( \beta \) is calculated as:

\[
L_n(\beta) = \prod_{i=1}^{n} \mathbb{P}(Y_i = Y_i^* | X_i)^{\delta_i} \mathbb{P}(Y_i \geq Y_i^* | X_i)^{1-\delta_i},
\]

from which we easily deduce the loglikelihood \( \ell_n(\beta) = \log L_n(\beta) \):

\[
\ell_n(\beta) = \sum_{i=1}^{n} \left\{ \delta_i \left( Y_i^* \beta^\top X_i - e^{\beta^\top X_i - \log(Y_i^*!)} \right) + (1 - \delta_i) \log \left( 1 - \sum_{k=0}^{Y_i^*-1} \frac{e^{-\exp(\beta^\top X_i) + k\beta^\top X_i}}{k!} \right) \right\}.
\]

By standard asymptotic theory, the maximum likelihood estimator \( \hat{\beta}_n = \arg \max_{\beta} \ell_n(\beta) \) is consistent and asymptotically normal with variance \(-\mathbb{E}[\partial^2 \ell_1(\beta) / \partial \beta \partial \beta^\top]\).

Now, we consider the situation where some additional uncertainty can arise in the observations. Precisely, we consider the situation where the censoring indicator \( \delta_i \) is missing for some individuals. Let \( \xi \) be a missingness indicator, that is, \( \xi = 1 \) if \( \delta \) is observed and \( \xi = 0 \) otherwise. Then, for individual \( i \in \{1, \ldots, n\} \), the observed data are

\[
(Y_i^*, X_i, \delta_i, \xi_i = 1) \quad \text{or} \quad (Y_i^*, X_i, \xi_i = 0).
\]

We consider a missing at random (MAR) mechanism, which means that \( \xi \) and \( \delta \) are independent given all other observed variables (a more restrictive assumption is that \( \xi \) and \( \delta \) are independent, which is called "missing completely at random"). In the next sections, we propose, investigate and compare several estimators of \( \beta \) in this context.

### 3. Regression calibration estimation

#### 3.1. The proposed estimator

Our first estimator is based on the regression calibration idea. It consists in replacing any missing \( \delta_i \) in (2.1) by its conditional expectation \( \mathbb{E}(\delta_i | W_i) \), where \( W_i \) contains the observed variables \( Y_i^* \) and \( X_i \) and eventually (if available) some observed surrogate variables \( V_i \) for \( \delta_i \). Thus, we let \( W_i = (Y_i^*, X_i^\top, V_i^\top)^\top \) (we denote by \( q \) the dimension of \( W_i \)). An approximated version of \( \delta_i \) can then be defined as:

\[
\hat{\delta}_i = \xi_i \delta_i + (1 - \xi_i) \mathbb{E}(\delta_i | W_i).
\]

The conditional expectation \( \mathbb{E}(\delta_i | W_i) \) (or conditional probability \( \mathbb{P}(\delta_i = 1 | W_i) \)) will generally be unknown and will have to be estimated. As is usual with the regression calibration approach, we assume that \( \mathbb{E}(\delta_i | W_i) \) can be specified by a parametric model \( m(W_i, \theta) \), where \( \theta \) is an unknown \( q \)-dimensional parameter with true value \( \theta_0 \).
Remark 1. A convenient candidate for \(m(\cdot, \cdot)\) is the logistic regression model \(m(W_i, \theta) = \log^{-1}(\theta^T W_i)\) but other choices, such as the probit, are possible. One may also allow for polynomial, spline and interaction terms in these models, in order to make them as flexible as desired. In what follows, we assume a general model \(m(W_i, \theta)\) with some regularity conditions stated in section 3.2.

At a first stage, we estimate \(\theta_0\) by maximizing a likelihood based on complete cases \(i \in \{1, \ldots, n|\xi_i = 1\}\) only:

\[
\hat{\theta}_n = \arg\max_\theta \prod_{i=1}^n m(W_i, \theta)^{\xi_i \delta_i} (1 - m(W_i, \theta))^{\xi_i (1-\delta_i)}. \tag{3.3}
\]

Let

\[
m(W_i, \theta) = \frac{\partial m(W_i, \theta)}{\partial \theta}, \quad \tilde{m}_i(\theta) = \frac{\hat{m}(W_i, \theta)}{m(W_i, \theta)(1 - m(W_i, \theta))},
\]

and

\[
\Theta(\theta) = \mathbb{E} \left[ \frac{\hat{m}^{\otimes 2}(W_i, \theta)}{m(W, \theta)(1 - m(W, \theta))} \right],
\]

where for any column vector \(u, u^{\otimes 2} = uu^T\). Then it is rather straightforward to see that \(\hat{\theta}_n\) is asymptotically linear with influence function \(\Theta^{-1}(\theta_0)\tilde{m}_i(\theta_0)\xi_i(\delta_i - m(W_i, \theta_0))\), that is:

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Theta^{-1}(\theta_0)\tilde{m}_i(\theta_0)\xi_i(\delta_i - m(W_i, \theta_0)) + o_P(1). \tag{3.4}
\]

Finally, it will be useful to note that if \(Y\) is distributed as Poisson with parameter \(\lambda\), then for any \(u \in \mathbb{N}, \mathbb{P}(Y \leq u) = \sum_{k=0}^u \exp(-\lambda)\lambda^k/k! = \Gamma(u + 1, \lambda)/u!\) where \(\Gamma(u, \lambda) = \int_\lambda^\infty t^{u-1} \exp(-t)dt\) is the incomplete gamma function, whose derivative with respect to \(\lambda\) is given by \(\partial \Gamma(u, \lambda)/\partial \lambda = -\exp(-\lambda)\lambda^{u-1}\).

Now, letting \(\hat{\delta}_i(\theta) = \xi_i \delta_i + (1 - \xi_i) m(W_i, \theta)\) be the approximation of \(\delta_i\) based on model \(m(W_i, \theta)\), we define our regression calibration estimator of \(\beta\) as

\[
\tilde{\beta}_n = \arg\max_\beta \tilde{\ell}_n(\beta, \hat{\theta}_n),
\]

where

\[
\tilde{\ell}_n(\beta, \hat{\theta}_n) = \sum_{i=1}^n \left\{ \hat{\delta}_i(\hat{\theta}_n) \left( (Y_i^* \beta^T X_i - e^{\beta^T X_i} - \log(Y_i^*)) + (1 - \hat{\delta}_i(\hat{\theta}_n)) \log \left( \frac{\Gamma(Y_i^*, e^{\beta^T X_i})}{(Y_i^* - 1)!} \right) \right) \right\}
\]

is an approximated version of (2.1).

3.2. Regularity conditions and asymptotic results

The following regularity conditions are needed to establish the asymptotic properties of the regression calibration estimator. We assume:

C1 The covariates vectors \(X_i\) and \(V_i\) are bounded, for every \(i = 1, 2, \ldots\)
The true parameter values $\beta_0$ and $\theta_0$ lie in the interior of some bounded sets $B \subset \mathbb{R}^p$ and $\Theta \subset \mathbb{R}^q$ respectively.

We have $\mathbb{P}(Y^* \geq 1 | \xi = 0) = 1$ and $\mathbb{P}(\delta = 1) > 0$.

The function $m(w, \theta)$ is differentiable with respect to $\theta$, for every $w$. For every $\theta, \hat{\theta} \in \Theta$, 
$$|m(w, \theta) - m(w, \hat{\theta})| \leq h(\mathbf{w})||\theta - \hat{\theta}||$$
for some bounded function $h$, with $\mathbb{E}[h(W)] = v$.

Remark 2. Condition C3 requires that a minimum amount of information is available on the count response when it is either censored ($\delta = 0$) or its censoring status is unknown ($\xi = 0$). Intuitively, the observation $\{Y^* = 0\}$ carries no information if it is unknown that $\delta = 1$ (i.e., that it is a "genuine" zero count), since all counts are non-negative.

Before stating the asymptotics of $\hat{\beta}_n$, we introduce some further notations. Let $h_\beta$ be the function defined by:

$$h_\beta(y, x) = \frac{e^{-e^{\beta^T x} + \beta^T x} y}{(y - 1)! - \Gamma(y, e^{\beta^T x})} \tag{3.5}$$

for any $\beta \in \mathbb{R}^p$, $x \in \mathbb{R}^p$ and $y \in \mathbb{N}\setminus\{0\}$. Let also $\pi(W) = \mathbb{P}(\xi = 1 | W)$ and define the matrices

$$\Sigma_1(\beta) = \mathbb{E} \left[ XX^\top \left( \delta e^{\beta^T X} + (\delta - 1) \left\{ Y^* - e^{\beta^T X} - h_\beta(Y^*, X) \right\} h_\beta(Y^*, X) \right) \right],$$

$$\Sigma_2(\beta, \theta) = \mathbb{E} \left[ X n^\top(W, \theta) \left( Y^* - e^{\beta^T X} - h_\beta(Y^*, X) \right) (1 - \pi(W)) \right],$$

$$\Sigma_3(\beta, \theta) = \mathbb{E} \left[ X n^\top(W, \theta) \left( Y^* - e^{\beta^T X} - h_\beta(Y^*, X) \right) \right].$$

We are now in position to state our first theorem. The proof is given in Appendix A.

Theorem 3.1. Assume that conditions C1-C4 hold. Then $\hat{\beta}_n \overset{\mathbb{P}}{\longrightarrow} \beta_0$ as $n \to \infty$ and $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is asymptotically normal with mean zero and variance $\Sigma$, where

$$\Sigma = \Sigma^{-1}_1(\beta_0) \left\{ \Sigma_1(\beta_0) + (2\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)) \Theta^{-1}(\theta_0) \Sigma_2^\top(\beta_0, \theta_0) \right\} \Sigma^{-1}_1(\beta_0).$$

Remark 3. If $\pi(W)$ is identically equal to 1 (that is, if there is no missing data), $\Sigma$ reduces to the asymptotic variance of the maximum likelihood estimator $\hat{\beta}_n$ in \[2.1\], which in turn reduces to the usual asymptotic variance $\left( \mathbb{E}[XX^\top e^{\beta_0^T X}] \right)^{-1}$ in Poisson regression if $m(W, \theta_0)$ is identically equal to 1 (that is, no censoring can affect the data).

A consistent estimator of $\Sigma$ is given by

$$\Sigma^{-1}_1(\hat{\beta}_n, \hat{\theta}_n) \left\{ \Sigma_{1,n}(\hat{\beta}_n, \hat{\theta}_n) + \left( 2\Sigma_{3,n}(\hat{\beta}_n, \hat{\theta}_n) - \Sigma_{2,n}(\hat{\beta}_n, \hat{\theta}_n) \right) \Theta^{-1}(\hat{\theta}_n) \Sigma_{2,n}^\top(\hat{\beta}_n, \hat{\theta}_n) \right\} \Sigma^{-1}_1(\hat{\beta}_n, \hat{\theta}_n).$$

5
The final multiple imputation estimator

where

\[ \Sigma_{1,n}(\beta, \theta) = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \left( \delta_i(\theta)e^{\beta^T X_i} + (\delta_i(\theta) - 1) \left\{ Y_i^* - e^{\beta^T X_i} - h_\beta(Y_i^*, X_i) \right\} h_\beta(Y_i^*, X_i) \right), \]

\[ \Sigma_{2,n}(\beta, \theta) = \frac{1}{n} \sum_{i=1}^{n} X_i \hat{m}^\top(W_i, \theta) \left( Y_i^* - e^{\beta^T X_i} - h_\beta(Y_i^*, X_i) \right) (1 - \xi_i), \]

\[ \Sigma_{3,n}(\beta, \theta) = \frac{1}{n} \sum_{i=1}^{n} X_i \hat{m}^\top(W_i, \theta) \left( Y_i^* - e^{\beta^T X_i} - h_\beta(Y_i^*, X_i) \right), \]

\[ \Theta_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{m} \otimes^2(W_i, \theta)}{m(W_i, \theta)(1 - m(W_i, \theta))} \xi_i. \]

The consistency proof of the variance estimator uses similar arguments as the proof of consistency of \( \hat{\beta}_n \), it is thus omitted. The estimator \( \hat{\beta}_n \) will be evaluated in the simulation study of Section 6.

Several methods have been proposed to address missing data problems in regression. Among them is the multiple imputation, which provides an alternative, popular and widely-used approach. The basic idea is to create several (say \( M \)) completed data sets, by filling in plausible values for the missing data. Then, each filled sample is analysed as if it were the complete data set. Finally, the \( M \) imputed-samples inferences are combined into a single overall inference. In the next section, we investigate this approach for estimating \( \beta \) in our problem.

4. Multiple imputation

In this section, we assume, as in Section 3, that the conditional expectation \( \mathbb{E}(\delta_i | W_i) \) can be specified by a parametric model \( m(W_i, \theta_0) \), and we denote by \( \hat{\theta}_n \) the maximum likelihood estimator of \( \theta_0 \) based on the complete cases \( i \in \{1, \ldots, n\} | \xi_i = 1 \).

The imputation procedure is as follows. Each missing \( \delta_i \) is replaced by a random draw from the Bernoulli distribution \( \mathcal{B}(m(W_i, \theta_0)) \). We obtain a completed data set. This procedure is repeated \( M \) times to form \( M \) imputed data sets. For a given \( \theta \), let \( D_{i,j}(\theta) \sim \mathcal{B}(m(W_i, \theta)) \) denote the imputation of \( \delta_i \) in the \( j \)-th completed data set (\( j = 1, \ldots, M \)). Let also

\[ \delta_{i,j}^*(\theta) = \xi_i \delta_i + (1 - \xi_i) D_{i,j}(\theta) \]

be the random variable which is equal to \( \delta_i \) is \( \xi_i = 1 \) (that is, if \( \delta_i \) is observed) and to \( D_{i,j}(\theta) \) if \( \xi_i = 0 \) (that is, if \( \delta_i \) is missing) (note the difference between the imputation method, where \( \delta_{i,j}^*(\theta) \in \{0, 1\} \), and the regression calibration approach, where \( \hat{\delta}_i(\theta) \in [0, 1] \)). A single-imputation estimator \( \hat{\beta}_{n,j}^* \) of \( \beta_0 \) is obtained by maximizing the imputed log-likelihood

\[ \ell_{i,j}^*(\beta, \hat{\theta}_n) = \sum_{i=1}^{n} \left\{ \delta_{i,j}^*(\hat{\theta}_n) \left( Y_i^* \beta^T X_i - e^{\beta^T X_i} - \log(Y_i^*) \right) + (1 - \delta_{i,j}^*(\hat{\theta}_n)) \log \left( 1 - \Gamma \left( \frac{Y_i^* \beta^T X_i}{Y_i^* - 1} \right) \right) \right\}. \]

The final multiple imputation estimator \( \hat{\beta}_n^* \) is obtained by averaging the \( M \) estimators \( \hat{\beta}_{n,j}^* \) as:

\[ \hat{\beta}_n^* = \frac{1}{M} \sum_{j=1}^{M} \hat{\beta}_{n,j}^*. \]
Theorem 4.1. For \( j = 1, \ldots, M \), let \( f_{\beta,\theta,j}(O_i) = X_i(\delta_{i,j}(\theta)[Y_i^* e^{\beta^T X_i} - h_\beta(Y_i^*, X_i)] + h_\beta(Y_i^*, X_i)) \), where \( O_i \) denotes the observation \((2.2)\). Let also \( \Sigma^*_1(\beta, \theta) = \text{var}(\frac{1}{M} \sum_{j=1}^{M} f_{\beta,\theta,j}(O_1)) \). If conditions C1-C4 hold, then \( \hat{\beta}_n \xrightarrow{p} \beta_0 \) as \( n \to \infty \) and \( \sqrt{n}(\hat{\beta}_n - \beta_0) \) is asymptotically normal with mean zero and variance \( \Sigma^* \), where

\[
\Sigma^* = \Sigma_1^{-1}(\beta_0) \left\{ \Sigma_1^*(\beta_0, \theta_0) + (2\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)) \right\} \Sigma_1^{-1}(\beta_0).
\]

A consistent estimator of \( \Sigma^* \) can be obtained as

\[
\tilde{\Sigma}_{1,n}^{-1}(\hat{\beta}_n, \hat{\theta}_n) \left\{ \Sigma_{1,n}^*(\hat{\beta}_n, \hat{\theta}_n) + (2\Sigma_{3,n}(\hat{\beta}_n, \hat{\theta}_n) - \Sigma_{2,n}(\hat{\beta}_n, \hat{\theta}_n)) \right\} \Theta_n^{-1}(\hat{\theta}_n) \Sigma_{2,n}^{-1}(\hat{\beta}_n, \hat{\theta}_n),
\]

where \( \Sigma_{1,n}^*(\beta, \theta) \) is the empirical covariance of the vectors \( \frac{1}{M} \sum_{j=1}^{M} f_{\beta,\theta,j}(O_i) \) \((i = 1, \ldots, n)\), \( \tilde{\Sigma}_{1,n}(\beta, \theta) \) is the average \( \frac{1}{M} \sum_{j=1}^{M} \tilde{\Sigma}_{1,n,j}(\beta, \theta) \), with

\[
\tilde{\Sigma}_{1,n,j}(\beta, \theta) = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \left( \delta_{i,j}(\theta) e^{\beta^T X_i} + (\delta_{i,j}(\theta) - 1) \right) \left( \beta^* e^{\beta^T X_i} - h_\beta(Y_i^*, X_i) \right),
\]

and \( \Sigma_{2,n}, \Sigma_{3,n} \) and \( \Theta_n \) are as given in Section 3.

Regression calibration and multiple imputation rely on the ability of the investigator to formulate an appropriate model for \( \mathbb{E}(\delta|W) \). Misspecifying this model is likely to yield biased estimates of the parameters of interest. An alternative approach is to specify the selection probabilities \( \pi(W_i) = \mathbb{P}(\xi_i = 1|W_i) \) and to use the inverse probability weighting (IPW) of complete-case technique of Horvitz and Thompson (1952). The basic idea of IPW is to adjust a complete-case analysis by weighting individuals with no missing data by the inverse of their selection probability. Selection probabilities are generally unknown and have to be estimated. Again, misspecifying the \( \pi(W_i) \) is likely to yield biased inference. Moreover, by discarding individuals with missing data, IPW is also known to yield loss of efficiency.

For these reasons, the augmented IPW approach (AIPW henceforth, see Robins et al. [1994]) was proposed to improve the basic IPW. Since its introduction, the method has been shown to be doubly robust in several models, such as the proportional hazards model (Wang and Chen [2001]), the single-index model (Guo et al. [2015]), the additive hazards model (Sun et al. [2017]) and the accelerated failure time model (Steingrimsson and Strawderman [2017]). Double robustness refers to the fact that the AIPW estimates are consistent as long as either the selection probability model or the conditional expectation of the missing data is correctly specified. In the next section, we propose an augmented IPW estimating equation adapted to our problem, and we investigate the asymptotic properties of the resulting estimator.

5. Augmented inverse probability weighted estimation

Inspired by Horvitz and Thompson (1952), the inverse probability weighting of complete cases has become a classical estimation method in missing data problems. One drawback of the method is that the observed variables of subjects with missing data are not fully used, except through the estimation of the unknown selection probabilities. The AIPW method improves IPW by introducing
an additional term involving contributions from individuals with some missing data (we refer to 
Tsiatis (2007) for a detailed account on the method and numerous references). Adapting this idea,
we propose the following augmented IPW estimating equation for $\beta$:

$$
\sum_{i=1}^{n} X_i \left[ \left\{ \frac{\xi_i \delta_i}{\pi(W_i)} + \left( 1 - \frac{\xi_i}{\pi(W_i)} \right) \mathbb{E}(\delta_i|W_i) \right\} \left( Y_i^* - e^{\beta^T X_i} - h_\beta(Y_i^*, Y_i^*) \right) + h_\beta(Y_i^*, X_i) \right].
$$

The quantities $\mathbb{E}(\delta_i|W_i)$ and $\pi(W_i)$ are unknown and have to be estimated. We assume that they can be specified by some parametric models $m(W_i, \theta)$ and $\pi(W_i, \gamma)$ respectively, where $\theta$ and $\gamma$ are unknown $q$-dimensional parameters with true values $\theta_0$ and $\gamma_0$. Let $\hat{\theta}_n$ and $\hat{\gamma}_n$ be the maximum likelihood estimates of $\theta_0$ and $\gamma_0$. $\hat{\theta}_n$ is given by (3.3). Similarly, $\hat{\gamma}_n$ can be obtained as

$$
\hat{\gamma}_n = \text{arg max}_{\gamma} \prod_{i=1}^{n} \pi(W_i, \gamma)^{\xi_i} (1 - \pi(W_i, \gamma))^{1-\xi_i}.$$

Finally, our AIPW estimator $\hat{\beta}_n$ of $\beta$ solves the estimating equation $\ell_n(\beta, \hat{\theta}_n, \hat{\gamma}_n) = 0$, where

$$
\ell_n(\beta, \hat{\theta}_n, \hat{\gamma}_n) = \sum_{i=1}^{n} X_i \left[ \left\{ \frac{\xi_i \delta_i}{\pi(W_i, \hat{\gamma}_n)} + \left( 1 - \frac{\xi_i}{\pi(W_i, \hat{\gamma}_n)} \right) m(W_i, \hat{\theta}_n) \right\} \left( Y_i^* - e^{\beta^T X_i} - h_\beta(Y_i^*, X_i) \right) + h_\beta(Y_i^*, X_i) \right].
$$

Before stating the asymptotic properties of $\hat{\beta}_n$, we introduce some further notations and regularity conditions. For any $\theta, \gamma \in \mathbb{R}^q$, we let

$$
\tilde{\chi}_i(\theta, \gamma) = \frac{\xi_i \delta_i}{\pi(W_i, \gamma)} + \left( 1 - \frac{\xi_i}{\pi(W_i, \gamma)} \right) m(W_i, \theta).
$$

Assuming the parametric model $\pi(W_i, \gamma)$ for the selection probabilities, the maximum likelihood estimator $\hat{\gamma}_n$ is asymptotically linear with influence function $\Sigma_\gamma^{-1}(\gamma_0)\pi_\gamma(\gamma_0)(\xi_i - \pi(W_i, \gamma_0))$, where

$$
\pi_\gamma(W_i, \gamma) = \frac{\partial \pi(W_i, \gamma)}{\partial \gamma}, \quad \pi_\gamma(\gamma) = \frac{\hat{\pi}(W_i, \gamma)}{\pi(W_i, \gamma)(1 - \pi(W_i, \gamma))},
$$

and

$$
\Sigma_\gamma(\gamma) = \mathbb{E} \left[ \frac{\hat{\pi}_\gamma(W_i, \gamma)}{\pi(W_i, \gamma)(1 - \pi(W_i, \gamma))} \right].
$$

That is:

$$
\sqrt{n}(\hat{\gamma}_n - \gamma_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma_\gamma^{-1}(\gamma_0)\pi_\gamma(\gamma_0)(\xi_i - \pi(W_i, \gamma_0)) + o_P(1). \quad (5.6)
$$

If the models $m(W_i, \theta)$ and $\pi(W_i, \gamma)$ are misspecified, then by White (1982), there exists $\theta^*$ and $\gamma^*$ such that $\hat{\theta}_n \xrightarrow{p} \theta^*$ and $\hat{\gamma}_n \xrightarrow{p} \gamma^*$. Moreover, the asymptotic linear expansions for $\hat{\theta}_n$ and $\hat{\gamma}_n$ are given by (3.4) and (5.6), with $\theta_0$ and $\gamma_0$ replaced by $\theta^*$ and $\gamma^*$ respectively. If the model $m(W_i, \theta)$ (respectively $\pi(W_i, \gamma)$) is correctly specified, then $\theta^* = \theta_0$ (respectively $\gamma^* = \gamma_0$).
Finally, let

\[ \Sigma_5(\beta, \theta, \gamma) = \mathbb{E} \left[ X (Y^* - e^{\beta^T X} - h_\beta(Y^*, X)) \left( 1 - \frac{\xi}{\pi(W, \gamma)} \right) m^\top(W, \theta) \right], \]

\[ \Sigma_6(\beta, \theta, \gamma) = \mathbb{E} \left[ X (Y^* - e^{\beta^T X} - h_\beta(Y^*, X)) \frac{\pi'(W, \gamma)}{\pi^2(W, \gamma)} (m(W, \theta) - \delta) \right], \]

\[ \Sigma_7(\beta, \theta, \gamma) = \Sigma_1(\beta) + (2\Sigma_3(\beta, \theta) - \Sigma_5(\beta, \theta, \gamma)) \Theta^{-1}(\theta) \Sigma_6^\top(\beta, \theta, \gamma), \]

and

\[ \Sigma_8(\beta, \theta, \gamma) = \Sigma_1(\beta) - \Sigma_6(\beta, \theta, \gamma) \Sigma_4^{-1}(\gamma) \Sigma_6^\top(\beta, \theta, \gamma), \]

We assume the following additional regularity conditions:

**C5** The parameter space for \( \gamma \) is a bounded set \( G \subset \mathbb{R}^q \) and the true parameter value \( \gamma_0 \) lies in the interior of \( G \).

**C6** The function \( \pi(w, \gamma) \) is strictly greater than 0 for all value of \( w \) in the support of \( W \) and all \( \gamma \in G \).

**C7** The function \( \pi(w, \gamma) \) is differentiable with respect to \( \gamma \), for every \( \gamma, \tilde{\gamma} \in G \),

\[ |\pi(w, \gamma) - \pi(w, \tilde{\gamma})| \leq g(w) \| \gamma - \tilde{\gamma} \| \]

for some bounded function \( g \) with \( \mathbb{E}[g(W)] = u \).

Conditions C5 and C7 for \( \gamma \) and \( \pi(\cdot, \cdot) \) are similar to conditions C2 and C4 for \( \theta \) and \( m(\cdot, \cdot) \). We are now in position to state the asymptotic properties of our AIPW estimator of \( \beta \). Proofs are given in Appendix C (consistency) and D (asymptotic normality).

**Theorem 5.1.** Assume that conditions C1-C7 hold. If either or both of the models \( m(W_i, \theta) \) and \( \pi(W_i, \gamma) \) are well specified, then \( \tilde{\beta}_n \xrightarrow{P} \beta_0 \) as \( n \to \infty \).

From this result, the proposed estimator \( \tilde{\beta}_n \) is doubly robust, in the sense that it estimates consistently \( \beta_0 \) as long as one of \( m(W_i, \theta) \) and \( \pi(W_i, \gamma) \) is correctly modeled. The next theorem describes the asymptotic distribution of \( \tilde{\beta}_n \).

**Theorem 5.2.** Assume that conditions C1-C7 hold. Then, as \( n \to \infty \), \( \sqrt{n}(\tilde{\beta}_n - \beta_0) \) converges in distribution to the Gaussian random vector \( \mathcal{N}(0, J) \), where

\[
J = \begin{cases} 
\Sigma_1^{-1}(\beta_0)\Sigma_7(\beta_0, \theta_0^*, \gamma^*)\Sigma_1^{-1}(\beta_0) & \text{if } m(W_i, \theta) \text{ is correctly specified}, \\
\Sigma_1^{-1}(\beta_0)\Sigma_8(\beta_0, \theta_0^*, \gamma_0)\Sigma_1^{-1}(\beta_0) & \text{if } \pi(W_i, \gamma) \text{ is correctly specified}, \\
\Sigma_1^{-1}(\beta_0) & \text{if both } m(W_i, \theta) \text{ and } \pi(W_i, \gamma) \text{ are correctly specified}.
\end{cases}
\]
In order to estimate the asymptotic variance of $\hat{\beta}_n$, let:

$$\Sigma_{1,n}(\beta, \theta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \left( \hat{\theta}_i(\theta, \gamma) e^{\beta^T \mathbf{X}_i} + (\hat{\delta}_i(\theta, \gamma) - 1) \left\{ Y_i^* - e^{\beta^T \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i) \right\} h_\beta(Y_i^*, \mathbf{X}_i) \right),$$

$$\Sigma_{4,n}(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\pi}^2(W_i, \gamma)}{\pi(W_i, \gamma)(1 - \pi(W_i, \gamma))},$$

$$\Sigma_{5,n}(\beta, \theta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \left( Y_i^* - e^{\beta^T \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i) \right) \left( 1 - \frac{\xi_i}{\pi(W_i, \gamma)} \right) m^T(W_i, \theta),$$

$$\Sigma_{6,n}(\beta, \theta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \left( Y_i^* - e^{\beta^T \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i) \right) \xi_i \frac{\pi^T(W_i, \gamma)}{\pi^2(W_i, \gamma)} (m(W_i, \theta) - \delta_i),$$

$$\Sigma_{7,n}(\beta, \theta, \gamma) = \tilde{\Sigma}_{1,n}(\beta, \theta, \gamma) + (2\Sigma_{3,n}(\beta, \theta) - \Sigma_{5,n}(\beta, \theta, \gamma)) \Theta_n^{-1}(\theta) \Sigma_{5,n}(\beta, \theta, \gamma),$$

and

$$\Sigma_{8,n}(\beta, \theta, \gamma) = \tilde{\Sigma}_{1,n}(\beta, \theta, \gamma) - \Sigma_{6,n}(\beta, \theta, \gamma) \tilde{\Sigma}_{4,n}^{-1}(\gamma) \Sigma_{6,n}(\beta, \theta, \gamma),$$

where $\Sigma_{3,n}$ and $\Theta_n$ are as given in Section 3. Then a consistent estimator of $\mathbf{J}$ is given by:

$$\mathbf{J}_n = \begin{cases} \tilde{\Sigma}_{1,n}^{-1}(\tilde{\beta}_n, \tilde{\theta}_n, \tilde{\gamma}_n) \Sigma_{7,n}(\tilde{\beta}_n, \tilde{\theta}_n, \tilde{\gamma}_n) \tilde{\Sigma}_{1,n}^{-1}(\tilde{\beta}_n, \tilde{\theta}_n, \tilde{\gamma}_n) & \text{if } m(W_i, \theta) \text{ is correctly specified}, \\ \tilde{\Sigma}_{1,n}^{-1}(\tilde{\beta}_n, \tilde{\theta}_n, \tilde{\gamma}_n) \Sigma_{8,n}(\tilde{\beta}_n, \tilde{\theta}_n, \tilde{\gamma}_n) \Sigma_{1,n}^{-1}(\tilde{\beta}_n, \tilde{\theta}_n, \tilde{\gamma}_n) & \text{if } \pi(W_i, \gamma) \text{ is correctly specified}, \\ \tilde{\Sigma}_{1,n}^{-1}(\tilde{\beta}_n, \tilde{\theta}_n, \tilde{\gamma}_n) & \text{if both } m(W_i, \theta) \text{ and } \pi(W_i, \gamma) \text{ are correctly specified.} \end{cases}$$

The proof is omitted.

6. A simulation study

In this section, we evaluate and compare the finite sample performance of the regression calibration (RC), multiple imputation (MI) and AIPW estimators. The simulation design is as follows. For each of $n$ individuals, the count response $Y$ is simulated from a Poisson regression model with parameter

$$\lambda = \exp(\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5),$$

where $\beta = (0.2, -0.1, 0.4, 0.3, 0.5)$, $X_2 \sim \mathcal{N}(0, 1)$, $X_3 \sim \text{Bernoulli}(0.3)$, $X_4 \sim \mathcal{N}(0, 1.5)$ and $X_5 \sim \text{uniform}[2, 5]$. The censoring and missingness mechanisms are set to be $\text{logit}(m(W, \theta)) = \theta_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4 + \theta_5 X_5 + \theta_6 Y$ and $\text{logit}(\pi(W, \gamma)) = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 Y$ respectively. In a first experiment, $\theta$ and $\gamma$ are chosen so that approximately 20% of the subjects have a censored count response and 20% have a missing censoring indicator. In a second experiment, approximately 20% of the subjects have a censored count and 40% have a missing censoring indicator. We take a sample size of $n = 500$. We compare the RC, MI and AIPW estimates under three scenario: (i) only $m(W, \theta)$ is correctly modeled, (ii) only $\pi(W, \gamma)$ is correctly modeled, (iii) both $m(W, \theta)$ and $\pi(W, \gamma)$ are correctly modeled. In the first scenario, $\pi(W, \gamma)$ is incorrectly modeled as $\text{logit}(\pi(W, \gamma)) = \gamma_1+
\( \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 Y^\ast \). In the second scenario, \( m(W, \theta) \) is incorrectly modeled as \( \logit(m(W, \theta)) = \theta_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 Y^\ast \).

Our simulation results are based on \( N = 1000 \) simulated samples. For each estimator, we report the average bias, average standard error (SE), empirical root mean square error (RMSE) and empirical coverage probability (CP) of 95%-level confidence intervals. MI estimates are obtained with \( M = 50 \) (from our numerical experiments, this is large enough to ensure stability of the estimates). To establish a benchmark for comparisons, we also include an estimator based on the full data set with no missing censoring indicators and the complete-case (CC) estimator which maximizes the log-likelihood on the subsample of complete cases only.

The results are summarized in Table 1 (experiment 1) and Table 2 (experiment 2). In the first scenario, the RC, MI and AIPW methods appear to have similar performance, with slightly smaller average SE and RMSE for the AIPW estimates. Coverage probabilities are close to the nominal confidence level, indicating that the asymptotic variances are appropriately estimated. In the second scenario, the bias of the RC and MI estimates increase substantially, resulting in coverage probabilities smaller than desired. This was expected due to misspecification of model \( m(W, \theta) \). On the other hand, the bias of the AIPW estimate stays moderate, and is similar to the first scenario, which is also expected due to the double robustness property stated in Theorem 5.1. The AIPW method generally achieves the smallest SE and RMSE but when the proportion of missing data increases, its coverage probabilities decrease around 85%-90%, which might indicate that the asymptotic variance is slightly under-estimated. Finally, when both models are correct, all three methods perform similarly (results for the RC and MI methods are the same as for the first scenario).

Overall, this simulation study confirms the theoretical results stated in the previous sections. The regression calibration, multiple imputation and robust IPW methods provide similar results when either \( m(W, \theta) \) or both \( m(W, \theta) \) and \( \pi(W, \gamma) \) are correctly specified. When \( m(W, \theta) \) is misspecified, the AIPW approach performs better than RC and MI in terms of point estimation (with substantially smaller bias for AIPW) but MI seems to provide better coverage probabilities (at the price of larger variance estimates). The CC estimates are outperformed by the three methods in all scenarios. Unreported simulation results with other sample sizes and censoring fractions provide similar observations.

7. Discussion

In this article, we have investigated several estimators of the regression parameter of the censored Poisson regression model when censoring indicators are partially missing. The regression calibration and multiple imputation estimates and their asymptotic variance estimators lead to reliable inferences when the model for the missing data given the observed variables is correctly specified, while the augmented inverse probability weighted estimator is robust against misspecification of either the model for the missing data or the missingness mechanism. Our estimators rely on parametric assumptions for the conditional models for the missing data and missingness mechanism. It is now important to assess the sensitivity of the statistical inference based on these estimates, with regard to deviations to the model assumptions. An alternative estimation strategy may rely on semiparametric or nonparametric estimation of the models for the missing data and mechanism. This stimulating topic is the subject for our next research.
Acknowledgements

Authors acknowledge financial support from the Hubert Curien "PHC-Utique" program (CMCU number : 20G1503 - Campus France number : 44172SL), implemented by Campus France.

Appendix A: Proof of Theorem 3.1

CONSISTENCY. The consistency of $\tilde{\beta}_n$ can be proved by verifying the conditions of the inverse function theorem (Foutz, 1977). We describe the main steps of the proof and omit calculation details.

Let $\hat{\ell}_n(\beta, \theta) := \partial \ell_n(\beta, \theta)/\partial \beta$. Straightforward calculations yield:

$$\hat{\ell}_n(\beta, \theta) = \sum_{i=1}^{n} X_i \left[ \hat{\delta}_i(\theta) \left( Y_i^* - e^{\beta^T X_i} \right) + (1 - \hat{\delta}_i(\theta)) h_\beta(Y_i^*, X_i) \right],$$

where $h_\beta(y, x)$ is given by (3.5). We first need to show that $\partial \hat{\ell}_n(\beta, \hat{\theta}_n)/\partial \beta^T$ exists and is continuous in a neighborhood of $\beta_0$. The map $\beta \mapsto \hat{\ell}_n(\beta, \hat{\theta}_n)$ is trivially differentiable with respect to $\beta$ and its derivative is given by:

$$\frac{\partial \hat{\ell}_n(\beta, \hat{\theta}_n)}{\partial \beta^T} = \sum_{i=1}^{n} X_i X_i^T \left( -\hat{\delta}_i(\theta_0) e^{\beta^T X_i} + (1 - \hat{\delta}_i(\hat{\theta}_n)) \left\{ Y_i^* - e^{\beta^T X_i} - h_\beta(Y_i^*, X_i) \right\} h_\beta(Y_i^*, X_i) \right),$$

which is continuous in $\beta$.

Secondly, we need to show that $n^{-1} \hat{\ell}_n(\beta_0, \hat{\theta}_n) = o_p(1)$. To see this, we decompose $n^{-1} \hat{\ell}_n(\beta_0, \hat{\theta}_n)$:

$$\frac{1}{n} \hat{\ell}_n(\beta_0, \hat{\theta}_n) = \frac{1}{n} \left( \hat{\ell}_n(\beta_0, \hat{\theta}_n) - \hat{\ell}_n(\beta_0, \theta_0) \right) + \frac{1}{n} \hat{\ell}_n(\beta_0, \theta_0).$$

By the weak law of large numbers, $n^{-1} \hat{\ell}_n(\beta_0, \theta_0)$ converges in probability to

$$\mathbb{E} \left[ X \left( \hat{\delta}(\theta_0)(Y^* - e^{\beta_0^T X}) + (1 - \hat{\delta}(\theta_0)) h_{\beta_0}(Y^*, X) \right) \right] = \mathbb{E} \left[ X \left( \mathbb{E}(\hat{\delta}(\theta_0)|W)(Y^* - e^{\beta_0^T X}) + (1 - \mathbb{E}(\hat{\delta}(\theta_0)|W)) h_{\beta_0}(Y^*, X) \right) \right], \quad (7.7)$$

where the second line follows by taking the conditional expectation given $W$. Under the missing at random assumption,

$$\mathbb{E}(\hat{\delta}(\theta_0)|W) = \mathbb{E}(\xi \delta + (1 - \xi)\mathbb{E}(\delta|W)|W) = \mathbb{E}(\xi|W)\mathbb{E}(\delta|W) + (1 - \mathbb{E}(\xi|W))\mathbb{E}(\delta|W) = \mathbb{E}(\delta|W).$$

Therefore, (7.7) is equal to

$$\mathbb{E} \left[ X \left( \mathbb{E}(\delta|W)(Y^* - e^{\beta_0^T X}) + (1 - \mathbb{E}(\delta|W)) h_{\beta_0}(Y^*, X) \right) \right] = \mathbb{E} \left[ X \left( \delta(Y^* - e^{\beta_0^T X}) + (1 - \delta) h_{\beta_0}(Y^*, X) \right) \right].$$
which is equal to 0 (this can be seen by taking successively the conditional expectations given \( \{ \delta = 1 \} \) and \( X \)). Convergence to 0 of \( n^{-1}(\hat{\ell}_n(\beta_0, \hat{\theta}_n) - \hat{\ell}_n(\beta_0, \theta_0)) \) is a consequence of the consistency of \( \hat{\theta}_n \) and of assumptions C1, C2, C4. Details are omitted.

Thirdly, we need to show that \( n^{-1}\partial \hat{\ell}_n(\beta, \hat{\theta}_n)/\partial \beta^\top \) converges in probability to a fixed matrix, uniformly in an open neighborhood of \( \beta_0 \). We have:

\[
\frac{1}{n} \frac{\partial \hat{\ell}_n(\beta, \theta)}{\partial \beta^\top} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \left( -\hat{\delta}_i(\theta) e^\beta^\top X_i + (1 - \hat{\delta}_i(\theta)) \left\{ Y_i^* - e^\beta^\top X_i - h_\beta(Y_i^*, X_i) \right\} h_\beta(Y_i^*, X_i) \right).
\]

We proceed as above and decompose \( n^{-1}\partial \hat{\ell}_n(\beta, \hat{\theta}_n)/\partial \beta^\top \) as

\[
\frac{1}{n} \frac{\partial \hat{\ell}_n(\beta, \hat{\theta}_n)}{\partial \beta^\top} = \frac{1}{n} \left( \frac{\partial \hat{\ell}_n(\beta, \hat{\theta}_n)}{\partial \beta^\top} - \frac{\partial \hat{\ell}_n(\beta, \theta_0)}{\partial \beta^\top} \right) + \frac{1}{n} \frac{\partial \hat{\ell}_n(\beta, \theta_0)}{\partial \beta^\top}.
\]

The first term converges to 0 (by the consistency of \( \hat{\theta}_n \) and assumptions C1, C2, C4) and \( n^{-1}\partial \hat{\ell}_n(\beta, \theta_0)/\partial \beta^\top \) converges in probability to \(-\Sigma_1(\beta)\) (by the weak law of large numbers). Therefore, \( n^{-1}\partial \hat{\ell}_n(\beta, \hat{\theta}_n)/\partial \beta^\top \) converges in probability to \(-\Sigma_1(\beta)\). Under conditions C1 and C2, the derivative of \( n^{-1}\partial \hat{\ell}_n(\beta, \hat{\theta}_n)/\partial \beta^\top \) with respect to \( \beta \) is bounded, for every \( n \). Hence the sequence \( (n^{-1}\partial \hat{\ell}_n(\beta, \hat{\theta}_n)/\partial \beta^\top) \) is equicontinuous. It follows from Ascoli theorem that the convergence of \( n^{-1}\partial \hat{\ell}_n(\beta, \hat{\theta}_n)/\partial \beta^\top \to -\Sigma_1(\beta) \) is uniform around \( \beta_0 \).

Having proved the conditions of the inverse function theorem, we conclude that \( \hat{\beta}_n \) converges in probability to \( \beta_0 \).

**ASYMPOTIC NORMALITY.** A Taylor’s expansion of \( \hat{\ell}_n(\hat{\beta}_n, \hat{\theta}_n) \) around \((\beta_0, \theta_0)\) yields

\[
\sqrt{n}(\hat{\beta}_n - \beta_0) = \left( -\frac{1}{n} \frac{\partial \hat{\ell}_n(\beta_0, \theta_0)}{\partial \beta^\top} \right)^{-1} \left( \frac{1}{\sqrt{n}} \hat{\ell}_n(\beta_0, \theta_0) + \frac{1}{n} \frac{\partial \hat{\ell}_n(\beta_0, \theta_0)}{\partial \theta^\top} \sqrt{n}(\hat{\theta}_n - \theta_0) \right) + o_P(1).
\]

We have:

\[
\frac{1}{n} \frac{\partial \hat{\ell}_n(\beta, \theta)}{\partial \theta^\top} = \frac{1}{n} \sum_{i=1}^{n} X_i \left( Y_i^* - e^\beta^\top X_i - h_\beta(Y_i^*, X_i) \right) (1 - \xi_i) m^\top(W_i, \theta)
\]

\[= \Sigma_2(\beta, \theta) + o_P(1).
\]

Combining this and (3.4), we can write:

\[
\sqrt{n}(\hat{\beta}_n - \beta_0) = \left( -\frac{1}{n} \frac{\partial \hat{\ell}_n(\beta_0, \theta_0)}{\partial \beta^\top} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ X_i \left( \hat{\delta}_i(\theta)(Y_i^* - e^{\beta_0^\top X_i}) + (1 - \hat{\delta}_i(\theta))h_{\beta_0}(Y_i^*, X_i) \right) \right.
\]

\[+ \Sigma_2(\beta_0, \theta_0) \Theta^{-1}(\theta_0) \tilde{m}(\theta_0) \xi_i (\hat{\delta}_i - m(W_i, \theta_0)) \]

\[+ o_P(1) \]

\[:= \left( -\frac{1}{n} \frac{\partial \hat{\ell}_n(\beta_0, \theta_0)}{\partial \beta^\top} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i + o_P(1).
\]
Now, note that
\[
\text{var} \left( X_i \left\{ \hat{\delta}_i(\theta_i)(Y_i - e^{\beta^T_0 X_i}) + (1 - \hat{\delta}_i(\theta_i))h_{\beta_0}(Y_i^*, X_i) \right\} \right) = \Sigma_1(\beta_0),
\]
and
\[
\text{var} \left( \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\tilde{m}_i(\theta_0)\xi_i(\delta - m(W_i, \theta_0)) \right) = \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)E \left[ \tilde{m}_i^{\otimes 2}(\theta_0)\xi_i(\delta - m(W_i, \theta_0))^2 \right]
\times \Theta^{-1}(\theta_0)\Sigma_2^T(\beta_0, \theta_0)
\]  
\[= \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\Sigma_2^T(\beta_0, \theta_0),
\]
since under the missing at random assumption, we have:
\[
E \left[ \tilde{m}_i^{\otimes 2}(\theta_0)\xi_i(\delta - m(W_i, \theta_0))^2 \right] = E \left[ \frac{\tilde{m}_i^{\otimes 2}(W_i, \theta_0)}{(m(W_i, \theta_0)(1 - m(W_i, \theta_0)))^2} E \left[ \xi_i(\delta - m(W_i, \theta_0))^2 \right] \right]
\times E \left[ \tilde{m}_i^{\otimes 2}(W_i, \theta_0)(1 - m(W_i, \theta_0)) \pi(W_i) \right]
\]
or
\[
\Theta(\theta_0).
\]
We consider now the covariance structure of \( \mathcal{U}_i \). We have
\[
\text{cov} \left( X_i \left\{ \hat{\delta}_i(\theta_i)(Y_i^* - e^{\beta^T_0 X_i}) + (1 - \hat{\delta}_i(\theta_i))h_{\beta_0}(Y_i^*, X_i) \right\}, \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\tilde{m}_i(\theta_0)\xi_i(\delta - m(W_i, \theta_0)) \right) = \E \left[ X_i \tilde{m}_i^T(\theta_0)E \left[ \left( \hat{\delta}_i(\theta_i)(Y_i^* - e^{\beta^T_0 X_i}) - h_{\beta_0}(Y_i^*, X_i) \right) + h_{\beta_0}(Y_i^*, X_i) \right] \xi_i(\delta - m(W_i, \theta_0)) \pi(W_i) \right]
\times \Theta^{-1}(\theta_0)\Sigma_2^T(\beta_0, \theta_0),
\]
and
\[
E \left[ \frac{\hat{\delta}_i(\theta_i)(Y_i^* - e^{\beta^T_0 X_i}) - h_{\beta_0}(Y_i^*, X_i)}{(1 - \hat{\delta}_i(\theta_i)(Y_i^*, X_i)) \xi_i(\delta - m(W_i, \theta_0)) \pi(W_i) \right]
= \E \left[ \xi_i(1 - m(W_i, \theta_0))(Y_i^* - e^{\beta^T_0 X_i} - h_{\beta_0}(Y_i^*, X_i)) + \xi_i h_{\beta_0}(Y_i^*, X_i)(\delta - m(W_i, \theta_0)) \pi(W_i) \right]
\]
\[
= (1 - m(W_i, \theta_0))(Y_i^* - e^{\beta^T_0 X_i} - h_{\beta_0}(Y_i^*, X_i))m(W_i, \theta_0)\pi(W_i),
\]
therefore,
\[
\text{cov} \left( X_i \left\{ \hat{\delta}_i(\theta_i)(Y_i^* - e^{\beta^T_0 X_i}) + (1 - \hat{\delta}_i(\theta_i))h_{\beta_0}(Y_i^*, X_i) \right\}, \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\tilde{m}_i(\theta_0)\xi_i(\delta - m(W_i, \theta_0)) \right) = \E \left[ X_i \tilde{m}_i^T(\theta_0)(1 - m(W_i, \theta_0))(Y_i^* - e^{\beta^T_0 X_i} - h_{\beta_0}(Y_i^*, X_i))m(W_i, \theta_0)\pi(W_i) \right] \Theta^{-1}(\theta_0)\Sigma_2^T(\beta_0, \theta_0)
\]
\[
= \E \left[ X_i \tilde{m}_i^T(\theta_0)(Y_i^* - e^{\beta^T_0 X_i} - h_{\beta_0}(Y_i^*, X_i))\pi(W_i) \right] \Theta^{-1}(\theta_0)\Sigma_2^T(\beta_0, \theta_0).
\]
It follows that
\[
\text{var}(\mathcal{U}_i) = \Sigma_1(\beta_0) + (2 \Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)) \Theta^{-1}(\theta_0)\Sigma_2^T(\beta_0, \theta_0).
\]
Finally, Theorem 3.1 follows from the multivariate central limit theorem and Slutsky’s theorem. \( \square \)
Appendix B: Proof of Theorem 4.1

Consistency can be proved in much the same way as \( \tilde{\beta}_n \); the proof is therefore omitted. We turn to asymptotic normality. A technical lemma is needed. For the class of functions \( G \) where \( \delta_f \),

**Proof of Lemma 1**

In this proof, for notational simplicity, we will write \( f_\theta \) instead of \( f_{\beta_0,\theta,j} \).

First, note that

\[
\frac{1}{\sqrt{n}} \left[ \hat{\ell}_{n,j}(\beta_0, \hat{\theta}_n) - n\mathbb{E}[^{\hat{\ell}}_{n,j}(\beta_0, \hat{\theta}_n)] - \left( \hat{\ell}_{n,j}(\beta_0, \theta_0) - n\mathbb{E}[^{\hat{\ell}}_{n,j}(\beta_0, \theta_0)] \right) \right] \xrightarrow{P} 0
\]

as \( n \to \infty \).

**Proof of Lemma 1**

In this proof, for notational simplicity, we will write \( f_\theta \) instead of \( f_{\beta_0,\theta,j} \).

First, note that

\[
\frac{1}{\sqrt{n}} \left[ \hat{\ell}_{n,j}(\beta_0, \theta) - n\mathbb{E}[^{\hat{\ell}}_{n,j}(\beta_0, \theta)] \right] = \frac{1}{\sqrt{n}} \left[ \sum_{i=1}^{n} f_{\theta}(O_i) - n\mathbb{E}[f_{\theta}(O_1)] \right]
\]

where \( \mathbb{G}_n f_\theta \) denotes the empirical process evaluated at \( f_\theta \). To prove the lemma, we first prove that the class of functions \( \{f_\theta : \theta \in \Theta\} \) is Donsker (see, for example, van der Vaart (2000) for a detailed account on empirical processes and Donsker classes). For that purpose, we decompose \( f_\theta \) in (7.8) as

\[
f_{\theta}(O_i) = X_i(f_{1,\theta}(O_i) + f_{2,\theta}(O_i) + f_{3,\theta}(O_i)),
\]

where \( f_{1,\theta}(O_i) = -\delta_{n,j}^*(\theta)e^{\beta_0}\mathbf{X}_i + h_{\beta_0}(Y_i^*, \mathbf{X}_i), \) \( f_{2,\theta}(O_i) = \delta_{n,j}^*(\theta)Y_i^* \) and \( f_{3,\theta}(O_i) = -\delta_{n,j}^*(\theta)h_{\beta_0}(Y_i^*, \mathbf{X}_i) \) and we show that the classes \( \mathcal{F}_1 := \{f_{1,\theta} : \theta \in \Theta\}, \mathcal{F}_2 := \{f_{2,\theta} : \theta \in \Theta\} \) and \( \mathcal{F}_3 := \{f_{3,\theta} : \theta \in \Theta\} \) are Donsker.

For illustration purpose, we show that \( \mathcal{F}_1 \) is Donsker. Here, it is useful to see that \( D_{n,j}(\theta) \sim \mathcal{B}(m(\mathbf{W}, \theta)) \) as the random variable \( 1_{U_i \leq m(\mathbf{W}, \theta)} \), where \( U_i \) is a uniform random variable on \([0,1]\), independent of \( O_i \).

Let \( d := \text{diam}(\Theta) \) denote the diameter of \( \Theta \subset \mathbb{R}^q \). Then the size of \( \Theta \) in every direction is at most \( d \) and thus, we can cover \( \Theta \) with fewer than \( (d/\kappa)^q \) cubes of length \( \kappa \). The circumscribed balls have radius a multiple \( \kappa^* := \alpha \kappa \) of \( \kappa \) \((\alpha > 0)\) and these balls also cover \( \Theta \). Now, for a given \( \theta \in \Theta \), consider the set

\[
\{f_{1,\tilde{\theta}} : \tilde{\theta} \in \Theta \cap \mathcal{B}(\theta, \kappa^*)\},
\]

where \( \mathcal{B}(\theta, \kappa^*) = \{\tilde{\theta} \in \mathbb{R}^q : \|\theta - \tilde{\theta}\| \leq \kappa^*\} \) is the ball of radius \( \kappa^* \) and center \( \theta \). If \( \tilde{\theta} \in \mathcal{B}(\theta, \kappa^*) \), condition C4 implies that

\[
|m(\mathbf{w}, \theta) - m(\mathbf{w}, \tilde{\theta})| \leq h(\mathbf{w})\kappa^*.
\]
hence \( m(w, \theta) - h(w) \kappa^* \leq m(w, \hat{\theta}) \leq m(w, \theta) + h(w) \kappa^* \) and thus \( 1_{\{U_i \leq m(w, \theta) - h(w) \kappa^*\}} \leq 1_{\{U_i \leq m(w, \hat{\theta})\}} \leq 1_{\{U_i \leq m(w, \theta) + h(w) \kappa^*\}} \). From this, we can see that

\[
 f^L_{\hat{\theta}}(O_i) \leq f^L_{1, \hat{\theta}}(O_i) \leq f^L_{\hat{\theta}}(O_i),
\]

where

\[
 f^L_{\hat{\theta}}(O_i) = h_{\beta_0}(Y^*_i, X_i) - (\xi_i \delta_i + (1 - \xi_i)1_{\{U_i \leq m(w, \theta) - h(w) \kappa^*\}})e^{\beta_0^T X_i},
\]

\[
 f^L_{\hat{\theta}}(O_i) = h_{\beta_0}(Y^*_i, X_i) - (\xi_i \delta_i + (1 - \xi_i)1_{\{U_i \leq m(w, \theta) - h(w) \kappa^*\}})e^{\beta_0^T X_i}.
\]

Moreover, under conditions C1, C2 and C4, there exists a finite positive constant \( c_1 \) such that

\[
 E \left[ (f^L_{\hat{\theta}}(O_i) - f^L_{1, \hat{\theta}}(O_i))^2 \right] \leq 2c_1 \kappa^* v.
\]

Therefore, \([f^L_{\hat{\theta}}, f^U_{\hat{\theta}}]\) is an \( \varepsilon \)-bracket for \( \{f_{1, \hat{\theta}} : \hat{\theta} \in \Theta \cap B(\theta, \kappa^*)\} \), with \( \varepsilon^2 = 2c_1 \kappa^* v \). Since we can cover \( \Theta \) with fewer than \( (d/\kappa)^q \) balls of radius \( \kappa^* \), we can cover \( \mathcal{F}_1 = \{f_{1, \hat{\theta}} : \hat{\theta} \in \Theta\} \) with fewer than \( (d/\kappa)^q \) \( \varepsilon \)-brackets \([f^L_{\hat{\theta}}, f^U_{\hat{\theta}}]\), with \( \varepsilon = \sqrt{2c_1 \kappa^* v} \). The number of such \( \varepsilon \)-brackets is thus bounded by \((ad/\kappa)^q = (2a_1 c_1 d v/\varepsilon^2)^q\), which is order \( \varepsilon^{-2q} \). Hence, the bracketing integral is of order \( \int_0^1 \sqrt{-2q \log \varepsilon} \, d\varepsilon \), which is finite. Therefore, the class of functions \( \mathcal{F}_1 \) is Donsker, by Theorem 19.5 of [van der Vaart (2000)].

By using similar arguments, we can prove that \( \mathcal{F}_2 \) and \( \mathcal{F}_3 \) are also Donsker classes. It follows that the class of functions \( \{f_{1, \theta} + f_{2, \theta} + f_{3, \theta} : \theta \in \Theta\} \) is Donsker (sums of Donsker classes are Donsker). Finally, \( \mathbf{X} \) is bounded (by condition C1), thus the class of functions \( \{f_{\theta} : \theta \in \Theta\} \) is Donsker.

It follows that the sequence of processes \( \{G_n f_{\theta} : \theta \in \Theta\} \) converges in distribution to a tight limit process, and as such, is stochastically equicontinuous. Thus, Lemma 14.3 of [Tsiatis (2007)] and the consistency of \( \hat{\theta}_n \) imply that \( G_n f_{\hat{\theta}_n} - G_n f_{\theta_0} \xrightarrow{P} 0 \), which is exactly (7.9). This concludes the proof.

\[
0 = \frac{1}{\sqrt{n}} \hat{\ell}^*_{n,j}(\hat{\beta}^*_{n,j}, \hat{\theta}_n) = \frac{1}{\sqrt{n}} \hat{\ell}^*_{n,j}(\beta_0, \hat{\theta}_n) + \frac{1}{n} \frac{\partial \hat{\ell}^*_{n,j}(\beta_0, \hat{\theta}_n)}{\partial \beta^\top} \sqrt{n}(\hat{\beta}^*_{n,j} - \beta_0) + o_P(1),
\]

Then, using Lemma [1] we obtain:

\[
0 = \frac{1}{\sqrt{n}} \hat{\ell}^*_{n,j}(\beta_0, \theta_0) - \sqrt{n}E[\hat{\ell}^*_{1,j}(\beta_0, \theta_0)] + \frac{1}{n} \frac{\partial \hat{\ell}^*_{n,j}(\beta_0, \theta_0)}{\partial \beta^\top} \sqrt{n}(\hat{\beta}^*_{n,j} - \beta_0) + o_P(1)
\]

\[
= \frac{1}{\sqrt{n}} \hat{\ell}^*_{n,j}(\beta_0, \theta_0) + \sqrt{n} \left( \frac{\partial E[\hat{\ell}^*_{1,j}(\beta_0, \theta_0)]}{\partial \theta^\top} (\theta_n - \theta_0) + o_P(||\hat{\theta}_n - \theta_0||) \right) + \frac{1}{n} \frac{\partial \hat{\ell}^*_{n,j}(\beta_0, \theta_0)}{\partial \beta^\top} \sqrt{n}(\hat{\beta}^*_{n,j} - \beta_0) + o_P(1),
\]

where the second line follows from a Taylor expansion of \( E[\hat{\ell}^*_{1,j}(\beta_0, \theta_0)] \) around \( \theta_0 \). Two technical lemmas are now needed.
Lemma 2. For \( j = 1, \ldots, M \), we have
\[
\frac{\partial \mathbb{E}[\dot{\ell}^*_1(j, \theta)]}{\partial \theta} = \Sigma_2(\beta, \theta).
\]

Proof of Lemma 2. First, we note that
\[
\mathbb{E}[\delta^*_1(j) | W_1] = \mathbb{E}[\xi_1 \delta_1 + (1 - \xi_1) D_{1,j}(\theta)] = \pi(W_1)m(W_1, \theta_0) + (1 - \pi(W_1))m(W_1, \theta). \tag{7.11}
\]
Hence, using (7.8) and iterating the expectation with conditioning on \( W_1 \), we obtain:
\[
\mathbb{E}[\dot{\ell}^*_1(j, \theta)] = \mathbb{E}\left[ X_1 \left( (\pi(W_1)m(W_1, \theta_0) + (1 - \pi(W_1))m(W_1, \theta)) \left[ Y_1^* - e^{\beta_0^T X_1} - h_\beta(Y_1^*, X_1) \right] + h_\beta(Y_1^*, X_1) \right) \right].
\]
Finally, straightforward calculations yield
\[
\frac{\partial \mathbb{E}[\dot{\ell}^*_1(j, \theta)]}{\partial \theta} = \mathbb{E}\left[ X_1 (1 - \pi(W_1))m^\top(W_1, \theta) \left( Y_1^* - e^{\beta_0^T X_1} - h_\beta(Y_1^*, X_1) \right) \right] = \Sigma_2(\beta, \theta).
\]

Lemma 3. For \( j = 1, \ldots, M \),
\[
\frac{1}{n} \frac{\partial \dot{\ell}^*_n(j, \theta_n)}{\partial \beta} \xrightarrow{p} -\Sigma_1(\beta_0).
\]

Proof of Lemma 3. Let \( j = 1, \ldots, M \). Straightforward calculations yield:
\[
\frac{\partial \dot{\ell}^*_n(j, \theta_n)}{\partial \beta} = \sum_{i=1}^n X_i X_i^\top \left[ -\delta^*_{i,j}(\theta_n)e^{\beta_0^T X_i} + (1 - \delta^*_{i,j}(\theta_n))h_{\beta_0}(Y_i^*, X_i) \left( Y_i^* - e^{\beta_0^T X_i} - h_{\beta_0}(Y_i^*, X_i) \right) \right].
\]
Then we decompose \( n^{-1} \partial \hat{\beta}_{n,j}(\beta_0, \hat{\theta}_n) / \partial \beta \) as:

\[
\frac{1}{n} \frac{\partial \hat{\beta}_{n,j}(\beta_0, \hat{\theta}_n)}{\partial \beta} = \left( \frac{1}{n} \frac{\partial \hat{\beta}_{n,j}(\beta_0, \hat{\theta}_n)}{\partial \beta} - \frac{1}{n} \frac{\partial \hat{\beta}_{n,j}(\beta_0, \theta_0)}{\partial \beta} \right) + \frac{1}{n} \frac{\partial \hat{\beta}_{n,j}(\beta_0, \theta_0)}{\partial \beta} \\
= \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \left[ -e_{\beta_0}^\top X_i (1 - \xi_i)(D_{i,j}(\hat{\theta}_n) - D_{i,j}(\theta_0)) \\
+ h_{\beta_0}(Y_i^*, X_i) \left( Y_i^* - e_{\beta_0}^\top X_i - h_{\beta_0}(Y_i^*, X_i) \right) (1 - \xi_i)(D_{i,j}(\theta_0) - D_{i,j}(\hat{\theta}_n)) \right] \\
+ \frac{1}{n} \frac{\partial \hat{\beta}_{n,j}(\beta_0, \theta_0)}{\partial \beta} \\
= \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top (1 - \xi_i) \left[ e_{\beta_0}^\top X_i + h_{\beta_0}(Y_i^*, X_i) \left( Y_i^* - e_{\beta_0}^\top X_i - h_{\beta_0}(Y_i^*, X_i) \right) \right] \\
\times (D_{i,j}(\theta_0) - D_{i,j}(\hat{\theta}_n)) + \frac{1}{n} \frac{\partial \hat{\beta}_{n,j}(\beta_0, \theta_0)}{\partial \beta}, \tag{7.12}
\]

where \( Z_i := X_i X_i^\top (1 - \xi_i) [e_{\beta_0}^\top X_i + h_{\beta_0}(Y_i^*, X_i)(Y_i^* - e_{\beta_0}^\top X_i - h_{\beta_0}(Y_i^*, X_i))] \) (in what follows, we will denote by \( Z_{i,(\ell,k)} \) the \((\ell, k)\)-th element of \( Z_i \)) and \( U_{i,j} \) is a uniform random variable on \([0, 1]\), independent of all other random variables.

Consider the first term in the right-hand side of (7.12). The random variable \( |1_{\{U_{i,j} \leq m(W_i, \theta_0)\}} - 1_{\{U_{i,j} \leq m(W_i, \hat{\theta}_n)\}}| \) is equal to 0 or 1 and takes the value 1 with probability \( |m(W_i, \theta_0) - m(W_i, \hat{\theta}_n)| \). Let \( \varepsilon > 0 \). Then, for \( \ell, k \in \{1, \ldots, p\} \), Markov’s inequality implies that

\[
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} Z_{i,(\ell,k)} \left( 1_{\{U_{i,j} \leq m(W_i, \theta_0)\}} - 1_{\{U_{i,j} \leq m(W_i, \hat{\theta}_n)\}} \right) \right| > \varepsilon \right) \\
\leq \frac{1}{\varepsilon} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} Z_{i,(\ell,k)} \left( 1_{\{U_{i,j} \leq m(W_i, \theta_0)\}} - 1_{\{U_{i,j} \leq m(W_i, \hat{\theta}_n)\}} \right) \right].
\]

Under conditions C1 and C2, there exists a finite positive constant \( c_2 \) such that \( |Z_{i,(\ell,k)}| \leq c_2 \). Thus,

\[
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} Z_{i,(\ell,k)} \left( 1_{\{U_{i,j} \leq m(W_i, \theta_0)\}} - 1_{\{U_{i,j} \leq m(W_i, \hat{\theta}_n)\}} \right) \right| > \varepsilon \right) \\
\leq \frac{c_2}{\varepsilon n} \sum_{i=1}^{n} |m(W_i, \theta_0) - m(W_i, \hat{\theta}_n)| \\
\leq \frac{c_2}{\varepsilon} \sum_{i=1}^{n} h(W_i) \| \theta_0 - \hat{\theta}_n \| \\
\leq \frac{c_2}{\varepsilon} \| \theta_0 - \hat{\theta}_n \| (v + o_P(1)),
\]

where the last two lines follow from the condition C4. Finally, consistency of \( \hat{\theta}_n \) implies that
\[ \frac{1}{n} \sum_{i=1}^{n} Z_{i,(t,k)} \left( I\{U_{i,j} \leq m(w_i, \theta_0)\} - I\{U_{i,j} \leq m(w_i, \tilde{\theta}_0)\} \right) \] converges in probability to 0, and the first term in the right-hand side of (7.12) also converges to 0.

We consider now the second term in the right-hand side of (7.12). By the weak law of large numbers, \( n^{-1} \partial \hat{\ell}_{n,j}^* (\beta_0, \theta_n) / \partial \beta^\top \) converges in probability to

\[
\mathbb{E} \left[ X_1 X_1^\top \left[ -\delta_{1,j}^* (\theta_0) e^\beta_i X_1 + \left( 1 - \delta_{1,j}^* (\theta_0) \right) h_{\beta_0} (Y_i^\ast, X_i) \left( Y_i^\ast - e^{\beta_i} X_1 - h_{\beta_0} (Y_i^\ast, X_i) \right) \right] \right].
\]

(7.13)

Using the fact that \( \mathbb{E} [\delta_{1,j}^* (\theta_0) | W_1] = m(W_1, \theta_0) \) (see (7.11)), and iterating the expectation in (7.13) with conditioning on \( W_1 \), we easily show that (7.13) is equal to \(-\Sigma_1 (\beta_0)\).

Thus, we have shown that \( n^{-1} \partial \hat{\ell}_{n,j}^* (\beta_0, \theta_n) / \partial \beta^\top \) converges in probability to \(-\Sigma_1 (\beta_0)\), which concludes the proof.

\[ \square \]

By combining (7.10) with Lemmas 2 and 3 we obtain the following approximation of \( \hat{\beta}_{n,j}^* \):

\[
\sqrt{n} (\hat{\beta}_{n,j}^* - \beta_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^{M} \left( \sqrt{n} (\hat{\beta}_{n,j}^* - \beta_0) \right)
\]

\[ = \Sigma_1^{-1} (\beta_0) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{1}{M} \sum_{j=1}^{M} X_i \left\{ \delta_{i,j}^* (\theta_0) [Y_i^\ast - e^{\beta_0^\top} X_i - h_{\beta_0} (Y_i^\ast, X_i)] + h_{\beta_0} (Y_i^\ast, X_i) \right\} \left[ \delta_{i,j}^* (\theta_0) \right] \right. \]

\[ + \left. \Sigma_2 (\beta_0, \theta_0) \Theta^{-1} (\theta_0) \tilde{m}_i (\theta_0) \xi_i (\delta_i - m(W_i, \theta_0)) \right\} + o_P(1), \]

(7.14)

which in turn implies the approximation of the multiple imputation estimator \( \hat{\beta}_n^* \):

\[
\sqrt{n} (\hat{\beta}_n^* - \beta_0) = \frac{1}{M} \sum_{j=1}^{M} \left( \sqrt{n} (\hat{\beta}_{n,j}^* - \beta_0) \right)
\]

\[ = \Sigma_1^{-1} (\beta_0) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{1}{M} \sum_{j=1}^{M} f_{\beta_0, \theta_0,j} (O_i) + V_i \right\} + o_P(1), \]

where \( V_i := \Sigma_2 (\beta_0, \theta_0) \Theta^{-1} (\theta_0) \tilde{m}_i (\theta_0) \xi_i (\delta_i - m(W_i, \theta_0)) \). We have already shown (see proof of Theorem 3.1) that

\[
\text{var}(V_i) = \Sigma_2 (\beta_0, \theta_0) \Theta^{-1} (\theta_0) \Sigma_2^\top (\beta_0, \theta_0).
\]

Similar calculations as in the proof of Theorem 3.1 yield:

\[
\text{cov}(f_{\beta_0, \theta_0,j} (O_i), V_i) = (\Sigma_3 (\beta_0, \theta_0) - \Sigma_2 (\beta_0, \theta_0)) \Theta^{-1} (\theta_0) \Sigma_2^\top (\beta_0, \theta_0).
\]

Therefore,

\[
\text{var} \left( \frac{1}{M} \sum_{j=1}^{M} f_{\beta_0, \theta_0,j} (O_i) + V_i \right) = \text{var} \left( \frac{1}{M} \sum_{j=1}^{M} f_{\beta_0, \theta_0,j} (O_i) \right) + \text{var}(V_i) + \frac{2}{M} \sum_{j=1}^{M} \text{cov}(f_{\beta_0, \theta_0,j} (O_i), V_i)
\]

\[ = \Sigma_1^* (\beta_0, \theta_0) + 2 \Sigma_2 (\beta_0, \theta_0) \Theta^{-1} (\theta_0) \Sigma_2^\top (\beta_0, \theta_0)
\]

\[ + 2 (\Sigma_3 (\beta_0, \theta_0) - \Sigma_2 (\beta_0, \theta_0)) \Theta^{-1} (\theta_0) \Sigma_2^\top (\beta_0, \theta_0)
\]

\[ = \Sigma_1^* (\beta_0, \theta_0) + (2 \Sigma_3 (\beta_0, \theta_0) - \Sigma_2 (\beta_0, \theta_0)) \Theta^{-1} (\theta_0) \Sigma_2^\top (\beta_0, \theta_0).
\]

(7.15)
Finally, it follows from (7.14), (7.15) and the multivariate central limit theorem that \( \sqrt{n}(\hat{\beta}_n^* - \beta_0) \)
converges in distribution to a Gaussian vector with mean zero and variance
\[
\Sigma^{-1}_1(\beta_0) \left\{ \Sigma^1_1(\beta_0, \theta_0) + (2\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)) \Theta^{-1}(\theta_0) \Sigma_2^\top(\beta_0, \theta_0) \right\} \Sigma^{-1}_1(\beta_0),
\]
which concludes the proof. \( \square \)

Appendix C: Proof of Theorem 5.1

Assume that the model \( m(W_i, \theta) \) is correctly specified. It is straightforward to check that the
map \( \beta \rightarrow \partial \ell_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)/\partial \beta^\top \) exists and is continuous in a neighborhood of \( \beta_0 \) (condition i).

Now, we show that \( n^{-1} \hat{\ell}_n(\beta_0, \hat{\theta}_n, \hat{\gamma}_n) = o_p(1) \) (condition ii). To see this, decompose \( n^{-1} \hat{\ell}_n(\beta_0, \hat{\theta}_n, \hat{\gamma}_n) \) as:
\[
\frac{1}{n} \hat{\ell}_n(\beta_0, \hat{\theta}_n, \hat{\gamma}_n) = \frac{1}{n} \sum_{i=1}^n x_i \left( \frac{\xi_i}{\pi(W_i, \hat{\gamma}_n)} (\delta_i - m(W_i, \hat{\theta}_n)) \left( Y_i^* - e^{\beta_0^\top X_i} \right) - h_{\beta_0}(Y_i^*, X_i) \right)

+ \frac{1}{n} \sum_{i=1}^n x_i \{ m(W_i, \hat{\theta}_n) \left( Y_i^* - e^{\beta_0^\top X_i} \right) + (1 - m(W_i, \hat{\theta}_n)) h_{\beta_0}(Y_i^*, X_i) \},
\]
\[
:= Q_n(1)(\hat{\theta}_n, \hat{\gamma}_n) + Q_n(2)(\hat{\theta}_n).
\]

First, we consider the term \( Q_n(1)(\hat{\theta}_n, \hat{\gamma}_n) \). Let \( Q_i \equiv x_i \xi_i \left( Y_i^* - e^{\beta_0^\top X_i} \right) - h_{\beta_0}(Y_i^*, X_i) \). We have:
\[
Q_n(1)(\hat{\theta}_n, \hat{\gamma}_n) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\pi(W_i, \hat{\gamma}_n)} (\delta_i - m(W_i, \hat{\theta}_n)) Q_i

= \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{\pi(W_i, \hat{\gamma}_n)} - \frac{1}{\pi(W_i, \gamma^*)} \right) (\delta_i - m(W_i, \hat{\theta}_n)) Q_i

+ \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{\pi(W_i, \gamma^*)} \right) (\delta_i - m(W_i, \theta_0)) Q_i

:= Q_{n,1}^{(1)} + Q_{n,2}^{(1)} + Q_{n,3}^{(1)}.
\]

Now, letting \( Q_{n,1,\ell}^{(1)} \) and \( Q_{i,\ell} \) denote the \( \ell \)-th component of the vectors \( Q_{n,1}^{(1)} \) and \( Q_i \), respectively (for \( \ell = 1, \ldots, p \)), we have:
\[
|Q_{n,1,\ell}^{(1)}| \leq \frac{1}{n} \sum_{i=1}^n \left| \frac{\pi(W_i, \gamma^*) - \pi(W_i, \hat{\gamma}_n)}{\pi(W_i, \gamma^*) \pi(W_i, \hat{\gamma}_n)} \right| |\delta_i - m(W_i, \hat{\theta}_n)||Q_{i,\ell}|.
\]

Conditions C1 and C6 ensure that there exists a finite positive constant \( c_3 \) such that
\[
|Q_{n,1,\ell}| \leq \frac{c_3}{n} \sum_{i=1}^n |\pi(W_i, \gamma^*) - \pi(W_i, \hat{\gamma}_n)|,
\]

20
and the condition C7 implies that
\[ |Q_{n,1,\ell}| \leq c_3 \frac{1}{n} \sum_{i=1}^{n} g(W_i) \left\| \gamma^* - \hat{\gamma}_n \right\|, \]
\[ \leq c_3 (u + o_{\mathbb{P}}(1)) \left\| \gamma^* - \hat{\gamma}_n \right\|. \]

Finally, the convergence of \( \hat{\gamma}_n \) to \( \gamma^* \) implies that \( Q_{n,1,\ell}^{(1)} (\ell = 1, \ldots, p) \), and thus \( Q_{n,1}^{(1)} \), converge to 0 as \( n \to \infty \). Similarly, under conditions C1 and C6, there exists a finite positive constant \( c_4 \) such that
\[ |Q_{n,3,\ell}| \leq \frac{c_4}{n} \sum_{i=1}^{n} m(W_i, \theta_0) - m(W_i, \hat{\theta}_n), \quad \ell = 1, \ldots, p, \]
and condition C4 implies
\[ |Q_{n,3,\ell}| \leq c_4 (v + o_{\mathbb{P}}(1)) \left\| \theta_0 - \hat{\theta}_n \right\|. \]

If the model \( m(W_i, \theta) \) is correctly specified (that is, if \( \hat{\theta}_n \) is consistent for \( \theta_0 \)), \( Q_{n,3,\ell}^{(1)} (\ell = 1, \ldots, p) \), and thus \( Q_{n,3}^{(1)} \), converge to 0 as \( n \to \infty \). Finally, by the law of large numbers, \( Q_{n,2}^{(1)} \) converges in probability to
\[ \mathbb{E} \left[ \frac{1}{\pi(W_i, \gamma^*)} (\delta_i - m(W_i, \theta_0)) Q_{1,\ell} \right] = \mathbb{E} \left[ \frac{X_i(Y_i^* - e^{\beta_0^T} X_i - h_{\beta_0}(Y_i^*, X_i))}{\pi(W_i, \gamma^*)} \mathbb{E}(\xi_i|W_i) (\mathbb{E}(\delta_i|W_i) - m(W_i, \theta_0)) \right], \]
which equals 0 if the model \( m(W_i, \theta) \) is correctly specified (in this case, \( m(W_i, \theta_0) = \mathbb{E}(\delta_i|W_i) \)). It follows that \( Q_{n,1}^{(1)} (\hat{\theta}_n, \hat{\gamma}_n) = o_{\mathbb{P}}(1) \). Therefore,
\[ \frac{1}{n} \hat{\ell}_n(\beta_0, \hat{\theta}_n, \hat{\gamma}_n) = Q_{n}^{(2)}(\hat{\theta}_n) + o_{\mathbb{P}}(1). \]

With obvious notations, we have \( Q_{n}^{(2)}(\hat{\theta}_n) = Q_{n}^{(2)}(\hat{\theta}_n) - Q_{n}^{(2)}(\theta_0) + Q_{n}^{(2)}(\theta_0) \). By the law of large numbers, \( Q_{n}^{(2)}(\theta_0) \) converges in probability to \( \mathbb{E}(X \{ m(W, \theta_0)(Y^* - e^{\beta_0^T} X) + (1 - m(W, \theta_0)) h_{\beta_0}(Y^*, X) \}) \), which is equal to 0 (see proof of Theorem 3.1). We also have
\[ Q_{n}^{(2)}(\hat{\theta}_n) - Q_{n}^{(2)}(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} X_i(m(W_i, \hat{\theta}_n) - m(W_i, \theta_0))(Y_i^* - e^{\beta_0^T} X_i - h_{\beta_0}(Y_i^*, X_i)), \]
and using similar arguments as for \( Q_{n,3}^{(1)} \), we can show that this converges to 0 if model \( m(W_i, \theta) \) is correctly specified. Finally, \( Q_{n}^{(2)}(\hat{\theta}_n) = o_{\mathbb{P}}(1) \), which concludes the proof of condition \( i \).

Now, we prove that \( n^{-1} \partial \ell_n(\beta, \hat{\theta}_n, \hat{\gamma}_n) / \partial \beta^T \) converges to \( -\Sigma_1(\beta) \), uniformly in a neighborhood of \( \beta_0 \) (condition \( iii \)). Letting \( Q_{i,\beta} = (Y_i^* - e^{\beta^T} X_i - h_{\beta}(Y_i^*, X_i)) h_{\beta}(Y_i^*, X_i) \), some easy calculations yield:
\[ \frac{1}{n} \partial \ell_n(\beta, \theta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \left[ -\delta_i(\theta, \gamma)(e^{\beta^T} X_i + Q_{i,\beta}) + Q_{i,\beta} \right], \]
\[ = -\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T (e^{\beta^T} X_i + Q_{i,\beta}) m(W_i, \theta) + \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T Q_{i,\beta} \]
\[ + \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \frac{\xi_i}{\pi(W_i, \gamma)} (m(W_i, \theta) - \delta_i)(e^{\beta^T} X_i + Q_{i,\beta}). \]
Now, decompose \( n^{-1} \partial \ell_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)/\partial \beta^T \) as
\[
\frac{1}{n} \frac{\partial \ell_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)}{\partial \beta^T} = \frac{1}{n} \frac{\partial \ell_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)}{\partial \beta^T} - \frac{1}{n} \frac{\partial \ell_n(\beta, \theta_0, \gamma^*)}{\partial \beta^T} + \frac{1}{n} \frac{\partial \ell_n(\beta, \theta_0, \gamma^*)}{\partial \beta^T},
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T (e^{ \beta^T X_i} + Q_{i,\beta})(m(W_i, \theta_0) - m(W_i, \hat{\theta}_n)) + \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \xi_i (e^{ \beta^T X_i} + Q_{i,\beta}) \frac{m(W_i, \theta_0) - \delta_i}{\pi(W_i, \hat{\gamma}_n)} - \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \xi_i (e^{ \beta^T X_i} + Q_{i,\beta}) \frac{m(W_i, \theta_0) - \delta_i}{\pi(W_i, \gamma^*)} + \frac{1}{n} \frac{\partial \ell_n(\beta, \theta_0, \gamma^*)}{\partial \beta^T},
\]
\[
= T_n^{(1)} + T_n^{(2)} + T_n^{(3)} + \frac{1}{n} \frac{\partial \ell_n(\beta, \theta_0, \gamma^*)}{\partial \beta^T}.
\]

Using similar arguments as for \( Q_{n,3}^{(1)} \) (respectively \( Q_n^{(1)} \) and \( Q_{n,2}^{(1)} \), we can show that \( T_n^{(1)} \) (respectively \( T_n^{(2)} \) and \( T_n^{(3)} \)) converge to 0 as \( n \to \infty \). Details are omitted. Now, \( n^{-1} \partial \ell_n(\beta, \theta_0, \gamma^*)/\partial \beta^T \) converges in probability to \( E[X_i X_i^T (\hat{\delta}_i(\theta_0, \gamma^*)(e^{ \beta^T X_i} + Q_{i,\beta}) + Q_{i,\beta})] \). If the model \( m(W_i, \theta) \) is correctly specified (that is, \( m(W_i, \theta_0) = E(\delta_i|W_i) \)), we have
\[
E[\hat{\delta}_i(\theta_0, \gamma^*)|W_i] = \frac{E[\xi_i W_i]E[\delta_i|W_i]}{\pi(W_i, \gamma^*)} + \left( 1 - \frac{E[\xi_i W_i]}{\pi(W_i, \gamma^*)} \right) m(W_i, \theta_0),
\]
\[
= E[\delta_i|W_i],
\]
thus
\[
E \left[ X_i X_i^T \left( -\hat{\delta}_i(\theta_0, \gamma^*)(e^{ \beta^T X_i} + Q_{i,\beta}) + Q_{i,\beta} \right) \right] = E \left[ X_i X_i^T \left( -E[\delta_i|W_i](e^{ \beta^T X_i} + Q_{i,\beta}) + Q_{i,\beta} \right) \right] = E \left[ X_i X_i^T \left( \delta_i e^{ \beta^T X_i} + (\delta_i - 1) Q_{i,\beta} \right) \right] = -\Sigma_1(\beta).
\]

It follows that \( n^{-1} \partial \ell_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)/\partial \beta^T \) converges in probability to \( -\Sigma_1(\beta) \). Uniformity of the convergence follows by the same arguments as in the proof of Theorem 3.1.

Finally, having proved conditions i, ii and iii, we apply the inverse function theorem of Foutz (1977) and conclude that \( \hat{\beta}_n \) converges in probability to \( \beta_0 \) if \( m(W_i, \theta) \) is correctly specified. The consistency proof of \( \hat{\beta}_n \) when model \( \pi(W_i, \gamma) \) is correctly specified proceeds along the same lines and is omitted.

**Appendix D: Proof of Theorem 5.2**

First, we have
\[
\frac{\partial \hat{\delta}_i(\theta, \gamma)}{\partial \theta^T} = \left( 1 - \frac{\xi_i}{\pi(W_i, \gamma)} \right) \hat{m}^T(W_i, \theta) \quad \text{and} \quad \frac{\partial \hat{\delta}_i(\theta, \gamma)}{\partial \gamma^T} = (m(W_i, \theta) - \delta_i) \xi_i \frac{\pi^T(W_i, \gamma)}{\pi^2(W_i, \gamma)}.
\]

Using this, it is straightforward to see that
\[
\frac{1}{n} \frac{\partial \ell_n(\beta_0, \theta^*, \gamma^*)}{\partial \theta^T} \xrightarrow{P} \Sigma_5(\beta_0, \theta^*, \gamma^*) \quad \text{and} \quad \frac{1}{n} \frac{\partial \ell_n(\beta_0, \theta^*, \gamma^*)}{\partial \gamma^T} \xrightarrow{P} \Sigma_6(\beta_0, \theta^*, \gamma^*) \quad (7.16)
\]
as \( n \to \infty \) (calculations are omitted). Moreover, if the model \( m(W_i, \theta) \) is correctly specified (that is, \( \theta^* = \theta_0 \)), then \( \Sigma_6(\beta_0, \theta_0, \gamma^*) = 0 \). Similarly, if model \( \pi(W_i, \gamma) \) is correctly specified (and thus, \( \gamma^* = \gamma_0 \)), then \( \Sigma_5(\beta_0, \theta_0, \gamma^*) = 0 \). Now, taking Taylor’s expansion of \( \hat{\ell}_n(\hat{\beta}_n, \hat{\theta}_n, \hat{\gamma}_n) \) around \((\beta_0, \theta^*, \gamma^*)\) gives

\[
\sqrt{n}(\hat{\beta}_n - \beta_0) = \left( -\frac{1}{n} \frac{\partial \hat{\ell}_n(\beta_0, \theta^*, \gamma^*)}{\partial \beta} \right)^{-1} \left( \frac{1}{\sqrt{n}} \hat{\ell}_n(\beta_0, \theta^*, \gamma^*) + \frac{1}{n} \frac{\partial \hat{\ell}_n(\beta_0, \theta^*, \gamma^*)}{\partial \theta} \sqrt{n}(\hat{\theta}_n - \theta^*) + \frac{1}{n} \frac{\partial \hat{\ell}_n(\beta_0, \theta^*, \gamma^*)}{\partial \gamma} \sqrt{n}(\hat{\gamma}_n - \gamma^*) \right) + o_P(1).
\]

Finally, combining (3.4), (5.6), (7.16) and (7.17) and using the limit central theorem yield the asymptotic distribution of \( \sqrt{n}(\hat{\beta}_n - \beta_0) \) when either \( m(W_i, \theta) \) or \( \pi(W_i, \gamma) \) is correctly specified. Formulas for the asymptotic variance follow from easy albeit tedious calculations.

If both \( m(W_i, \theta) \) and \( \pi(W_i, \gamma) \) are correctly specified, \( \Sigma_7(\beta_0, \theta_0, \gamma_0) = \Sigma_8(\beta_0, \theta_0, \gamma_0) = \Sigma_1(\beta_0) \) and the asymptotic variance of \( \beta_n \) reduces to \( \Sigma_1^{-1}(\beta_0) \), which concludes the proof. \( \square \)

References


<table>
<thead>
<tr>
<th>estimator</th>
<th>correct $m(W, \theta)$ / incorrect $\pi(W, \gamma)$</th>
<th>incorrect $m(W, \theta)$ / correct $\pi(W, \gamma)$</th>
<th>both models correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_3$</td>
</tr>
<tr>
<td>full data</td>
<td>bias</td>
<td>-0.0060</td>
<td>-0.0008</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0766</td>
<td>0.0150</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1099</td>
<td>0.0215</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.9460</td>
<td>0.9490</td>
</tr>
<tr>
<td>CC</td>
<td>bias</td>
<td>0.0688</td>
<td>-0.0038</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0866</td>
<td>0.0163</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1421</td>
<td>0.0238</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.8676</td>
<td>0.9388</td>
</tr>
<tr>
<td>RC</td>
<td>bias</td>
<td>-0.0022</td>
<td>-0.0006</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.1070</td>
<td>0.0173</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1114</td>
<td>0.0218</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.9490</td>
<td>0.9490</td>
</tr>
<tr>
<td>AIPW</td>
<td>bias</td>
<td>-0.0078</td>
<td>-0.0009</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0765</td>
<td>0.0150</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1106</td>
<td>0.0217</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.9410</td>
<td>0.9450</td>
</tr>
<tr>
<td>MI</td>
<td>bias</td>
<td>-0.0026</td>
<td>-0.0006</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0772</td>
<td>0.0150</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1109</td>
<td>0.0216</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.9410</td>
<td>0.9450</td>
</tr>
</tbody>
</table>

Table 1: Simulation results for experiment 1. SE: average standard error. RMSE: root mean square error. CP: empirical coverage probability of 95%-level confidence intervals.
<table>
<thead>
<tr>
<th>estimator</th>
<th>correct $m(W, \theta)$ / incorrect $\pi(W, \gamma)$</th>
<th>incorrect $m(W, \theta)$ / correct $\pi(W, \gamma)$</th>
<th>both models correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_3$</td>
</tr>
<tr>
<td>full data</td>
<td>bias</td>
<td>-0.0015</td>
<td>-0.0001</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0765</td>
<td>0.0150</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1100</td>
<td>0.0213</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.9370</td>
<td>0.9450</td>
</tr>
<tr>
<td>CC</td>
<td>bias</td>
<td>0.1224</td>
<td>0.0121</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0993</td>
<td>0.0186</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1864</td>
<td>0.0289</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.7500</td>
<td>0.8940</td>
</tr>
<tr>
<td>RC</td>
<td>bias</td>
<td>0.0065</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0793</td>
<td>0.0154</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1134</td>
<td>0.0217</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.9430</td>
<td>0.9480</td>
</tr>
<tr>
<td>AIPW</td>
<td>bias</td>
<td>-0.0052</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0765</td>
<td>0.0150</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1115</td>
<td>0.0215</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.9390</td>
<td>0.9440</td>
</tr>
<tr>
<td>MI</td>
<td>bias</td>
<td>0.0056</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0777</td>
<td>0.0150</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1124</td>
<td>0.0215</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>0.9440</td>
<td>0.9420</td>
</tr>
</tbody>
</table>

Table 2: Simulation results for experiment 2. SE: average standard error. RMSE: root mean square error. CP: empirical coverage probability of 95%-level confidence intervals.