Electronic Journal of Statistics
Vol. 15 (2021) 4343–4383
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1897

Censored count data regression with missing censoring information*

Bilel Bousselmi and Jean-François Dupuy†

Univ Rennes, INSA Rennes, CNRS, IRMAR – UMR 6625, F-35000 Rennes, France
e-mail: Bilel.Bousselmi@insa-rennes.fr; Jean-Francois.Dupuy@insa-rennes.fr

Abderrazek Karoui

University of Carthage, Department of Mathematics, Faculty of Sciences of Bizerte, Tunisia
e-mail: Abderrazek.Karoui@fsb.rnu.tn

Abstract: We investigate estimation in 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. These estimates are illustrated on a real data set.

MSC2020 subject classifications: Primary 62J12; secondary 62F12.
Keywords and phrases: Poisson regression, asymptotic properties, missing data, regression calibration, multiple imputation, augmented inverse probability weighting, simulations.

Received December 2020.

Contents

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

†Corresponding author.

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 censored Poisson regression and extensions was therefore addressed by several authors; see, for example, [39], [8], [11], [50], [26] for censored generalized Poisson regression, [22] for finite mixtures of censored Poisson regressions and [33], [29] 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, [24] and [36, 37] address estimation of the survival function of a random survival time with missing censoring indicators. [27], [9] and [5] consider estimation in the proportional hazards and additive hazard regression models with missing censoring indicators. [45], [46] and [6] propose various nonparametric estimates of the hazard and conditional hazard functions with censoring indicators missing at random. [44] and [52] estimate the linear regression and partially linear single-index models for survival data with missing censoring indicators. A similar issue arises in competing risks data analysis with missing cause of failure, see for example [2, 3, 51, 28].

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 adapted estimation methods have been proposed. A common and simple approach, called complete-case analysis, is to exclude individuals with missing data. This method can induce bias and substantial variance increase. Two alternatives are regression calibration and multiple imputation. Regression calibration is a general method for handling missing or mismeasured variables. It consists in replacing missing data by their conditional expectation given observed data. We refer to [7] for a detailed account of this method, which has been used in a variety of contexts, including the linear regression model [19], proportional hazards regression model for survival data [43, 10, 25], generalized linear models [16, 47]. 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. Multiple imputation was also used in a number of settings, including linear regression [17], generalized linear models [21], proportional hazards regression [49, 20] and count data models [23]. Both regression calibration and multiple imputation 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 [34] 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, [40] 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. We also illustrate the proposed estimates on a real data set. 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 $\mathbf{X} = (1, X_2, \ldots, X_p)^\top$ be a p-vector of covariates ($\top$ denotes the transpose operator). We assume that the conditional distribution of Y given $\mathbf{X}$ is given by a Poisson regression model with parameter $\lambda = \exp(\beta^\top \mathbf{X})$, 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 δ (called censoring indicator thereafter) which is equal to 1 if $Y < C$ and 0 if $Y \geq C$. In what follows, we assume that Y and C are independent conditionally on $\mathbf{X}$ and that the distribution of C does not depend on β. These conditions are reminiscent of survival analysis, where they are called the independent censoring and non-informative censoring hypotheses respectively. These hypotheses are reasonable if censoring is due to an external event (i.e., the value of C is not directly driven by the value of Y) or is fixed by design. Otherwise, one needs to model the joint distribution of (Y, C). 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 $(Y_i^*, \mathbf{X}_i, \delta_i)$ (with $i \in \{1, \ldots, n\}$). Under the above hypotheses, the likelihood of β is calculated as:

$$L_n(\beta) = \prod_{i=1}^{n} \mathbb{P}(Y_i = Y_i^* | \mathbf{X}_i)^{\delta_i} \mathbb{P}(Y_i \geq Y_i^* | \mathbf{X}_i)^{1-\delta_i},$$

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

$$\begin{aligned} \ell_n(\beta) \quad = \quad & \sum_{i=1}^{n} \left\{ \delta_i \left(Y_i^* \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Y_i^*!) \right) + \right. \\ & \left. (1-\delta_i) \log \left(1 - \sum_{k=0}^{Y_i^*-1} \frac{e^{-\exp(\beta^\top \mathbf{X}_i) + k\beta^\top \mathbf{X}_i}}{k!} \right) \right\} \end{aligned} \tag{2.1}$$

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

Remark. The censoring variable C does not have to be a count. For example, if $Y_i = 4$ and $C_i = 2.5$, then $Y_i^* = 2.5$ (and $\delta_i = 0$) and the contribution of the observation $(Y_i^*, \delta_i) = (2.5, 0)$ to the likelihood is $\mathbb{P}(Y_i \geq 2.5 | \mathbf{X}_i)$. But Y_i is discrete, hence $\mathbb{P}(Y_i \geq 2.5 | \mathbf{X}_i) = \mathbb{P}(Y_i \geq 3 | \mathbf{X}_i)$ and (2.1) is still valid if we write the contribution of a censored observation as $\mathbb{P}(Y_i \geq \lceil Y_i^* \rceil | \mathbf{X}_i)$, where $\lceil Y_i^* \rceil$ denotes the smallest integer not less than Y_i^*. If C is continuous, contributions of uncensored observations are unchanged since these observations are integer values.

Overall, the loglikelihood (2.1) remains valid when C is continuous, with the appropriate change of notation $\mathbb{P}(Y_i \geq \lceil Y_i^* \rceil | \mathbf{X}_i)$ for censored observations. In

order to keep notations simple, and without loss of generality, we assume that C is discrete (as is also the case in most applications).

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

$$(Y_i^*, \mathbf{X}_i, \delta_i, \xi_i = 1) \quad \text{or} \quad (Y_i^*, \mathbf{X}_i, \xi_i = 0). \tag{2.2}$$

We consider a missing at random (MAR) mechanism, which means that ξ and δ are independent given all other observed variables (a more restrictive assumption is that ξ and δ are independent, which is called "missing completely at random"). In the next sections, we propose, investigate and compare several estimators of β 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 δ_i in (2.1) by its conditional expectation $\mathbb{E}(\delta_i|\mathbf{W}_i)$, where $\mathbf{W}_i$ contains the observed variables Y_i^* and $\mathbf{X}_i$ and eventually (if available) some observed surrogate variables $\mathbf{V}_i$ for δ_i. Thus, we let $\mathbf{W}_i = (Y_i^*, \mathbf{X}_i^\top, \mathbf{V}_i^\top)^\top$ (we denote by q the dimension of $\mathbf{W}_i$). An approximated version of δ_i can then be defined as:

$$\hat{\delta}_i = \xi_i \delta_i + (1 - \xi_i)\mathbb{E}(\delta_i|\mathbf{W}_i).$$

The conditional expectation $\mathbb{E}(\delta_i|\mathbf{W}_i)$ (or conditional probability $\mathbb{P}(\delta_i = 1|\mathbf{W}_i)$) will generally be unknown and will have to be estimated. As is usual with the regression calibration approach, we assume that $\mathbb{E}(\delta_i|\mathbf{W}_i)$ can be specified by a parametric model $m(\mathbf{W}_i, \theta)$, where θ is an unknown q-dimensional parameter with true value θ_0.

Remark. A convenient candidate for $m(\cdot, \cdot)$ is the logistic regression model $m(\mathbf{W}_i, \theta) = \text{logit}^{-1}(\theta^\top \mathbf{W}_i)$ 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 $m(\mathbf{W}_i, \theta)$ with some regularity conditions stated in section 3.2.

At a first stage, we estimate θ_0 by maximizing a likelihood based on complete cases $i \in \{1, \ldots, n|\xi_i = 1\}$ only:

$$\hat{\theta}_n = \arg\max_\theta \prod_{i=1}^n m(\mathbf{W}_i, \theta)^{\xi_i \delta_i} (1 - m(\mathbf{W}_i, \theta))^{\xi_i (1-\delta_i)}. \tag{3.1}$$

Let

$$\dot{m}(\mathbf{W}_i,\theta)=\frac{\partial m(\mathbf{W}_i,\theta)}{\partial\theta},\quad \widetilde{m_i}(\theta)=\frac{\dot{m}(\mathbf{W}_i,\theta)}{m(\mathbf{W}_i,\theta)(1-m(\mathbf{W}_i,\theta))},$$

and

$$\Theta(\theta)=\mathbb{E}\left[\frac{\dot{m}^{\otimes 2}(\mathbf{W},\theta)}{m(\mathbf{W},\theta)(1-m(\mathbf{W},\theta))}\xi\right],$$

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

$$\sqrt{n}(\hat{\theta}_n-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Theta^{-1}(\theta_0)\widetilde{m_i}(\theta_0)\xi_i(\delta_i-m(\mathbf{W}_i,\theta_0))+o_{\mathbb{P}}(1). \tag{3.2}$$

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

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

$$\tilde{\beta}_n=\arg\max_\beta \widetilde{\ell_n}(\beta,\hat{\theta}_n),$$

where

$$\begin{aligned}\widetilde{\ell_n}(\beta,\hat{\theta}_n) &= \sum_{i=1}^{n}\left\{\hat{\delta}_i(\hat{\theta}_n)\left(Y_i^*\beta^\top\mathbf{X}_i-e^{\beta^\top\mathbf{X}_i}-\log(Y_i^*!)\right)\right.\\ &\quad\left.+(1-\hat{\delta}_i(\hat{\theta}_n))\log\left(1-\frac{\Gamma\left(Y_i^*,e^{\beta^\top\mathbf{X}_i}\right)}{(Y_i^*-1)!}\right)\right\}\end{aligned}$$

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 $\mathbf{X}_i$ and $\mathbf{V}_i$ are bounded, for every $i=1,2,\ldots$

C2 The true parameter values β_0 and θ_0 lie in the interior of some bounded sets $\mathcal{B}\subset\mathbb{R}^p$ and $\Theta\subset\mathbb{R}^q$ respectively.

C3 We have $\mathbb{P}(Y^*\ge 1|\xi\delta=0)=1$ and $\mathbb{P}(\delta=1)>0$.

C4 The function $m(\mathbf{w},\theta)$ is differentiable with respect to θ, for every $\mathbf{w}$. For every $\theta,\tilde{\theta}\in\Theta$, $|m(\mathbf{w},\theta)-m(\mathbf{w},\tilde{\theta})|\le h(\mathbf{w})\|\theta-\tilde{\theta}\|$ for some bounded function h, with $\mathbb{E}[h(\mathbf{W})]=v$.

Remark. 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 $\tilde{\beta}_n$, we introduce some further notations. Let h_β be the function defined by:

$$h_\beta(y, x) = \frac{e^{-e^{\beta^\top x} + \beta^\top xy}}{(y-1)! - \Gamma(y, e^{\beta^\top x})} \tag{3.3}$$

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

$$\begin{aligned}
\Sigma_1(\beta) &= \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top \left(\delta e^{\beta^\top \mathbf{X}} + (\delta - 1)\left\{Y^* - e^{\beta^\top \mathbf{X}} - h_\beta(Y^*, \mathbf{X})\right\} h_\beta(Y^*, \mathbf{X})\right)\right], \\
\Sigma_2(\beta, \theta) &= \mathbb{E}\left[\mathbf{X}\dot{m}^\top(\mathbf{W}, \theta)\left(Y^* - e^{\beta^\top \mathbf{X}} - h_\beta(Y^*, \mathbf{X})\right)(1 - \pi(\mathbf{W}))\right], \\
\Sigma_3(\beta, \theta) &= \mathbb{E}\left[\mathbf{X}\dot{m}^\top(\mathbf{W}, \theta)\left(Y^* - e^{\beta^\top \mathbf{X}} - h_\beta(Y^*, \mathbf{X})\right)\right].
\end{aligned}$$

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* $\tilde{\beta}_n \xrightarrow{\mathbb{P}} \beta_0$ *as* $n \to \infty$ *and* $\sqrt{n}(\tilde{\beta}_n - \beta_0)$ *is asymptotically normal with mean zero and variance* Σ*, 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. If $\pi(\mathbf{W})$ is identically equal to 1 (that is, if there is no missing data), Σ 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 $(\mathbb{E}[\mathbf{X}\mathbf{X}^\top e^{\beta_0^\top \mathbf{X}}])^{-1}$ in Poisson regression if $m(\mathbf{W}, \theta_0)$ is identically equal to 1 (that is, no censoring can affect the data).

A consistent estimator of Σ is given by

$$\begin{aligned}
\Sigma_n \;=\; & \Sigma_{1,n}^{-1}(\tilde{\beta}_n, \hat{\theta}_n)\left\{\Sigma_{1,n}(\tilde{\beta}_n, \hat{\theta}_n) + \right. \\
& \left.\left(2\Sigma_{3,n}(\tilde{\beta}_n, \hat{\theta}_n) - \Sigma_{2,n}(\tilde{\beta}_n, \hat{\theta}_n)\right)\Theta_n^{-1}(\hat{\theta}_n)\Sigma_{2,n}^\top(\tilde{\beta}_n, \hat{\theta}_n)\right\}\Sigma_{1,n}^{-1}(\tilde{\beta}_n, \hat{\theta}_n),
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_{1,n}(\beta, \theta) \;=\; & \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top \left(\hat{\delta}_i(\theta)e^{\beta^\top \mathbf{X}_i} + \right. \\
& \left.(\hat{\delta}_i(\theta) - 1)\left\{Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)\right\} h_\beta(Y_i^*, \mathbf{X}_i)\right), \\
\Sigma_{2,n}(\beta, \theta) \;=\; & \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\dot{m}^\top(\mathbf{W}_i, \theta)\left(Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)\right)(1 - \xi_i),
\end{aligned}$$

$$\Sigma_{3,n}(\beta,\theta) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\dot{m}^{\top}(\mathbf{W}_i,\theta)\left(Y_i^* - e^{\beta^{\top}\mathbf{X}_i} - h_{\beta}(Y_i^*,\mathbf{X}_i)\right),$$

$$\Theta_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\dot{m}^{\otimes 2}(\mathbf{W}_i,\theta)}{m(\mathbf{W}_i,\theta)(1-m(\mathbf{W}_i,\theta))}\xi_i.$$

The consistency proof of the variance estimator uses similar arguments as the proof of consistency of $\tilde{\beta}_n$, it is thus omitted. The estimator $\tilde{\beta}_n$ 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 β in our problem.

4. Multiple imputation

In this section, we assume, as in Section 3, that the conditional expectation $\mathbb{E}(\delta_i|\mathbf{W}_i)$ can be specified by a parametric model $m(\mathbf{W}_i,\theta_0)$, and we denote by $\hat{\theta}_n$ the maximum likelihood estimator of θ_0 based on the complete cases $i \in \{1,\ldots,n|\xi_i = 1\}$.

The imputation procedure is as follows. Each missing δ_i is replaced by a random draw from the Bernoulli distribution $\mathcal{B}(m(\mathbf{W}_i,\hat{\theta}_n))$. We obtain a completed data set. This procedure is repeated M times to form M imputed data sets. For a given θ, let $D_{i,j}(\theta) \sim \mathcal{B}(m(\mathbf{W}_i,\theta))$ denote the imputation of δ_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 δ_i is $\xi_i = 1$ (that is, if δ_i is observed) and to $D_{i,j}(\theta)$ if $\xi_i = 0$ (that is, if δ_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 β_0 is obtained by maximizing the imputed log-likelihood

$$\ell_{n,j}^*(\beta,\hat{\theta}_n) = \sum_{i=1}^{n}\left\{\delta_{i,j}^*(\hat{\theta}_n)\left(Y_i^*\beta^{\top}\mathbf{X}_i - e^{\beta^{\top}\mathbf{X}_i} - \log(Y_i^*!)\right)\right.$$
$$\left. +(1-\delta_{i,j}^*(\hat{\theta}_n))\log\left(1 - \frac{\Gamma\left(Y_i^*, e^{\beta^{\top}\mathbf{X}_i}\right)}{(Y_i^*-1)!}\right)\right\}.$$

The final multiple imputation estimator $\hat{\beta}_n^*$ is obtained by averaging the M

estimators $\hat{\beta}^*_{n,j}$ as:

$$\hat{\beta}^*_n = \frac{1}{M}\sum_{j=1}^{M}\hat{\beta}^*_{n,j}.$$

The next theorem gives the asymptotic properties of $\hat{\beta}^*_n$. Its proof is given in Appendix B.

Theorem 4.1. *For $j = 1,\ldots,M$, let $f_{\beta,\theta,j}(\mathcal{O}_i) = \mathbf{X}_i\{\delta^*_{i,j}(\theta)[Y^*_i - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y^*_i,\mathbf{X}_i)] + h_\beta(Y^*_i,\mathbf{X}_i)\}$, where $\mathcal{O}_i$ denotes the observation (2.2). Let also $\Sigma^*_1(\beta,\theta) = var(\frac{1}{M}\sum_{j=1}^{M} f_{\beta,\theta,j}(\mathcal{O}_1)$. If conditions C1-C4 hold, then $\hat{\beta}^*_n \xrightarrow{\mathbb{P}} \beta_0$ as $n\to\infty$ and $\sqrt{n}(\hat{\beta}^*_n - \beta_0)$ is asymptotically normal with mean zero and variance Σ^*, where $\Sigma^* =$*

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

A consistent estimator of Σ^* can be obtained as

$$\begin{aligned}\Sigma^*_n \;=\; & \bar{\Sigma}^{-1}_{1,n}(\hat{\beta}^*_n,\hat{\theta}_n)\left\{\Sigma^*_{1,n}(\hat{\beta}^*_n,\hat{\theta}_n) + \right.\\ & \left.\left(2\Sigma_{3,n}(\hat{\beta}^*_n,\hat{\theta}_n) - \Sigma_{2,n}(\hat{\beta}^*_n,\hat{\theta}_n)\right)\Theta_n^{-1}(\hat{\theta}_n)\Sigma^\top_{2,n}(\hat{\beta}^*_n,\hat{\theta}_n)\right\}\bar{\Sigma}^{-1}_{1,n}(\hat{\beta}^*_n,\hat{\theta}_n),\end{aligned}$$

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

$$\begin{aligned}\Sigma_{1,n,j}(\beta,\theta) \;=\; & \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\left(\delta^*_{i,j}(\theta)e^{\beta^\top\mathbf{X}_i}\right.\\ & \left.+(\delta^*_{i,j}(\theta)-1)\left\{Y^*_i - e^{\beta^\top\mathbf{X}_i} - h_\beta(Y^*_i,\mathbf{X}_i)\right\}h_\beta(Y^*_i,\mathbf{X}_i)\right),\end{aligned}$$

and $\Sigma_{2,n},\Sigma_{3,n}$ and Θ_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 $\mathbb{E}(\delta|\mathbf{W})$. Misspecifying this model is likely to yield biased estimates of the parameters of interest. An alternative approach is to specify the selection probabilities $\pi(\mathbf{W}_i) = \mathbb{P}(\xi_i = 1|\mathbf{W}_i)$ and to use the inverse probability weighting (IPW) of complete-case technique of [18]. 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 $\pi(\mathbf{W}_i), i = 1,\ldots,n$ 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 32] 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 [42], the single-index model [14], the additive hazards model [38] and the accelerated failure time model [35]. 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 [18], 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 [40] for a detailed account on the method and numerous references). Adapting this idea, we propose the following augmented IPW estimating equation for β:

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

The quantities $\mathbb{E}(\delta_i|\mathbf{W}_i)$ and $\pi(\mathbf{W}_i)$ are unknown and have to be estimated. We assume that they can be specified by some parametric models $m(\mathbf{W}_i, \theta)$ and $\pi(\mathbf{W}_i, \gamma)$ respectively, where θ and γ are unknown q-dimensional parameters with true values θ_0 and γ_0. Let $\hat{\theta}_n$ and $\hat{\gamma}_n$ be the maximum likelihood estimates of θ_0 and γ_0. $\hat{\theta}_n$ is given by (3.1). Similarly, $\hat{\gamma}_n$ can be obtained as

$$\hat{\gamma}_n = \arg\max_{\gamma} \prod_{i=1}^{n} \pi(\mathbf{W}_i, \gamma)^{\xi_i} (1 - \pi(\mathbf{W}_i, \gamma))^{1-\xi_i}.$$

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

$$\breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n) = \sum_{i=1}^{n} \mathbf{X}_i \left[\left\{ \frac{\xi_i \delta_i}{\pi(\mathbf{W}_i, \hat{\gamma}_n)} + \left(1 - \frac{\xi_i}{\pi(\mathbf{W}_i, \hat{\gamma}_n)}\right) m(\mathbf{W}_i, \hat{\theta}_n) \right\} \times \left(Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i) \right) + h_\beta(Y_i^*, \mathbf{X}_i) \right].$$

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

$$\breve{\delta}_i(\theta, \gamma) = \frac{\xi_i \delta_i}{\pi(\mathbf{W}_i, \gamma)} + \left(1 - \frac{\xi_i}{\pi(\mathbf{W}_i, \gamma)}\right) m(\mathbf{W}_i, \theta).$$

Assuming the parametric model $\pi(\mathbf{W}_i, \gamma)$ for the selection probabilities, the maximum likelihood estimator $\hat{\gamma}_n$ is asymptotically linear with influence function $\Sigma_4^{-1}(\gamma_0)\widetilde{\pi}_i(\gamma_0)(\xi_i - \pi(\mathbf{W}_i, \gamma_0))$, where

$$\dot{\pi}(\mathbf{W}_i, \gamma) = \frac{\partial \pi(\mathbf{W}_i, \gamma)}{\partial \gamma}, \quad \widetilde{\pi}_i(\gamma) = \frac{\dot{\pi}(\mathbf{W}_i, \gamma)}{\pi(\mathbf{W}_i, \gamma)(1 - \pi(\mathbf{W}_i, \gamma))},$$

and

$$\Sigma_4(\gamma) = \mathbb{E}\left[\frac{\dot{\pi}^{\otimes 2}(\mathbf{W}, \gamma)}{\pi(\mathbf{W}, \gamma)(1 - \pi(\mathbf{W}, \gamma))}\right].$$

That is:

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma_4^{-1}(\gamma_0)\widetilde{\pi}_i(\gamma_0)(\xi_i - \pi(\mathbf{W}_i, \gamma_0)) + o_{\mathbb{P}}(1). \tag{5.1}$$

If the models $m(\mathbf{W}_i, \theta)$ and $\pi(\mathbf{W}_i, \gamma)$ are misspecified, then by [48], there exists θ^* and γ^* such that $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$ and $\hat{\gamma}_n \xrightarrow{\mathbb{P}} \gamma^*$. Moreover, the asymptotic linear expansions for $\hat{\theta}_n$ and $\hat{\gamma}_n$ are given by (3.2) and (5.1), with θ_0 and γ_0 replaced by θ^* and γ^* respectively. If the model $m(\mathbf{W}_i, \theta)$ (respectively $\pi(\mathbf{W}_i, \gamma)$) is correctly specified, then $\theta^* = \theta_0$ (respectively $\gamma^* = \gamma_0$).

Finally, let

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

$$\Sigma_6(\beta, \theta, \gamma) = \mathbb{E}\left[\mathbf{X}\left(Y^* - e^{\beta^\top \mathbf{X}} - h_\beta(Y^*, \mathbf{X})\right)\xi\frac{\dot{\pi}^\top(\mathbf{W}, \gamma)}{\pi^2(\mathbf{W}, \gamma)}\left(m(\mathbf{W}, \theta) - \delta\right)\right],$$

$$\Sigma_7(\beta, \theta, \gamma) = \Sigma_1(\beta) + \left(2\Sigma_3(\beta, \theta) - \Sigma_5(\beta, \theta, \gamma)\right)\Theta^{-1}(\theta)\Sigma_5^\top(\beta, \theta, \gamma),$$

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 γ is a bounded set $\mathcal{G} \subset \mathbb{R}^q$ and the true parameter value γ_0 lies in the interior of $\mathcal{G}$.

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

C7 The function $\pi(\mathbf{w}, \gamma)$ is differentiable with respect to γ, for every $\mathbf{w}$. For every $\gamma, \tilde{\gamma} \in \mathcal{G}$, $|\pi(\mathbf{w}, \gamma) - \pi(\mathbf{w}, \tilde{\gamma})| \leq g(\mathbf{w})\|\gamma - \tilde{\gamma}\|$ for some bounded function g with $\mathbb{E}[g(\mathbf{W})] = u$.

Conditions C5 and C7 for γ and $\pi(\cdot, \cdot)$ are similar to conditions C2 and C4 for θ and $m(\cdot, \cdot)$. We are now in position to state the asymptotic properties of our AIPW estimator of β.

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

From this result, the proposed estimator $\breve{\beta}_n$ is doubly robust, in the sense that it estimates consistently β_0 as long as one of $m(\mathbf{W}_i, \theta)$ and $\pi(\mathbf{W}_i, \gamma)$ is correctly modeled.

Remark. The basic idea of regression calibration and multiple imputation is to replace a missing δ_i by an approximation whose conditional expectation given observed variables is equal to $\mathbb{E}(\delta_i|\mathbf{W}_i)$ (one can check that $\mathbb{E}(\hat{\delta}_i(\theta_0)|\mathbf{W}_i) = \mathbb{E}(\delta^*_{i,j}(\theta_0)|\mathbf{W}_i) = \mathbb{E}(\delta_i|\mathbf{W}_i)$), so that the expectation of the corresponding estimating equations coincide with the expectation of the estimating equation with no missing data. Similarly, one can easily check that if $m(\mathbf{W}_i, \theta)$ (respectively $\pi(\mathbf{W}_i, \gamma)$) is correctly specified, then $\mathbb{E}(\breve{\delta}_i(\theta_0, \gamma^*)|\mathbf{W}_i) = \mathbb{E}(\delta_i|\mathbf{W}_i)$ (respectively $\mathbb{E}(\breve{\delta}_i(\theta^*, \gamma_0)|\mathbf{W}_i) = \mathbb{E}(\delta_i|\mathbf{W}_i)$). Here is the intuition underlying the AIPW method, and the seemingly complicated expression of $\breve{\delta}_i(\theta, \gamma)$.

The proof of Theorem 5.1 is given in Appendix C. The next theorem describes the asymptotic distribution of $\breve{\beta}_n$. Its proof is given in Appendix D.

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

$$\mathbf{J} = \begin{cases} \Sigma_1^{-1}(\beta_0)\Sigma_7(\beta_0, \theta_0, \gamma^*)\Sigma_1^{-1}(\beta_0) \text{ if } m(\mathbf{W}_i, \theta) \text{ is correctly specified,} \\ \Sigma_1^{-1}(\beta_0)\Sigma_8(\beta_0, \theta^*, \gamma_0)\Sigma_1^{-1}(\beta_0) \text{ if } \pi(\mathbf{W}_i, \gamma) \text{ is correctly specified,} \\ \Sigma_1^{-1}(\beta_0) \text{ if both } m(\mathbf{W}_i, \theta) \text{ and } \pi(\mathbf{W}_i, \gamma) \text{ are correctly specified.} \end{cases}$$

In order to estimate the asymptotic variance of $\breve{\beta}_n$, let:

$$\begin{aligned}
\breve{\Sigma}_{1,n}(\beta, \theta, \gamma) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i^\top \Big(\breve{\delta}_i(\theta, \gamma)e^{\beta^\top \mathbf{X}_i} \\
&\quad + (\breve{\delta}_i(\theta, \gamma) - 1)\left\{Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)\right\} h_\beta(Y_i^*, \mathbf{X}_i)\Big), \\
\Sigma_{4,n}(\gamma) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{\pi}^{\otimes 2}(\mathbf{W}_i, \gamma)}{\pi(\mathbf{W}_i, \gamma)(1 - \pi(\mathbf{W}_i, \gamma))}, \\
\Sigma_{5,n}(\beta, \theta, \gamma) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \left(Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)\right) \\
&\quad \times \left(1 - \frac{\xi_i}{\pi(\mathbf{W}_i, \gamma)}\right) \dot{m}^\top(\mathbf{W}_i, \theta), \\
\Sigma_{6,n}(\beta, \theta, \gamma) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \left(Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)\right) \xi_i \\
&\quad \times \frac{\dot{\pi}^\top(\mathbf{W}_i, \gamma)}{\pi^2(\mathbf{W}_i, \gamma)} \left(m(\mathbf{W}_i, \theta) - \delta_i\right),
\end{aligned}$$

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

and

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

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

$$\mathbf{J}_n = \begin{cases} \breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n)\Sigma_{7,n}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n)\breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n) \\ \quad \text{if } m(\mathbf{W}_i,\theta) \text{ is correctly specified,} \\ \breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n)\Sigma_{8,n}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n)\breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n) \\ \quad \text{if } \pi(\mathbf{W}_i,\gamma) \text{ is correctly specified,} \\ \breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n) \\ \quad \text{if both } m(\mathbf{W}_i,\theta) \text{ and } \pi(\mathbf{W}_i,\gamma) \text{ are correctly specified.} \end{cases} \tag{5.2}$$

The proof of consistency of $\mathbf{J}_n$ is omitted.

6. Numerical results

6.1. A simulation study

6.1.1. Simulation design

In this section, we investigate 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 $\beta = (0.2, -0.1, 0.4, 0.3, 0.5)$, $X_2 \sim \mathcal{N}(0,1)$, $X_3 \sim$ Bernoulli(0.3), $X_4 \sim \mathcal{N}(0,1.5)$ and $X_5 \sim$ uniform$[2,5]$. The censoring and missingness mechanisms are set to be $\text{logit}(m(\mathbf{W},\theta)) = \theta_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4 + \theta_5 X_5 + \theta_6 Y$ and $\text{logit}(\pi(\mathbf{W},\gamma)) = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 Y^*$ respectively, where θ and γ are chosen to yield the desired fractions of censored and missing data. We run a series of experiments in order to assess the effect of the sample size, censoring rate (CR) and missing rate (MR) on estimation:

- experiment 1: we take $n = 250$, CR $= 20\%$, MR $= 20\%$,
- experiment 2: we take $n = 500$, CR $= 20\%$, MR $= 20\%$,
- experiment 3: we take $n = 500$, CR $= 20\%$, MR $= 40\%$,
- experiment 4: we take $n = 500$, CR $= 40\%$, MR $= 20\%$,

We can assess the effect of the sample size by comparing results of experiments 1 and 2. Similarly, by comparing experiments 2 and 3 (respectively 2 and 4), we can assess the effect of the missing rate (respectively censoring rate). Within each experiment, we compare the RC, MI and AIPW estimates under three scenario: (i) only $m(\mathbf{W},\theta)$ is correctly modeled, (ii) only $\pi(\mathbf{W},\gamma)$ is correctly modeled, (iii) both $m(\mathbf{W},\theta)$ and $\pi(\mathbf{W},\gamma)$ are correctly modeled. In the first scenario, $\pi(\mathbf{W},\gamma)$ is incorrectly modeled as $\text{logit}(\pi(\mathbf{W},\gamma)) = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 Y^*$. In the second scenario, $m(\mathbf{W},\theta)$ is incorrectly modeled as $\text{logit}(m(\mathbf{W},\theta)) = \theta_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 Y^*$.

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 (2.1) on the subsample of complete cases only. Results of experiment j are summarized in Table j, for $j = 1, \ldots, 4$. All estimates are obtained using the Newton-Raphson algorithm, implemented in `R` [31].

Remark. The EM algorithm is a popular tool for calculating maximum likelihood estimates in missing data problems. In the context of Poisson regression with missing data, it has been used by several authors. For example, [12], [4] and [22] use EM in finite mixtures of Poisson, bivariate Poisson and censored Poisson regression models. In these works, the EM algorithm is motivated by the missing data formulation of mixture models, where the unknown mixture component indicator is treated as the missing data. EM was also used in zero-inflated Poisson regression [15]. Here, the missing data is the unobserved state variable (zero state vs Poisson state). [1] use EM in bivariate Poisson regression with missing outcome. The EM algorithm could also be used in our setting, and it would be interesting to investigate the convergence rate of the sequence of EM estimates. This, however, falls beyond the scope of our paper and constitutes a topic for future work.

6.1.2. Results

As expected, the performance of the estimators improve when sample size increases. In the first scenario of each experiment, the RC, MI and AIPW methods appear to have similar performance. Coverage probabilities are close to the nominal confidence level, indicating that the asymptotic variances are appropriately estimated.

In the second scenario, the AIPW method generally achieves the smallest SE and RMSE, while the bias of the RC and MI estimates increase substantially, resulting in coverage probabilities smaller than desired (this is particularly noticeable when the censoring rate is large, see Table 4). This result was expected

since $m(\mathbf{W},\theta)$ is misspecified. On the other hand, when censoring is moderate (Table 1-Table 3), the bias of the AIPW estimate stays moderate and of the same order of magnitude (but generally slightly larger, see our explanation below) as in the first scenario, which is also expected due to the double robustness property stated in Theorem 5.1. When the censoring rate is high, the bias of the AIPW estimate is more important and coverage probabilities can be affected (but less than for the RC and MI methods). This suggests that in finite samples, the AIPW estimator is more sensitive to a misspecification of $m(\mathbf{W},\theta)$ than of $\pi(\mathbf{W},\gamma)$. This can be explained by the expression of $\breve{\delta}_i(\theta,\gamma)$, which is equal to $m(\mathbf{W}_i,\theta)$ if $\xi_i = 0$ and to $m(\mathbf{W}_i,\theta) + \frac{(\delta_i - m(\mathbf{W}_i,\theta)}{\pi(\mathbf{W}_i,\gamma)}$ if $\xi_i = 1$. Indeed, when $m(\mathbf{W},\theta)$ is wrong, every individual i contributes to the likelihood with a misspecified term, whatever ξ_i is. On the other hand, when $\pi(\mathbf{W},\gamma)$ is wrong, only individuals with $\xi_i = 1$ contribute to the likelihood with a misspecified term, since $\pi(\mathbf{W},\gamma)$ does not appear in the contribution of individuals with $\xi_i = 0$. This unbalance may explain the greater sensitivity of the AIPW estimate to a misspecification of $m(\mathbf{W},\theta)$.

Finally, when both models are correct (third scenario), 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 $m(\mathbf{W},\theta)$ or both $m(\mathbf{W},\theta)$ and $\pi(\mathbf{W},\gamma)$ are correctly specified. When $m(\mathbf{W},\theta)$ is misspecified, the AIPW approach performs better than RC and MI, in particular in terms of point estimation (with substantially smaller bias for AIPW). The CC estimates are outperformed by the three methods in all scenarios.

6.1.3. Asymptotic variance estimation

Estimation of Σ, Σ^* and $\mathbf{J}$ (the asymptotic variances of the RC, MI and AIPW estimates respectively) is a crucial issue for statistical inference purpose. In this section, we investigate the accuracy of their respective estimates Σ_n, Σ_n^* and $\mathbf{J}_n$.

First, note that although Σ, Σ^* and $\mathbf{J}$ have explicit expressions, they cannot be calculated analytically, due to the complex expectations involved in the Σ_j, $j = 1,\dots,8$. Therefore, we propose to compare Σ_n (respectively $\Sigma_n^*, \mathbf{J}_n$) to some "oracle" estimate Σ^{or} (respectively $\Sigma^{*,or}, \mathbf{J}^{or}$), which is obtained as follows: we simulate a very large number (here, 15000) of observations $(Y_i^*, \mathbf{X}_i, \delta_i, \xi_i)$, and we calculate empirical versions of the Σ_j where all expectations are replaced by sample averages and parameters are fixed at their true value (hence the name oracle). We expect these oracles to be as close as possible of the true unknown asymptotic variances. Comparisons between $\Sigma_n, \Sigma_n^*, \mathbf{J}_n$ and the oracles are based on the results of the above simulation study.

For each experiment and each of the RC, MI and AIPW method, we calculate the relative differences $100 \times |\Sigma_{n,(j,j)}^{1/2} - (\Sigma_{(j,j)}^{or})^{1/2}|/(\Sigma_{(j,j)}^{or})^{1/2}$, $100 \times |(\Sigma_{n,(j,j)}^{*})^{1/2} - (\Sigma_{(j,j)}^{*,or})^{1/2}|/(\Sigma_{(j,j)}^{*,or})^{1/2}$ and $100 \times |\mathbf{J}_{n,(j,j)}^{1/2} - (\mathbf{J}_{(j,j)}^{or})^{1/2}|/(\mathbf{J}_{(j,j)}^{or})^{1/2}$

(where $A_{(j,j)}$ is the j-th diagonal element of a matrix A) between the estimated and oracle standard deviations of β_j $(j = 1, \ldots, 5)$ (we calculate the relative error for standard deviations rather than for variance, since standard deviations are used to obtain confidence intervals and Wald test statistics, which are the cornerstones of statistical inference in regression models). Results are averaged over the N simulated samples and reported in Table 5. We also report the RMSE of the RC, MI and AIPW variance estimates (the corresponding oracle variance estimates are used as reference). For example, the RMSE of the RC variance estimate of β_j is calculated as

$$\sqrt{\frac{1}{N}\sum_{\ell=1}^{N}\left(\Sigma^{(\ell)}_{n,(j,j)} - \Sigma^{or}_{(j,j)}\right)^2},$$

where $\Sigma^{(\ell)}_{n,(j,j)}$ denotes the RC variance estimate of β_j in the ℓ-th simulated sample. Results are given in Table 6. Regarding AIPW, we evaluate the three variance estimates given in (5.2) (results are reported at line "AIPW1" for misspecified $\pi(\mathbf{W}, \gamma)$, "AIPW2" for misspecified $m(\mathbf{W}, \theta)$, "AIPW3" when both models are correct, in both Tables 5 and 6).

From these results, it appears that both RMSE and relative differences between estimated and oracle standard deviations decrease when sample size increases and censoring and missing rates decrease. Relative errors and RMSE are the smallest for the AIPW estimate when both $m(\mathbf{W}, \theta)$ and $\pi(\mathbf{W}, \gamma)$ are correctly specified (line AIPW3). In each scenario, the relative errors and RMSE of AIPW are smaller when $\pi(\mathbf{W}, \gamma)$ is misspecified than when $m(\mathbf{W}, \theta)$ is misspecified. In fact, relative errors and RMSE show little sensitivity to misspecification of $\pi(\mathbf{W}, \gamma)$ (lines AIPW1 and AIPW3 are close to each other). On the other hand, when $m(\mathbf{W}, \theta)$ is misspecified and censoring is high, the AIPW relative error can be substantial (around 20%, yielding the low coverage probabilities – around 75% – reported for $\beta_1, \beta_4, \beta_5$ in Table 4). But overall, all variance estimates essentially provide low relative errors (most of them being less than 6%). When $m(\mathbf{W}, \theta)$ is well specified and $\pi(\mathbf{W}, \gamma)$ is misspecified: *i*) the RC (respectively MI) variance estimate performs slightly better (respectively a little less well) than AIPW in terms of relative error, *ii*) AIPW variance estimate performs better than RC and MI in terms of RMSE. These observations suggest that the AIPW variance estimator is superior to RC and MI when both $m(\mathbf{W}, \theta)$ and $\pi(\mathbf{W}, \gamma)$ are correct, and that AIPW and RC variance estimates have similar performance when $\pi(\mathbf{W}, \gamma)$ is misspecified.

6.2. A real data analysis

We apply the proposed estimates to a data set from a survey of daily fruits and vegetables intake. The data were collected by the Office for National Statistics (UK), as part of a larger opinion survey. Respondents were asked about their usual daily intake of fruits and vegetables. Precisely, we have the number

of portions of fruits and vegetables eaten by each respondent the day before the survey, and we know whether this number coincides with the respondent's usual intake or whether it is less than the usual intake. In this latter case, the usual intake is right-censored. The total sample size is $n = 928$. The censoring information is missing for 228 respondents (that is, 24.6% of the sample) and 29.6% of the respondents with known censoring information have a right-censored daily intake. Covariates are gender, age, marital status (married vs single/divorced/separated), educational level (with three levels: "General Certificate of Secondary Education (GCSE) or no qualification", "A-level or equivalent", "higher education") and a factor coding respondents appreciation of their daily intake of fruits and vegetables ("enough", "not enough", "more than enough"). We use logistic regression models (with covariates the number of portions reported by the respondents and the five variables mentioned above) for the conditional expectation of the censoring indicator and the selection probabilities. A forward-and-backward elimination strategy based on the AIC is used to select the final models. Finally, we calculate the RC, MI (with $M = 50$ completed data sets) and AIPW estimates in a Poisson regression model for the usual daily intake of fruits and vegetables.

Results are presented in Table 7 (in this table, *gender*=1 for male and 0 for female; *single*=1 for a single/divorced/separated respondent and 0 for a married respondent; *GCSE/no qualif.*=1 if the respondent has either no qualification or has obtained a GCSE, and 0 otherwise, *A-level or equiv.*=1 if the respondent has obtained a A-level or an equivalent diploma; *more than enough*=1 if the respondent considers that her/his daily intake of fruits and vegetables is more than enough and 0 otherwise, *not enough*=1 if the respondent considers that her/his daily intake is not enough and 0 otherwise).

All methods conclude that age has a significant effect on the daily intake of fruits and vegetables, with older people consuming more than younger ones. The gender effect is not significant (at level 5%) for the CC, RC and AIPW analysis but is significant for the MI method, with women consuming more fruits and vegetables than men. All methods find that being married is associated with increased fruits and vegetables intakes, while being single, separated or divorced is associated with lower consumption. For example, using the MI estimate, we find that on average, being single yields a $(1 - e^{-0.0944}) \times 100 \approx 9\%$ decrease in the daily intake (holding fixed all the other effects). Our results also suggest that individuals with higher education have a higher consumption of fruits and vegetables than those with lower education (the reference level in Table 7 is "higher education"). The difference in daily intake between respondents with an A-level or equivalent diploma and respondents with a higher degree is not significant (although not being far from it, for all methods except the complete-case analysis) but there is a very significant difference between respondents with a GCSE or no qualification and those with a high degree. Finally, respondents who perceive their intake as more than enough (respectively not enough) indeed consume more (respectively less) fruits and vegetables, which may reflect the fact that respondents are well-informed on the usual recommandations about fruits and vegetables intake. Our results are coherent with the findings of previous stud-

TABLE 1. *Simulation results for $n = 250$, censoring rate = 20%, missing rate = 20%. SE: average standard error. RMSE: root mean square error. CP: empirical coverage probability of 95%-level confidence intervals.*

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	-0.0097	-0.0001	0.0005	0.0009	0.0021	-0.0097	-0.0001	0.0005	0.0009	0.0021	-0.0097	-0.0001	0.0005	0.0009	0.0021
	SE	0.1092	0.0213	0.0459	0.0164	0.0267	0.1092	0.0213	0.0459	0.0164	0.0267	0.1092	0.0213	0.0459	0.0164	0.0267
	RMSE	0.1549	0.0307	0.0643	0.0232	0.0380	0.1549	0.0307	0.0643	0.0232	0.0380	0.1549	0.0307	0.0643	0.0232	0.0380
	CP	0.9571	0.9397	0.9510	0.9540	0.9581	0.9571	0.9397	0.9510	0.9540	0.9581	0.9571	0.9397	0.9510	0.9540	0.9581
CC	bias	0.0624	-0.0033	-0.0062	-0.0096	-0.0082	0.0624	-0.0033	-0.0062	-0.0096	-0.0082	0.0624	-0.0033	-0.0062	-0.0096	-0.0082
	SE	0.1235	0.0233	0.0500	0.0187	0.0295	0.1235	0.0233	0.0500	0.0187	0.0295	0.1235	0.0233	0.0500	0.0187	0.0295
	RMSE	0.1842	0.0334	0.0698	0.0281	0.0423	0.1842	0.0334	0.0698	0.0281	0.0423	0.1842	0.0334	0.0698	0.0281	0.0423
	CP	0.9326	0.9418	0.9571	0.9142	0.9540	0.9326	0.9418	0.9571	0.9142	0.9540	0.9326	0.9418	0.9571	0.9142	0.9540
RC	bias	-0.0062	0.0001	-0.0003	0.0004	0.0014	0.0256	0.0026	-0.0069	-0.0042	-0.0061	-0.0062	0.0001	-0.0003	0.0004	0.0014
	SE	0.1111	0.0217	0.0469	0.0167	0.0273	0.1095	0.0217	0.0470	0.0164	0.0268	0.1111	0.0217	0.0469	0.0167	0.0273
	RMSE	0.1568	0.0311	0.0653	0.0235	0.0386	0.1615	0.0316	0.0666	0.0242	0.0396	0.1568	0.0311	0.0653	0.0235	0.0386
	CP	0.9540	0.9438	0.9510	0.9540	0.9540	0.9387	0.9336	0.9428	0.9234	0.9305	0.9540	0.9438	0.9510	0.9540	0.9540
AIPW	bias	-0.0119	-0.0002	0.0008	0.0012	0.0026	-0.0133	-0.0002	0.0014	0.0016	0.0029	-0.0100	-0.0001	0.0005	0.0010	0.0021
	SE	0.1089	0.0213	0.0458	0.0162	0.0267	0.1058	0.0211	0.0453	0.0159	0.0260	0.1092	0.0213	0.0460	0.0164	0.0268
	RMSE	0.1557	0.0308	0.0646	0.0232	0.0383	0.1609	0.0316	0.0660	0.0243	0.0395	0.1557	0.0308	0.0648	0.0233	0.0383
	CP	0.9510	0.9397	0.9428	0.9499	0.9459	0.9152	0.9183	0.9275	0.9122	0.9142	0.9520	0.9397	0.9459	0.9520	0.9489
MI	bias	-0.0069	0.0000	-0.0001	0.0005	0.0015	0.0241	0.0024	-0.0065	-0.0040	-0.0058	-0.0069	0.0000	-0.0001	0.0005	0.0015
	SE	0.1091	0.0212	0.0457	0.0163	0.0267	0.1124	0.0219	0.0476	0.0168	0.0274	0.1091	0.0212	0.0457	0.0163	0.0267
	RMSE	0.1556	0.0308	0.0646	0.0233	0.0383	0.1633	0.0317	0.0671	0.0245	0.0399	0.1556	0.0308	0.0646	0.0233	0.0383
	CP	0.9520	0.9418	0.9479	0.9489	0.9489	0.9459	0.9397	0.9510	0.9356	0.9408	0.9520	0.9418	0.9479	0.9489	0.9489

TABLE 2. *Simulation results for $n = 500$, censoring rate = 20%, missing rate = 20%.*

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
	bias	-0.0060	-0.0008	0.0014	0.0005	0.0013	-0.0060	-0.0008	0.0014	0.0005	0.0013	-0.0060	-0.0008	0.0014	0.0005	0.0013
full data	SE	0.0766	0.0150	0.0323	0.0115	0.0188	0.0766	0.0150	0.0323	0.0115	0.0188	0.0766	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1099	0.0215	0.0453	0.0162	0.0269	0.1099	0.0215	0.0453	0.0162	0.0269	0.1099	0.0215	0.0453	0.0162	0.0269
	CP	0.9460	0.9490	0.9500	0.9560	0.9420	0.9460	0.9490	0.9500	0.9560	0.9420	0.9460	0.9490	0.9500	0.9560	0.9420
	bias	0.0688	-0.0038	-0.0051	-0.0109	-0.0092	0.0688	-0.0038	-0.0051	-0.0109	-0.0092	0.0688	-0.0038	-0.0051	-0.0109	-0.0092
CC	SE	0.0866	0.0163	0.0350	0.0131	0.0207	0.0866	0.0163	0.0350	0.0131	0.0207	0.0866	0.0163	0.0350	0.0131	0.0207
	RMSE	0.1421	0.0238	0.0493	0.0213	0.0310	0.1421	0.0238	0.0493	0.0213	0.0310	0.1421	0.0238	0.0493	0.0213	0.0310
	CP	0.8676	0.9388	0.9519	0.8656	0.9188	0.8676	0.9388	0.9519	0.8656	0.9188	0.8676	0.9388	0.9519	0.8656	0.9188
	bias	-0.0022	-0.0006	0.0008	0.0000	0.0005	0.0311	0.0021	-0.0065	-0.0048	-0.0073	-0.0022	-0.0006	0.0008	0.0000	0.0005
RC	SE	0.0780	0.0153	0.0329	0.0117	0.0192	0.0769	0.0153	0.0330	0.0115	0.0188	0.0780	0.0153	0.0329	0.0117	0.0192
	RMSE	0.1114	0.0218	0.0463	0.0163	0.0273	0.1177	0.0220	0.0476	0.0174	0.0286	0.1114	0.0218	0.0463	0.0163	0.0273
	CP	0.9490	0.9490	0.9520	0.9570	0.9430	0.9100	0.9450	0.9330	0.9160	0.9120	0.9490	0.9490	0.9520	0.9570	0.9430
	bias	-0.0078	-0.0009	0.0020	0.0007	0.0017	-0.0069	-0.0008	0.0018	0.0009	0.0015	-0.0060	-0.0008	0.0017	0.0006	0.0013
AIPW	SE	0.0765	0.0150	0.0322	0.0114	0.0187	0.0747	0.0149	0.0319	0.0112	0.0183	0.0766	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1106	0.0217	0.0459	0.0161	0.0271	0.1151	0.0221	0.0474	0.0170	0.0280	0.1106	0.0217	0.0459	0.0162	0.0271
	CP	0.9410	0.9450	0.9430	0.9500	0.9390	0.8990	0.9280	0.9290	0.9150	0.9070	0.9420	0.9420	0.9450	0.9520	0.9420
	bias	-0.0026	-0.0006	0.0009	0.0000	0.0006	0.0301	0.0019	-0.0062	-0.0047	-0.0071	-0.0026	-0.0006	0.0009	0.0000	0.0006
MI	SE	0.0772	0.0150	0.0324	0.0115	0.0189	0.0805	0.0155	0.0340	0.0120	0.0196	0.0772	0.0150	0.0324	0.0115	0.0189
	RMSE	0.1109	0.0216	0.0460	0.0162	0.0272	0.1199	0.0222	0.0482	0.0177	0.0290	0.1109	0.0216	0.0460	0.0162	0.0272
	CP	0.9410	0.9450	0.9430	0.9460	0.9380	0.9340	0.9500	0.9470	0.9330	0.9300	0.9410	0.9450	0.9430	0.9460	0.9380

TABLE 3. *Simulation results for $n = 500$, censoring rate $= 20\%$, missing rate $= 40\%$.*

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	-0.0015	-0.0001	-0.0001	0.0006	0.0003	-0.0015	-0.0001	-0.0001	0.0006	0.0003	-0.0015	-0.0001	-0.0001	0.0006	0.0003
	SE	0.0765	0.0150	0.0323	0.0115	0.0188	0.0765	0.0150	0.0323	0.0115	0.0188	0.0765	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1100	0.0213	0.0447	0.0162	0.0270	0.1100	0.0213	0.0447	0.0162	0.0270	0.1100	0.0213	0.0447	0.0162	0.0270
	CP	0.9370	0.9450	0.9520	0.9540	0.9440	0.9370	0.9450	0.9520	0.9540	0.9440	0.9370	0.9450	0.9520	0.9540	0.9440
CC	bias	0.1224	0.0121	-0.0128	-0.0177	-0.0152	0.1224	0.0121	-0.0128	-0.0177	-0.0152	0.1224	0.0121	-0.0128	-0.0177	-0.0152
	SE	0.0993	0.0186	0.0391	0.0151	0.0233	0.0993	0.0186	0.0391	0.0151	0.0233	0.0993	0.0186	0.0391	0.0151	0.0233
	RMSE	0.1864	0.0289	0.0562	0.0275	0.0363	0.1864	0.0289	0.0562	0.0275	0.0363	0.1864	0.0289	0.0562	0.0275	0.0363
	CP	0.7500	0.8940	0.9400	0.7800	0.8920	0.7500	0.8940	0.9400	0.7800	0.8920	0.7500	0.8940	0.9400	0.7800	0.8920
RC	bias	0.0065	0.0004	-0.0021	-0.0006	-0.0013	0.0779	0.0010	-0.0173	-0.0113	-0.0181	0.0065	0.0004	-0.0021	-0.0006	-0.0013
	SE	0.0793	0.0154	0.0336	0.0119	0.0196	0.0764	0.0153	0.0335	0.0115	0.0187	0.0793	0.0154	0.0336	0.0119	0.0196
	RMSE	0.1134	0.0217	0.0465	0.0167	0.0281	0.1387	0.0220	0.0504	0.0204	0.0335	0.1134	0.0217	0.0465	0.0167	0.0281
	CP	0.9430	0.9480	0.9580	0.9490	0.9460	0.7750	0.9420	0.9200	0.8080	0.7920	0.9430	0.9480	0.9580	0.9490	0.9460
AIPW	bias	-0.0052	0.0000	0.0004	0.0010	0.0012	-0.0061	-0.0005	0.0008	0.0014	0.0013	-0.0017	0.0000	-0.0002	0.0007	0.0003
	SE	0.0765	0.0150	0.0322	0.0112	0.0188	0.0698	0.0149	0.0310	0.0105	0.0173	0.0765	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1115	0.0215	0.0455	0.0163	0.0276	0.1185	0.0225	0.0480	0.0177	0.0291	0.1113	0.0216	0.0455	0.0164	0.0275
	CP	0.9390	0.9400	0.9410	0.9470	0.9370	0.8501	0.9080	0.8925	0.8273	0.8635	0.9410	0.9410	0.9430	0.9550	0.9420
MI	bias	0.0056	0.0004	-0.0018	-0.0005	-0.0011	0.0763	0.0010	-0.0168	-0.0110	-0.0177	0.0056	0.0004	-0.0018	-0.0005	-0.0011
	SE	0.0777	0.0150	0.0326	0.0116	0.0191	0.0822	0.0156	0.0352	0.0125	0.0200	0.0777	0.0150	0.0326	0.0116	0.0191
	RMSE	0.1124	0.0215	0.0458	0.0165	0.0278	0.1412	0.0222	0.0514	0.0209	0.0340	0.1124	0.0215	0.0458	0.0165	0.0278
	CP	0.9440	0.9420	0.9440	0.9470	0.9420	0.8330	0.9450	0.9390	0.8690	0.8440	0.9440	0.9420	0.9440	0.9470	0.9420

TABLE 4. *Simulation results for $n = 500$, censoring rate = 40%, missing rate = 20%.*

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	0.0039	0.0002	-0.0005	-0.0002	-0.0007	0.0039	0.0002	-0.0005	-0.0002	-0.0007	0.0039	0.0002	-0.0005	-0.0002	-0.0007
	SE	0.0907	0.0180	0.0402	0.0144	0.0233	0.0907	0.0180	0.0402	0.0144	0.0233	0.0907	0.0180	0.0402	0.0144	0.0233
	RMSE	0.1285	0.0256	0.0573	0.0199	0.0327	0.1285	0.0256	0.0573	0.0199	0.0327	0.1285	0.0256	0.0573	0.0199	0.0327
	CP	0.9587	0.9518	0.9420	0.9676	0.9538	0.9587	0.9518	0.9420	0.9676	0.9538	0.9587	0.9518	0.9420	0.9676	0.9538
CC	bias	0.0764	-0.0033	-0.0062	-0.0116	-0.0115	0.0764	-0.0033	-0.0062	-0.0116	-0.0115	0.0764	-0.0033	-0.0062	-0.0116	-0.0115
	SE	0.1049	0.0202	0.0451	0.0168	0.0263	0.1049	0.0202	0.0451	0.0168	0.0263	0.1049	0.0202	0.0451	0.0168	0.0263
	RMSE	0.1668	0.0289	0.0645	0.0261	0.0387	0.1668	0.0289	0.0645	0.0261	0.0387	0.1668	0.0289	0.0645	0.0261	0.0387
	CP	0.8702	0.9440	0.9479	0.8899	0.9272	0.8702	0.9440	0.9479	0.8899	0.9272	0.8702	0.9440	0.9479	0.8899	0.9272
RC	bias	0.0221	0.0012	-0.0037	-0.0027	-0.0051	0.1838	0.0123	-0.0361	-0.0271	-0.0464	0.0221	0.0012	-0.0037	-0.0027	-0.0051
	SE	0.0960	0.0190	0.0429	0.0152	0.0250	0.0903	0.0192	0.0441	0.0142	0.0231	0.0960	0.0190	0.0429	0.0152	0.0250
	RMSE	0.1356	0.0269	0.0605	0.0208	0.0349	0.2279	0.0299	0.0720	0.0346	0.0577	0.1356	0.0269	0.0605	0.0208	0.0349
	CP	0.9548	0.9548	0.9469	0.9676	0.9587	0.4808	0.9036	0.8673	0.5152	0.5034	0.9548	0.9548	0.9469	0.9676	0.9587
AIPW	bias	-0.0036	-0.0004	0.0022	0.0010	0.0011	0.0199	0.0017	-0.0038	-0.0028	-0.0047	0.0046	0.0002	0.0003	-0.0002	-0.0010
	SE	0.0899	0.0178	0.0398	0.0138	0.0231	0.0681	0.0169	0.0366	0.0110	0.0173	0.0907	0.0180	0.0402	0.0144	0.0233
	RMSE	0.1296	0.0263	0.0589	0.0198	0.0332	0.1332	0.0273	0.0617	0.0216	0.0340	0.1302	0.0264	0.0589	0.0201	0.0333
	CP	0.9508	0.9292	0.9272	0.9489	0.9489	0.7443	0.8741	0.8348	0.7443	0.7345	0.9479	0.9292	0.9361	0.9626	0.9459
MI	bias	0.0203	0.0010	-0.0031	-0.0025	-0.0046	0.1781	0.0117	-0.0344	-0.0262	-0.0449	0.0203	0.0010	-0.0031	-0.0025	-0.0046
	SE	0.0933	0.0183	0.0412	0.0147	0.0241	0.1011	0.0199	0.0467	0.0162	0.0258	0.0933	0.0183	0.0412	0.0147	0.0241
	RMSE	0.1337	0.0264	0.0594	0.0205	0.0343	0.2281	0.0301	0.0728	0.0348	0.0577	0.1337	0.0264	0.0594	0.0205	0.0343
	CP	0.9430	0.9390	0.9381	0.9587	0.9508	0.5821	0.9145	0.9046	0.6332	0.5929	0.9430	0.9390	0.9381	0.9587	0.9508

TABLE 5. *Relative errors (in %) of estimated standard deviations for the RC, MI and AIPW methods (with oracle standard deviations as reference values).*

	experiment 1					experiment 2				
	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
RC	3.5064	4.5296	3.5551	4.3983	3.6484	2.4176	3.3005	2.4370	3.0294	2.4596
MI	5.8060	7.7976	6.0465	6.9825	6.0021	3.9850	5.7020	4.3539	5.0137	4.1130
AIPW1	3.5505	4.6329	3.5394	4.5765	3.6785	2.4617	3.3519	2.4448	3.1441	2.5167
AIPW2	4.4184	4.6137	3.7427	5.2623	4.2667	2.8580	3.4715	2.5475	3.6508	2.7213
AIPW3	3.2736	4.4388	3.4035	4.2700	3.3965	2.2022	3.2673	2.3002	2.9241	2.2328
	experiment 3					experiment 4				
	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
RC	2.5299	3.2919	2.7640	3.1681	2.6332	3.1954	3.7548	3.4727	3.6572	3.3338
MI	4.1904	5.4829	4.4390	5.0557	4.2105	4.9915	5.6509	5.0467	5.6463	5.0811
AIPW1	2.5886	3.3626	2.7770	3.6558	2.7379	3.1904	4.0247	3.5270	4.1102	3.4390
AIPW2	6.2583	3.2878	3.7213	7.3639	5.5936	21.6887	5.0476	6.9504	20.6814	21.9974
AIPW3	2.2943	3.2301	2.4348	2.9428	2.3428	2.7317	3.5375	3.0144	3.3137	2.8459

TABLE 6. *Root mean square errors of the RC, MI and AIPW variance estimates (with oracle variances as reference values).*

	experiment 1					experiment 2				
	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
RC	0.2761	0.0141	0.0512	0.0078	0.0173	0.1837	0.0098	0.0342	0.0052	0.0114
MI	0.4431	0.0227	0.0835	0.0117	0.0277	0.3012	0.0162	0.0584	0.0084	0.0186
AIPW1	0.2710	0.0138	0.0483	0.0078	0.0169	0.1801	0.0096	0.0328	0.0051	0.0111
AIPW2	0.3556	0.0133	0.0515	0.0091	0.0206	0.2327	0.0100	0.0353	0.0065	0.0128
AIPW3	0.2467	0.0129	0.0464	0.0072	0.0153	0.1615	0.0094	0.0309	0.0048	0.0099
	experiment 3					experiment 4				
	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
RC	0.2039	0.0100	0.0398	0.0059	0.0128	0.3637	0.0173	0.0829	0.0106	0.0261
MI	0.3242	0.0158	0.0599	0.0089	0.0198	0.5520	0.0241	0.1112	0.0155	0.0376
AIPW1	0.1927	0.0096	0.0371	0.0059	0.0122	0.3254	0.0163	0.0725	0.0098	0.0234
AIPW2	0.4568	0.0094	0.0531	0.0114	0.0249	0.9898	0.0185	0.1106	0.0248	0.0650
AIPW3	0.1694	0.0093	0.0323	0.0050	0.0105	0.2801	0.0146	0.0637	0.0085	0.0193

TABLE 7. *Analysis results for the daily fruits and vegetables intake.*

	CC			RC			MI			AIPW		
	est	se	p-value	est	se	p-value	est	se	p-value	est	se	p-value
constant	1.6372	0.0897	0.0000	1.6042	0.0785	0.0000	1.5980	0.0632	0.0000	1.6187	0.0766	0.0000
gender	-0.0721	0.0452	0.1107	-0.0699	0.0402	0.0818	-0.0701	0.0353	0.0470	-0.0715	0.0385	0.0636
age	0.0027	0.0013	0.0343	0.0027	0.0011	0.0163	0.0028	0.0009	0.0033	0.0026	0.0011	0.0239
single	-0.1175	0.0444	0.0082	-0.0979	0.0382	0.0104	-0.0944	0.0329	0.0041	-0.1039	0.0382	0.0065
GCSE/no qualif.	-0.2600	0.0535	0.0000	-0.2392	0.0477	0.0000	-0.2389	0.0405	0.0000	-0.2389	0.0455	0.0000
A-level or equiv.	-0.1196	0.0797	0.1335	-0.1182	0.0710	0.0961	-0.1195	0.0630	0.0577	-0.1236	0.0675	0.0669
more than enough	0.3749	0.0679	0.0000	0.3746	0.0610	0.0000	0.3753	0.0582	0.0000	0.3712	0.0574	0.0000
not enough	-0.5611	0.0546	0.0000	-0.5045	0.0478	0.0000	-0.5029	0.0424	0.0000	-0.5072	0.0459	0.0000

ies that investigated factors that influence fruits and vegetables consumption, see [30] for example. Although here, the complete-case analysis yields the same conclusions as the RC, MI and AIPW methods, we observe some differences between the CC estimates and the RC, MI and AIPW estimates, which might reflect the bias of the CC method observed in the simulation study. Moreover, the CC estimates have usually larger standard errors, which reflects the loss of efficiency of the method.

In this example, the three AIPW variance estimates given in (5.2) are equal up to 3 digits (for this reason, only one standard error is reported in Table 7). This suggests that both the missingness model and conditional model for the censoring indicators given observed variables are correctly specified. For this reason and in view of the conclusions of the simulation study, we would recommand to use the AIPW estimate for further statistical inference on this data set.

Remark. A naive (but easy to implement, using standard statistical softwares) estimation method consists in fitting an uncensored Poisson regression model to the data (that is, the censored intakes are treated as if they were uncensored). Using this method, the estimated constant is 1.338 (from Table 7, a consensus estimate is around 1.6). As expected, this naive method underestimates the baseline level of fruits and vegetables consumption.

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 asymptotically robust against misspecification of either the model for the missing data or the missingness mechanism. In finite samples, the AIPW estimator seems to be be more sensitive to a misspecification of the censoring mechanism than of the missingness mechanism.

Now, several issues deserve attention. First, in this work, we considered missing censoring indicators in the Poisson regression model, which assumes equidispersion. A similar issue may arise with under- or over-dispersed counts. The generalized Poisson regression model (see [11] for example) is an appealing model for such data. The negative binomial regression model is an other option for modeling over-dispersed counts. When over-dispersion is due to zero-inflation, zero-inflated regression models (such as zero-inflated Poisson, zero-inflated generalized Poisson or zero-inflated negative binomial models) are appropriate. The estimates proposed in our paper may be adapted to these models and similar techniques could be used to investigate their asymptotic properties.

An other topic for further research is as follows. Our estimators rely on parametric models for the missing data and missingness mechanism. It is important

to assess the sensitivity of the statistical inference to deviations to these models. An alternative estimation strategy may use semiparametric or nonparametric estimation of the models for missing data and missingness mechanism, and is also the topic for our future work.

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 [13]. We describe the main steps of the proof and omit calculation details.

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

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

where $h_\beta(y,x)$ is given by (3.3). We first need to show that $\partial\dot{\widetilde{\ell}}_n(\beta,\hat{\theta}_n)/\partial\beta^\top$ exists and is continuous in a neighborhood of β_0. The map $\beta \mapsto \dot{\widetilde{\ell}}_n(\beta,\hat{\theta}_n)$ is trivially differentiable with respect to β and its derivative is given by:

$$\begin{aligned}\frac{\partial\dot{\widetilde{\ell}}_n(\beta,\hat{\theta}_n)}{\partial\beta^\top} &= \sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\Big(-\hat{\delta}_i(\hat{\theta}_n)e^{\beta^\top\mathbf{X}_i} \\ &\quad +(1-\hat{\delta}_i(\hat{\theta}_n))\left\{Y_i^* - e^{\beta^\top\mathbf{X}_i} - h_\beta(Y_i^*,\mathbf{X}_i)\right\}h_\beta(Y_i^*,\mathbf{X}_i)\Big),\end{aligned}$$

which is continuous in β.

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

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

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

$$\begin{aligned}&\mathbb{E}\left[\mathbf{X}\left(\hat{\delta}(\theta_0)(Y^* - e^{\beta_0^\top\mathbf{X}}) + (1-\hat{\delta}(\theta_0))h_{\beta_0}(Y^*,\mathbf{X})\right)\right] \\ &= \mathbb{E}\left[\mathbf{X}\left(\mathbb{E}(\hat{\delta}(\theta_0)|\mathbf{W})(Y^* - e^{\beta_0^\top\mathbf{X}}) + (1-\mathbb{E}(\hat{\delta}(\theta_0)|\mathbf{W}))h_{\beta_0}(Y^*,\mathbf{X})\right)\right] \quad (7.1)\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}(\hat{\delta}(\theta_0)|\mathbf{W}) &= \mathbb{E}(\xi\delta + (1-\xi)\mathbb{E}(\delta|\mathbf{W})|\mathbf{W}) \\ &= \mathbb{E}(\xi|\mathbf{W})\mathbb{E}(\delta|\mathbf{W}) + (1-\mathbb{E}(\xi|\mathbf{W}))\mathbb{E}(\delta|\mathbf{W}) \\ &= \mathbb{E}(\delta|\mathbf{W}).\end{aligned}$$

Therefore, (7.1) is equal to

$$\mathbb{E}\left[\mathbf{X}\left(\mathbb{E}(\delta|\mathbf{W})(Y^* - e^{\beta_0^\top \mathbf{X}}) + (1 - \mathbb{E}(\delta|\mathbf{W}))h_{\beta_0}(Y^*, \mathbf{X})\right)\right]$$
$$= \mathbb{E}\left[\mathbf{X}\left(\delta(Y^* - e^{\beta_0^\top \mathbf{X}}) + (1-\delta)h_{\beta_0}(Y^*, \mathbf{X})\right)\right],$$

which is equal to 0 (this can be seen by taking successively the conditional expectations given $\{\delta = 1\}$ and $\mathbf{X}$). Convergence to 0 of $n^{-1}(\dot{\tilde{\ell}}_n(\beta_0, \hat{\theta}_n) - \dot{\tilde{\ell}}_n(\beta_0, \theta_0))$ 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 $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta, \hat{\theta}_n)/\partial\beta^\top$ converges in probability to a fixed matrix, uniformly in an open neighborhood of β_0. We have:

$$\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta,\theta)}{\partial\beta^\top} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top \Big(-\hat{\delta}_i(\theta)e^{\beta^\top\mathbf{X}_i} + (1-\hat{\delta}_i(\theta))\left\{Y_i^* - e^{\beta^\top\mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)\right\} h_\beta(Y_i^*, \mathbf{X}_i)\Big).$$

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

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

The first term converges to 0 (by the consistency of $\hat{\theta}_n$ and assumptions C1, C2, C4) and $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta, \theta_0)/\partial\beta^\top$ converges in probability to $-\Sigma_1(\beta)$ (by the weak law of large numbers). Therefore, $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta, \hat{\theta}_n)/\partial\beta^\top$ converges in probability to $-\Sigma_1(\beta)$. Under conditions C1 and C2, the derivative of $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta, \hat{\theta}_n)/\partial\beta^\top$ with respect to β is bounded, for every n. Hence the sequence $(n^{-1}\partial\dot{\tilde{\ell}}_n(\beta, \hat{\theta}_n)/\partial\beta^\top)_n$ is equicontinuous. It follows from Ascoli theorem that the convergence of $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta, \hat{\theta}_n)/\partial\beta^\top$ to $-\Sigma_1(\beta)$ is uniform around β_0.

Having proved the conditions of the inverse function theorem, we conclude that $\tilde{\beta}_n$ converges in probability to β_0.

Asymptotic normality. A Taylor's expansion of $\dot{\tilde{\ell}}_n(\tilde{\beta}_n, \hat{\theta}_n)$ around (β_0, θ_0) yields

$$\sqrt{n}(\tilde{\beta}_n - \beta_0) = \left(-\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta_0,\theta_0)}{\partial\beta^\top}\right)^{-1} \times \left(\frac{1}{\sqrt{n}}\dot{\tilde{\ell}}_n(\beta_0,\theta_0) + \frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta_0,\theta_0)}{\partial\theta^\top}\sqrt{n}(\hat{\theta}_n - \theta_0)\right) + o_{\mathbb{P}}(1).$$

We have:

$$\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta,\theta)}{\partial\theta^\top} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\left(Y_i^* - e^{\beta^\top\mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)\right)(1-\xi_i)\dot{m}^\top(\mathbf{W}_i, \theta)$$

$$= \quad \Sigma_2(\beta,\theta) + o_{\mathbb{P}}(1).$$

Combining this and (3.2), we can write:

$$\begin{aligned}\sqrt{n}(\tilde{\beta}_n - \beta_0) &= \left(-\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta_0,\theta_0)}{\partial\beta^\top}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\mathbf{X}_i\left\{\hat{\delta}_i(\theta_0)(Y_i^* - e^{\beta_0^\top\mathbf{X}_i})\right.\\ &\quad\left.+(1-\hat{\delta}_i(\theta_0))h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right\}\\ &\quad+\Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\widetilde{m_i}(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i,\theta_0))\Big] + o_{\mathbb{P}}(1)\\ &:= \left(-\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta_0,\theta_0)}{\partial\beta^\top}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{U}_i + o_{\mathbb{P}}(1).\end{aligned}$$

Now, note that

$$\operatorname{var}\left(\mathbf{X}_i\left\{\hat{\delta}_i(\theta_0)(Y_i^* - e^{\beta_0^\top\mathbf{X}_i}) + (1-\hat{\delta}_i(\theta_0))h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right\}\right) = \Sigma_1(\beta_0),$$

and

$$\begin{aligned}&\operatorname{var}\left(\Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\widetilde{m_i}(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i,\theta_0))\right)\\ &\qquad= \Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\mathbb{E}\left[\widetilde{m_i}^{\otimes 2}(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i,\theta_0))^2\right]\\ &\qquad\qquad\times\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0)\\ &\qquad= \Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0),\end{aligned}$$

since under the missing at random assumption, we have:

$$\begin{aligned}&\mathbb{E}\left[\widetilde{m_i}^{\otimes 2}(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i,\theta_0))^2\right]\\ &= \mathbb{E}\left[\frac{\dot{m}^{\otimes 2}(\mathbf{W}_i,\theta_0)}{\{m(\mathbf{W}_i,\theta_0)(1-m(\mathbf{W}_i,\theta_0))\}^2}\mathbb{E}\left[\xi_i(\delta_i - m(\mathbf{W}_i,\theta_0))^2|\mathbf{W}_i\right]\right]\\ &= \mathbb{E}\left[\frac{\dot{m}^{\otimes 2}(\mathbf{W}_i,\theta_0)}{\{m(\mathbf{W}_i,\theta_0)(1-m(\mathbf{W}_i,\theta_0))\}^2}\mathbb{E}\left[\xi_i|\mathbf{W}_i\right]\mathbb{E}\left[\delta_i - 2\delta_i m(\mathbf{W}_i,\theta_0)\right.\right.\\ &\qquad\left.\left.+m^2(\mathbf{W}_i,\theta_0)|\mathbf{W}_i\right]\right]\\ &= \mathbb{E}\left[\frac{\dot{m}^{\otimes 2}(\mathbf{W}_i,\theta_0)}{m(\mathbf{W}_i,\theta_0)(1-m(\mathbf{W}_i,\theta_0))}\pi(\mathbf{W}_i)\right]\\ &= \Theta(\theta_0).\end{aligned}$$

We consider now the covariance structure of $\mathcal{U}_i$. We have

$$\begin{aligned}&\operatorname{cov}\left(\mathbf{X}_i\left\{\hat{\delta}_i(\theta_0)(Y_i^* - e^{\beta_0^\top\mathbf{X}_i}) + (1-\hat{\delta}_i(\theta_0))h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right\},\right.\\ &\qquad\left.\Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\widetilde{m_i}(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i,\theta_0))\right)\\ =\ &\mathbb{E}\left[\mathbf{X}_i\widetilde{m_i}^\top(\theta_0)\mathbb{E}\left[\left(\hat{\delta}_i(\theta_0)(Y_i^* - e^{\beta_0^\top\mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i)) + h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right)\right.\end{aligned}$$

$$\times\ \xi_i(\delta_i - m(\mathbf{W}_i,\theta_0))|\mathbf{W}_i]] \times \Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0),$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\delta}_i(\theta_0)(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i)) + h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right)\xi_i(\delta_i - m(\mathbf{W}_i,\theta_0))|\mathbf{W}_i\right]\\
&\qquad = \mathbb{E}\left[\xi_i\delta_i(1-m(\mathbf{W}_i,\theta_0))(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i))\right.\\
&\qquad\qquad \left. + \xi_i h_{\beta_0}(Y_i^*,\mathbf{X}_i)(\delta_i - m(\mathbf{W}_i,\theta_0))|\mathbf{W}_i\right]\\
&\qquad = (1-m(\mathbf{W}_i,\theta_0))(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i))m(\mathbf{W}_i,\theta_0)\pi(\mathbf{W}_i),
\end{aligned}$$

therefore,

$$\begin{aligned}
&\operatorname{cov}\left(\mathbf{X}_i\left\{\hat{\delta}_i(\theta_0)(Y_i^* - e^{\beta_0^\top \mathbf{X}_i}) + (1-\hat{\delta}_i(\theta_0))h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right\},\right.\\
&\qquad\qquad \left.\Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\widetilde{m_i}(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i,\theta_0))\right)\\
=\ &\mathbb{E}\left[\mathbf{X}_i\widetilde{m_i}^\top(\theta_0)(1-m(\mathbf{W}_i,\theta_0))(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i))\right.\\
&\qquad \left.\times m(\mathbf{W}_i,\theta_0)\pi(\mathbf{W}_i)\right]\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0)\\
=\ &\mathbb{E}\left[\mathbf{X}_i\dot{m}_i{}^\top(\mathbf{W}_i,\theta_0)(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i))\pi(\mathbf{W}_i)\right]\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0)\\
=\ &\left(\Sigma_3(\beta_0,\theta_0) - \Sigma_2(\beta_0,\theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0).
\end{aligned}$$

It follows that

$$\operatorname{var}(\mathcal{U}_i) = \Sigma_1(\beta_0) + \left(2\Sigma_3(\beta_0,\theta_0) - \Sigma_2(\beta_0,\theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0).$$

Finally, Theorem 3.1 follows from the multivariate central limit theorem and Slutsky's theorem. □

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 $j = 1,\ldots,M$, let

$$\begin{aligned}
\dot{\ell}_{n,j}^*(\beta,\theta) &= \frac{\partial \ell_{n,j}^*(\beta,\theta)}{\partial\beta}\\
&= \sum_{i=1}^n \mathbf{X}_i\left(\delta_{i,j}^*(\theta)\left[Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*,\mathbf{X}_i)\right] + h_\beta(Y_i^*,\mathbf{X}_i)\right) \quad (7.2)\\
&:= \sum_{i=1}^n f_{\beta,\theta,j}(\mathcal{O}_i).
\end{aligned}$$

Then the following holds:

Lemma 7.1. *Under conditions C1, C2 and C4:*

$$\frac{1}{\sqrt{n}}\left[\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n) - n\mathbb{E}[\dot{\ell}^*_{1,j}(\beta_0,\hat{\theta}_n)] - \left(\dot{\ell}^*_{n,j}(\beta_0,\theta_0) - n\mathbb{E}[\dot{\ell}^*_{1,j}(\beta_0,\theta_0)]\right)\right] \xrightarrow{\mathbb{P}} 0 \quad (7.3)$$

as $n \to \infty$.

Proof of Lemma 7.1. In this proof, for notational simplicity, we will write f_θ instead of $f_{\beta_0,\theta,j}$. First, note that

$$\begin{aligned}\frac{1}{\sqrt{n}}\left[\dot{\ell}^*_{n,j}(\beta_0,\theta) - n\mathbb{E}[\dot{\ell}^*_{1,j}(\beta_0,\theta)]\right] &= \frac{1}{\sqrt{n}}\left[\sum_{i=1}^n f_\theta(\mathcal{O}_i) - n\mathbb{E}[f_\theta(\mathcal{O}_1)]\right] \\ &= \mathbb{G}_n f_\theta,\end{aligned}$$

where $\mathbb{G}_n f_\theta$ denotes the empirical process evaluated at f_θ. To prove the lemma, we first prove that the class of functions $\{f_\theta : \theta \in \Theta\}$ is Donsker (see, for example, [41] for a detailed account on empirical processes and Donsker classes). For that purpose, we decompose f_θ in (7.2) as $f_\theta(\mathcal{O}_i) = \mathbf{X}_i(f_{1,\theta}(\mathcal{O}_i) + f_{2,\theta}(\mathcal{O}_i) + f_{3,\theta}(\mathcal{O}_i))$, where $f_{1,\theta}(\mathcal{O}_i) = -\delta^*_{i,j}(\theta)e^{\beta_0^\top \mathbf{X}_i} + h_{\beta_0}(Y_i^*, \mathbf{X}_i)$, $f_{2,\theta}(\mathcal{O}_i) = \delta^*_{i,j}(\theta)Y_i^*$ and $f_{3,\theta}(\mathcal{O}_i) = -\delta^*_{i,j}(\theta)h_{\beta_0}(Y_i^*, \mathbf{X}_i)$ 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 $D_{i,j}(\theta) \sim \mathcal{B}(m(\mathbf{W}_i,\theta))$ as the random variable $1_{\{U_i \le m(\mathbf{W}_i,\theta)\}}$, where U_i is a uniform random variable on $[0,1]$, independent of $\mathcal{O}_i$.

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

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

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

$$|m(\mathbf{w},\theta) - m(\mathbf{w},\tilde{\theta})| \le h(\mathbf{w})\kappa^*,$$

hence $m(\mathbf{w},\theta) - h(\mathbf{w})\kappa^* \le m(\mathbf{w},\tilde{\theta}) \le m(\mathbf{w},\theta) + h(\mathbf{w})\kappa^*$ and thus we have $1_{\{U_i \le m(\mathbf{w},\theta) - h(\mathbf{w})\kappa^*\}} \le 1_{\{U_i \le m(\mathbf{w},\tilde{\theta})\}} \le 1_{\{U_i \le m(\mathbf{w},\theta) + h(\mathbf{w})\kappa^*\}}$. From this, we can see that

$$f^L_\theta(\mathcal{O}_i) \le f_{1,\tilde{\theta}}(\mathcal{O}_i) \le f^U_\theta(\mathcal{O}_i),$$

where

$$\begin{aligned} f^L_\theta(\mathcal{O}_i) &= h_{\beta_0}(Y_i^*, \mathbf{X}_i) - (\xi_i\delta_i + (1-\xi_i)1_{\{U_i \le m(\mathbf{W}_i,\theta) + h(\mathbf{W}_i)\kappa^*\}})e^{\beta_0^\top \mathbf{X}_i}, \\ f^U_\theta(\mathcal{O}_i) &= h_{\beta_0}(Y_i^*, \mathbf{X}_i) - (\xi_i\delta_i + (1-\xi_i)1_{\{U_i \le m(\mathbf{W}_i,\theta) - h(\mathbf{W}_i)\kappa^*\}})e^{\beta_0^\top \mathbf{X}_i}. \end{aligned}$$

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

$$\mathbb{E}\left[\left(f_\theta^U(\mathcal{O}_i) - f_\theta^L(\mathcal{O}_i)\right)^2\right] \leq 2c_1\kappa^* v.$$

Therefore, $[f_\theta^L, f_\theta^U]$ is an ε-bracket for $\{f_{1,\tilde{\theta}} : \tilde{\theta} \in \Theta \cap \mathcal{B}(\theta, \kappa^*)\}$, with $\varepsilon^2 = 2c_1\kappa^* v$. Since we can cover Θ with fewer than $(d/\kappa)^q$ balls of radius κ^*, we can cover $\mathcal{F}_1 = \{f_{1,\tilde{\theta}} : \tilde{\theta} \in \Theta\}$ with fewer than $(d/\kappa)^q$ ε-brackets $[f_\theta^L, f_\theta^U]$, with $\varepsilon = \sqrt{2c_1\kappa^* v}$. The number of such ε-brackets is thus bounded by $(\alpha d/\kappa^*)^q = (2\alpha c_1 dv/\varepsilon^2)^q$, which is order ε^{-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 [41].

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 $\{\mathbb{G}_n f_\theta : \theta \in \Theta\}$ converges in distribution to a tight limit process, and as such, is stochastically equicontinuous. Thus, Lemma 14.3 of [40] and the consistency of $\hat{\theta}_n$ imply that $\mathbb{G}_n f_{\hat{\theta}_n} - \mathbb{G}_n f_{\theta_0} \xrightarrow{\mathbb{P}} 0$, which is exactly (7.3). This concludes the proof. □

We come back to the proof of asymptotic normality. By a Taylor expansion of $\dot{\ell}^*_{n,j}(\hat{\beta}^*_{n,j}, \hat{\theta}_n)$ around β_0 (for $j = 1, \ldots, M$), we have:

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\dot{\ell}^*_{n,j}(\hat{\beta}^*_{n,j}, \hat{\theta}_n) \\ &= \frac{1}{\sqrt{n}}\dot{\ell}^*_{n,j}(\beta_0, \hat{\theta}_n) + \frac{1}{n}\frac{\partial \dot{\ell}^*_{n,j}(\beta_0, \hat{\theta}_n)}{\partial \beta^\top}\sqrt{n}(\hat{\beta}^*_{n,j} - \beta_0) + o_{\mathbb{P}}(1). \end{aligned}$$

Then, using Lemma 7.1, we obtain:

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\dot{\ell}^*_{n,j}(\beta_0, \theta_0) - \sqrt{n}\mathbb{E}[\dot{\ell}^*_{1,j}(\beta_0, \theta_0)] + \sqrt{n}\mathbb{E}[\dot{\ell}^*_{1,j}(\beta_0, \hat{\theta}_n)] \\ &\qquad + \frac{1}{n}\frac{\partial \dot{\ell}^*_{n,j}(\beta_0, \hat{\theta}_n)}{\partial \beta^\top}\sqrt{n}(\hat{\beta}^*_{n,j} - \beta_0) + o_{\mathbb{P}}(1) \\ &= \frac{1}{\sqrt{n}}\dot{\ell}^*_{n,j}(\beta_0, \theta_0) + \sqrt{n}\left(\frac{\partial \mathbb{E}[\dot{\ell}^*_{1,j}(\beta_0, \theta_0)]}{\partial \theta^\top}(\hat{\theta}_n - \theta_0) + o_{\mathbb{P}}(\|\hat{\theta}_n - \theta_0\|)\right) \\ &\qquad + \frac{1}{n}\frac{\partial \dot{\ell}^*_{n,j}(\beta_0, \hat{\theta}_n)}{\partial \beta^\top}\sqrt{n}(\hat{\beta}^*_{n,j} - \beta_0) + o_{\mathbb{P}}(1), \end{aligned} \tag{7.4}$$

where the second line follows from a Taylor expansion of $\mathbb{E}[\dot{\ell}^*_{1,j}(\beta_0, \hat{\theta}_n)]$ around θ_0. Two technical lemmas are now needed:

Lemma 7.2. *For $j = 1, \ldots, M$, we have*

$$\frac{\partial \mathbb{E}[\dot{\ell}^*_{1,j}(\beta, \theta)]}{\partial \theta^\top} = \Sigma_2(\beta, \theta).$$

Proof of Lemma 7.2. First, we note that

$$\begin{aligned}\mathbb{E}[\delta^*_{1,j}(\theta)|\mathbf{W}_1] &= \mathbb{E}[\xi_1\delta_1 + (1-\xi_1)D_{1,j}(\theta)] \\ &= \pi(\mathbf{W}_1)m(\mathbf{W}_1,\theta_0) + (1-\pi(\mathbf{W}_1))m(\mathbf{W}_1,\theta). \qquad (7.5)\end{aligned}$$

Hence, using (7.2) and iterating the expectation with conditioning on $\mathbf{W}_1$, we obtain:

$$\begin{aligned}\mathbb{E}[\dot{\ell}^*_{1,j}(\beta,\theta)] &= \mathbb{E}\left[\mathbf{X}_1\left(\delta^*_{1,j}(\theta)\left[Y_1^* - e^{\beta^\top\mathbf{X}_1} - h_\beta(Y_1^*,\mathbf{X}_1)\right] + h_\beta(Y_1^*,\mathbf{X}_1)\right)\right] \\ &= \mathbb{E}\left[\mathbf{X}_1\left((\pi(\mathbf{W}_1)m(\mathbf{W}_1,\theta_0) + (1-\pi(\mathbf{W}_1))m(\mathbf{W}_1,\theta)\right)\right. \\ &\quad \left.\times\left[Y_1^* - e^{\beta^\top\mathbf{X}_1} - h_\beta(Y_1^*,\mathbf{X}_1)\right] + h_\beta(Y_1^*,\mathbf{X}_1)\right)\Big].\end{aligned}$$

Finally, straightforward calculations yield

$$\begin{aligned}\frac{\partial\mathbb{E}[\dot{\ell}^*_{1,j}(\beta,\theta)]}{\partial\theta^\top} &= \mathbb{E}\left[\mathbf{X}_1(1-\pi(\mathbf{W}_1))\dot{m}^\top(\mathbf{W}_1,\theta)\left(Y_1^* - e^{\beta^\top\mathbf{X}_1} - h_\beta(Y_1^*,\mathbf{X}_1)\right)\right] \\ &= \Sigma_2(\beta,\theta).\end{aligned}$$

□

Lemma 7.3. *For $j = 1,\ldots,M$,*

$$\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n)}{\partial\beta^\top} \xrightarrow{\mathbb{P}} -\Sigma_1(\beta_0).$$

Proof of Lemma 7.3. Let $j = 1,\ldots,M$. Straightforward calculations yield:

$$\begin{aligned}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n)}{\partial\beta^\top} &= \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\left[-\delta^*_{i,j}(\hat{\theta}_n)e^{\beta_0^\top\mathbf{X}_i}\right. \\ &\quad \left.+(1-\delta^*_{i,j}(\hat{\theta}_n))h_{\beta_0}(Y_i^*,\mathbf{X}_i)\left(Y_i^* - e^{\beta_0^\top\mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right)\right].\end{aligned}$$

Then we decompose $n^{-1}\partial\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n)/\partial\beta^\top$ as:

$$\begin{aligned}\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n)}{\partial\beta^\top} &= \left(\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n)}{\partial\beta^\top} - \frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\theta_0)}{\partial\beta^\top}\right) + \frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\theta_0)}{\partial\beta^\top} \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\left[-e^{\beta_0^\top\mathbf{X}_i}(1-\xi_i)(D_{i,j}(\hat{\theta}_n) - D_{i,j}(\theta_0))\right. \\ &\quad +h_{\beta_0}(Y_i^*,\mathbf{X}_i)\left(Y_i^* - e^{\beta_0^\top\mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right) \\ &\quad \left.\times(1-\xi_i)(D_{i,j}(\theta_0) - D_{i,j}(\hat{\theta}_n))\right] + \frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\theta_0)}{\partial\beta^\top} \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top(1-\xi_i)\left[e^{\beta_0^\top\mathbf{X}_i} + h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right.\end{aligned}$$

$$\times \left(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)\right)\Big]$$
$$\times (D_{i,j}(\theta_0) - D_{i,j}(\hat{\theta}_n)) + \frac{1}{n}\frac{\partial \dot{\ell}^*_{n,j}(\beta_0, \theta_0)}{\partial \beta^\top}$$
$$= \frac{1}{n}\sum_{i=1}^n \mathcal{Z}_i(1_{\{U_{i,j} \le m(\mathbf{W}_i, \theta_0)\}} - 1_{\{U_{i,j} \le m(\mathbf{W}_i, \hat{\theta}_n)\}})$$
$$+ \frac{1}{n}\frac{\partial \dot{\ell}^*_{n,j}(\beta_0, \theta_0)}{\partial \beta^\top}, \tag{7.6}$$

where $\mathcal{Z}_i := \mathbf{X}_i\mathbf{X}_i^\top(1-\xi_i)[e^{\beta_0^\top \mathbf{X}_i} + h_{\beta_0}(Y_i^*, \mathbf{X}_i)(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i))]$ (in what follows, we will denote by $\mathcal{Z}_{i,(\ell,k)}$ the (ℓ,k)-th element of $\mathcal{Z}_i$) 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 (7.6). The random variable $|1_{\{U_{i,j} \le m(\mathbf{W}_i, \theta_0)\}} - 1_{\{U_{i,j} \le m(\mathbf{W}_i, \hat{\theta}_n)\}}|$ is equal to 0 or 1 and takes the value 1 with probability $|m(\mathbf{W}_i, \theta_0) - m(\mathbf{W}_i, \hat{\theta}_n)|$. Let $\varepsilon > 0$. Then, for $\ell, k \in \{1, \dots, p\}$, Markov's inequality implies that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{i,(\ell,k)}(1_{\{U_{i,j} \le m(\mathbf{W}_i, \theta_0)\}} - 1_{\{U_{i,j} \le m(\mathbf{W}_i, \hat{\theta}_n)\}})\right| > \varepsilon\right)$$
$$\le \frac{1}{\varepsilon}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{i,(\ell,k)}(1_{\{U_{i,j} \le m(\mathbf{W}_i, \theta_0)\}} - 1_{\{U_{i,j} \le m(\mathbf{W}_i, \hat{\theta}_n)\}})\right|\right].$$

Under conditions C1 and C2, there exists a finite positive constant c_2 such that $|\mathcal{Z}_{i,(\ell,k)}| \le c_2$. Thus,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{i,(\ell,k)}(1_{\{U_{i,j} \le m(\mathbf{W}_i, \theta_0)\}} - 1_{\{U_{i,j} \le m(\mathbf{W}_i, \hat{\theta}_n)\}})\right| > \varepsilon\right)$$
$$\le \frac{c_2}{\varepsilon n}\sum_{i=1}^n |m(\mathbf{W}_i, \theta_0) - m(\mathbf{W}_i, \hat{\theta}_n)|$$
$$\le \frac{c_2}{\varepsilon n}\sum_{i=1}^n h(\mathbf{W}_i)\|\theta_0 - \hat{\theta}_n\|$$
$$\le \frac{c_2}{\varepsilon}\|\theta_0 - \hat{\theta}_n\|(v + o_{\mathbb{P}}(1)),$$

where the last two lines follow from the condition C4. Finally, consistency of $\hat{\theta}_n$ implies that $\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{i,(\ell,k)}(1_{\{U_{i,j} \le m(\mathbf{W}_i, \theta_0)\}} - 1_{\{U_{i,j} \le m(\mathbf{W}_i, \hat{\theta}_n)\}})$ converges in probability to 0, and the first term in the right-hand side of (7.6) also converges to 0.

We consider now the second term in the right-hand side of (7.6). By the weak law of large numbers, $n^{-1}\partial \dot{\ell}^*_{n,j}(\beta_0, \theta_0)/\partial \beta^\top$ converges in probability to

$$\mathbb{E}\left[\mathbf{X}_1\mathbf{X}_1^\top\left[-\delta^*_{1,j}(\theta_0)e^{\beta_0^\top \mathbf{X}_1} + (1 - \delta^*_{1,j}(\theta_0))h_{\beta_0}(Y_1^*, \mathbf{X}_1)\right.\right.$$

$$\times \left(Y_1^* - e^{\beta_0^\top \mathbf{X}_1} - h_{\beta_0}(Y_1^*, \mathbf{X}_1)\right)\Big]\Big]. \tag{7.7}$$

Using the fact that $\mathbb{E}[\delta_{1,j}^*(\theta_0)|\mathbf{W}_1] = m(\mathbf{W}_1, \theta_0)$ (see (7.5)), and iterating the expectation in (7.7) with conditioning on $\mathbf{W}_1$, we easily show that (7.7) is equal to $-\Sigma_1(\beta_0)$.

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

By combining (7.4) with Lemmas 7.2 and 7.3, we obtain the following approximation of $\hat{\beta}_{n,j}^*$:

$$\begin{aligned}
&\sqrt{n}(\hat{\beta}_{n,j}^* - \beta_0) \\
&= \Sigma_1^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\mathbf{X}_i\left\{\delta_{i,j}^*(\theta_0)[Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)] + h_{\beta_0}(Y_i^*, \mathbf{X}_i)\right\} \\
&\qquad + \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i, \theta_0))\Big] + o_{\mathbb{P}}(1),
\end{aligned}$$

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

$$\begin{aligned}
&\sqrt{n}(\hat{\beta}_n^* - \beta_0) \\
= \;& \frac{1}{M}\sum_{j=1}^{M}\left(\sqrt{n}(\hat{\beta}_{n,j}^* - \beta_0)\right) \\
= \;& \Sigma_1^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{1}{M}\sum_{j=1}^{M}\mathbf{X}_i\left\{\delta_{i,j}^*(\theta_0)[Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)]\right.\right. \\
& \left.\left. + h_{\beta_0}(Y_i^*, \mathbf{X}_i)\right\} + \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i, \theta_0))\Big] + o_{\mathbb{P}}(1) \\
:= \;& \Sigma_1^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{1}{M}\sum_{j=1}^{M} f_{\beta_0,\theta_0,j}(\mathcal{O}_i) + \mathcal{V}_i\right] + o_{\mathbb{P}}(1),
\end{aligned} \tag{7.8}$$

where $\mathcal{V}_i := \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i, \theta_0))$. We have already shown (see proof of Theorem 3.1) that

$$\mathrm{var}(\mathcal{V}_i) = \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0).$$

Similar calculations as in the proof of Theorem 3.1 yield:

$$\mathrm{cov}(f_{\beta_0,\theta_0,j}(\mathcal{O}_i), \mathcal{V}_i) = \left(\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0).$$

Therefore,

$$\mathrm{var}\left(\frac{1}{M}\sum_{j=1}^{M} f_{\beta_0,\theta_0,j}(\mathcal{O}_i) + \mathcal{V}_i\right)$$

$$
\begin{aligned}
&= \operatorname{var}\left(\frac{1}{M}\sum_{j=1}^{M} f_{\beta_0,\theta_0,j}(\mathcal{O}_i)\right) + \operatorname{var}(\mathcal{V}_i) + \frac{2}{M}\sum_{j=1}^{M}\operatorname{cov}(f_{\beta_0,\theta_0,j}(\mathcal{O}_i),\mathcal{V}_i)\\
&= \Sigma_1^*(\beta_0,\theta_0) + \Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0)\\
&\qquad +2\left(\Sigma_3(\beta_0,\theta_0) - \Sigma_2(\beta_0,\theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0)\\
&= \Sigma_1^*(\beta_0,\theta_0) + \left(2\Sigma_3(\beta_0,\theta_0) - \Sigma_2(\beta_0,\theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0). \qquad (7.9)
\end{aligned}
$$

Finally, it follows from (7.8), (7.9) 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

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

which concludes the proof. □

Appendix C: Proof of Theorem 5.1

Assume that the model $m(\mathbf{W}_i,\theta)$ is correctly specified. It is straightforward to check that the map $\beta \mapsto \partial\breve{\ell}_n(\beta,\hat{\theta}_n,\hat{\gamma}_n)/\partial\beta$ exists and is continuous in a neighborhood of β_0 (condition i).

Now, we show that $n^{-1}\breve{\ell}_n(\beta_0,\hat{\theta}_n,\hat{\gamma}_n) = o_{\mathbb{P}}(1)$ (condition ii). To see this, decompose $n^{-1}\breve{\ell}_n(\beta_0,\hat{\theta}_n,\hat{\gamma}_n)$ as:

$$
\begin{aligned}
\frac{1}{n}\breve{\ell}_n(\beta_0,\hat{\theta}_n,\hat{\gamma}_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\frac{\xi_i}{\pi(\mathbf{W}_i,\hat{\gamma}_n)}\left(\delta_i - m(\mathbf{W}_i,\hat{\theta}_n)\right)\\
&\qquad\times\left(Y_i^* - e^{\beta_0^\top\mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right)\\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left\{m(\mathbf{W}_i,\hat{\theta}_n)\left(Y_i^* - e^{\beta_0^\top\mathbf{X}_i}\right)\right.\\
&\qquad\left. + \left(1 - m(\mathbf{W}_i,\hat{\theta}_n)\right)h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right\},\\
&:= Q_n^{(1)}(\hat{\theta}_n,\hat{\gamma}_n) + Q_n^{(2)}(\hat{\theta}_n).
\end{aligned}
$$

First, we consider the term $Q_n^{(1)}(\hat{\theta}_n,\hat{\gamma}_n)$. Let $\mathcal{Q}_i \equiv \mathbf{X}_i\xi_i(Y_i^* - e^{\beta_0^\top\mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i))$. We have:

$$
\begin{aligned}
Q_n^{(1)}(\hat{\theta}_n,\hat{\gamma}_n) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi(\mathbf{W}_i,\hat{\gamma}_n)}(\delta_i - m(\mathbf{W}_i,\hat{\theta}_n))\mathcal{Q}_i,\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\pi(\mathbf{W}_i,\hat{\gamma}_n)} - \frac{1}{\pi(\mathbf{W}_i,\gamma^*)} + \frac{1}{\pi(\mathbf{W}_i,\gamma^*)}\right)\\
&\qquad\times(\delta_i - m(\mathbf{W}_i,\hat{\theta}_n))\mathcal{Q}_i,\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\pi(\mathbf{W}_i,\hat{\gamma}_n)} - \frac{1}{\pi(\mathbf{W}_i,\gamma^*)}\right)(\delta_i - m(\mathbf{W}_i,\hat{\theta}_n))\mathcal{Q}_i
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi(\mathbf{W}_i,\gamma^*)}(\delta_i - m(\mathbf{W}_i,\theta_0))\mathcal{Q}_i \\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi(\mathbf{W}_i,\gamma^*)}(m(\mathbf{W}_i,\theta_0) - m(\mathbf{W}_i,\hat{\theta}_n))\mathcal{Q}_i \\
:= \quad & Q_{n,1}^{(1)} + Q_{n,2}^{(1)} + Q_{n,3}^{(1)}.
\end{aligned}
$$

Now, letting $Q_{n,1,\ell}^{(1)}$ and $\mathcal{Q}_{i,\ell}$ denote the ℓ-th component of the vectors $Q_{n,1}^{(1)}$ and $\mathcal{Q}_i$ respectively (for $\ell = 1,\ldots,p$), we have:

$$
|Q_{n,1,\ell}^{(1)}| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\pi(\mathbf{W}_i,\gamma^*) - \pi(\mathbf{W}_i,\hat{\gamma}_n)}{\pi(\mathbf{W}_i,\hat{\gamma}_n)\pi(\mathbf{W}_i,\gamma^*)}\right||\delta_i - m(\mathbf{W}_i,\hat{\theta}_n)||\mathcal{Q}_{i,\ell}|.
$$

Conditions C1 and C6 ensure that there exists a finite positive constant c_3 such that

$$
|Q_{n,1,\ell}^{(1)}| \leq \frac{c_3}{n}\sum_{i=1}^{n}\left|\pi(\mathbf{W}_i,\gamma^*) - \pi(\mathbf{W}_i,\hat{\gamma}_n)\right|,
$$

and the condition C7 implies that

$$
\begin{aligned}
|Q_{n,1,\ell}^{(1)}| &\leq \frac{c_3}{n}\sum_{i=1}^{n} g(\mathbf{W}_i)\|\gamma^* - \hat{\gamma}_n\|, \\
&\leq c_3(u + o_{\mathbb{P}}(1))\|\gamma^* - \hat{\gamma}_n\|.
\end{aligned}
$$

Finally, the convergence of $\hat{\gamma}_n$ to γ^* 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

$$
|Q_{n,3,\ell}^{(1)}| \leq \frac{c_4}{n}\sum_{i=1}^{n}\left|m(\mathbf{W}_i,\theta_0) - m(\mathbf{W}_i,\hat{\theta}_n)\right|, \quad \ell = 1,\ldots,p,
$$

and condition C4 implies

$$
|Q_{n,3,\ell}^{(1)}| \leq c_4(v + o_{\mathbb{P}}(1))\|\theta_0 - \hat{\theta}_n\|.
$$

If the model $m(\mathbf{W}_i,\theta)$ is correctly specified (that is, if $\hat{\theta}_n$ is consistent for θ_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

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{\pi(\mathbf{W}_i,\gamma^*)}(\delta_i - m(\mathbf{W}_i,\theta_0))\mathcal{Q}_i\right] \\
&= \mathbb{E}\left[\frac{\mathbf{X}_i(Y_i^* - e^{\beta_0^\top\mathbf{X}_i} - h_{\beta_0}(Y_i^*,\mathbf{X}_i))}{\pi(\mathbf{W}_i,\gamma^*)}\mathbb{E}(\xi_i|\mathbf{W}_i)(\mathbb{E}(\delta_i|\mathbf{W}_i) - m(\mathbf{W}_i,\theta_0))\right],
\end{aligned}
$$

which equals 0 if the model $m(\mathbf{W}_i, \theta)$ is correctly specified (in this case, $m(\mathbf{W}_i, \theta_0) = \mathbb{E}(\delta_i|\mathbf{W}_i)$). It follows that $Q_n^{(1)}(\hat{\theta}_n, \hat{\gamma}_n) = o_{\mathbb{P}}(1)$. Therefore,

$$\frac{1}{n}\breve{\ell}_n(\beta_0, \hat{\theta}_n, \hat{\gamma}_n) = Q_n^{(2)}(\hat{\theta}_n) + o_{\mathbb{P}}(1).$$

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 $\mathbb{E}[\mathbf{X}\{m(\mathbf{W}, \theta_0)(Y^* - e^{\beta_0^\top \mathbf{X}}) + (1 - m(\mathbf{W}, \theta_0))h_{\beta_0}(Y^*, \mathbf{X})\}]$, which is equal to 0 (see proof of Theorem 3.1). We also have

$$\begin{aligned} &Q_n^{(2)}(\hat{\theta}_n) - Q_n^{(2)}(\theta_0) \\ &\quad = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i(m(\mathbf{W}_i, \hat{\theta}_n) - m(\mathbf{W}_i, \theta_0))(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)), \end{aligned}$$

and using similar arguments as for $Q_{n,3}^{(1)}$, we can show that this converges to 0 if model $m(\mathbf{W}_i, \theta)$ is correctly specified. Finally, $Q_n^{(2)}(\hat{\theta}_n) = o_{\mathbb{P}}(1)$, which concludes the proof of condition *ii*.

Now, we prove that $n^{-1}\partial\breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)/\partial\beta^\top$ converges to $-\Sigma_1(\beta)$, uniformly in a neighborhood of β_0 (condition *iii*). To see this, let $\mathcal{Q}_{i,\beta} = (Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i))h_\beta(Y_i^*, \mathbf{X}_i)$. Some easy calculations yield:

$$\begin{aligned} \frac{1}{n}\frac{\partial\breve{\ell}_n(\beta, \theta, \gamma)}{\partial\beta^\top} &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\left[-\breve{\delta}_i(\theta, \gamma)(e^{\beta^\top \mathbf{X}_i} + \mathcal{Q}_{i,\beta}) + \mathcal{Q}_{i,\beta}\right], \\ &= -\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top(e^{\beta^\top \mathbf{X}_i} + \mathcal{Q}_{i,\beta})m(\mathbf{W}_i, \theta) + \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top \mathcal{Q}_{i,\beta} \\ &\quad + \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top \frac{\xi_i}{\pi(\mathbf{W}_i, \gamma)}(m(\mathbf{W}_i, \theta) - \delta_i)(e^{\beta^\top \mathbf{X}_i} + \mathcal{Q}_{i,\beta}). \end{aligned}$$

Now, decompose $n^{-1}\partial\breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)/\partial\beta^\top$ as

$$\begin{aligned} \frac{1}{n}\frac{\partial\breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)}{\partial\beta^\top} &= \frac{1}{n}\frac{\partial\breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)}{\partial\beta^\top} - \frac{1}{n}\frac{\partial\breve{\ell}_n(\beta, \theta_0, \gamma^*)}{\partial\beta^\top} + \frac{1}{n}\frac{\partial\breve{\ell}_n(\beta, \theta_0, \gamma^*)}{\partial\beta^\top}, \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top(e^{\beta^\top \mathbf{X}_i} + \mathcal{Q}_{i,\beta})(m(\mathbf{W}_i, \theta_0) - m(\mathbf{W}_i, \hat{\theta}_n)) \\ &\quad + \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top \xi_i(e^{\beta^\top \mathbf{X}_i} + \mathcal{Q}_{i,\beta})\frac{m(\mathbf{W}_i, \hat{\theta}_n) - \delta_i}{\pi(\mathbf{W}_i, \hat{\gamma}_n)} \\ &\quad - \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top \xi_i(e^{\beta^\top \mathbf{X}_i} + \mathcal{Q}_{i,\beta})\frac{m(\mathbf{W}_i, \theta_0) - \delta_i}{\pi(\mathbf{W}_i, \gamma^*)} \\ &\quad + \frac{1}{n}\frac{\partial\breve{\ell}_n(\beta, \theta_0, \gamma^*)}{\partial\beta^\top}, \end{aligned}$$

$$\equiv \quad T_n^{(1)} + T_n^{(2)} + T_n^{(3)} + \frac{1}{n}\frac{\partial \breve{\ell}_n(\beta, \theta_0, \gamma^*)}{\partial \beta^\top}.$$

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, $n^{-1}\partial \breve{\ell}_n(\beta, \theta_0, \gamma^*)/\partial \beta^\top$ converges to $\mathbb{E}[\mathbf{X}_i\mathbf{X}_i^\top(-\breve{\delta}_i(\theta_0, \gamma^*)(e^{\beta^\top \mathbf{X}_i} + \mathcal{Q}_{i,\beta}) + \mathcal{Q}_{i,\beta})]$ in probability. If the model $m(\mathbf{W}_i, \theta)$ is correctly specified (that is, $m(\mathbf{W}_i, \theta_0) = \mathbb{E}(\delta_i|\mathbf{W}_i)$), we have

$$\begin{aligned}\mathbb{E}[\breve{\delta}_i(\theta_0, \gamma^*)|\mathbf{W}_i] &= \frac{\mathbb{E}[\xi_i|\mathbf{W}_i]\mathbb{E}[\delta_i|\mathbf{W}_i]}{\pi(\mathbf{W}_i, \gamma^*)} + \left(1 - \frac{\mathbb{E}[\xi_i|\mathbf{W}_i]}{\pi(\mathbf{W}_i, \gamma^*)}\right) m(\mathbf{W}_i, \theta_0), \\ &= \mathbb{E}[\delta_i|\mathbf{W}_i],\end{aligned}$$

thus

$$\begin{aligned}&\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top\left(-\breve{\delta}_i(\theta_0, \gamma^*)(e^{\beta^\top \mathbf{X}_i} + \mathcal{Q}_{i,\beta}) + \mathcal{Q}_{i,\beta}\right)\right] \\ &\quad= \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top\left(-\mathbb{E}[\delta_i|\mathbf{W}_i](e^{\beta^\top \mathbf{X}_i} + \mathcal{Q}_{i,\beta}) + \mathcal{Q}_{i,\beta}\right)\right] \\ &\quad= -\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top\left(\delta_i e^{\beta^\top \mathbf{X}_i} + (\delta_i - 1)\mathcal{Q}_{i,\beta}\right)\right] \\ &\quad= -\Sigma_1(\beta).\end{aligned}$$

It follows that $n^{-1}\partial \breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)/\partial \beta^\top$ 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 [13] and conclude that $\breve{\beta}_n$ converges in probability to β_0 if $m(\mathbf{W}_i, \theta)$ is correctly specified. The consistency proof of $\breve{\beta}_n$ when model $\pi(\mathbf{W}_i, \gamma)$ is correctly specified proceeds along the same lines and is omitted. □

Appendix D: Proof of Theorem 5.2

First, we have

$$\frac{\partial \breve{\delta}_i(\theta, \gamma)}{\partial \theta^\top} = \left(1 - \frac{\xi_i}{\pi(\mathbf{W}_i, \gamma)}\right) \dot{m}^\top(\mathbf{W}_i, \theta)$$

and

$$\frac{\partial \breve{\delta}_i(\theta, \gamma)}{\partial \gamma^\top} = (m(\mathbf{W}_i, \theta) - \delta_i)\xi_i \frac{\dot{\pi}^\top(\mathbf{W}_i, \gamma)}{\pi^2(\mathbf{W}_i, \gamma)}.$$

Using this, it is straightforward to see that

$$\frac{1}{n}\frac{\partial \breve{\ell}_n(\beta_0, \theta^*, \gamma^*)}{\partial \theta^\top} \xrightarrow{\mathbb{P}} \Sigma_5(\beta_0, \theta^*, \gamma^*), \frac{1}{n}\frac{\partial \breve{\ell}_n(\beta_0, \theta^*, \gamma^*)}{\partial \gamma^\top} \xrightarrow{\mathbb{P}} \Sigma_6(\beta_0, \theta^*, \gamma^*) \quad (7.10)$$

as $n \to \infty$ (calculations are omitted). Moreover, if the model $m(\mathbf{W}_i, \theta)$ is correctly specified (that is, $\theta^* = \theta_0$), then $\Sigma_6(\beta_0, \theta_0, \gamma^*) = 0$. Similarly, if model $\pi(\mathbf{W}_i, \gamma)$ is correctly specified (and thus, $\gamma^* = \gamma_0$), then $\Sigma_5(\beta_0, \theta_0, \gamma^*) = 0$. Now, taking Taylor's expansion of $\breve{\ell}_n(\breve{\beta}_n, \hat{\theta}_n, \hat{\gamma}_n)$ around $(\beta_0, \theta^*, \gamma^*)$ gives

$$\begin{aligned}\sqrt{n}(\breve{\beta}_n - \beta_0) &= \left(-\frac{1}{n}\frac{\partial \breve{\ell}_n(\beta_0, \theta^*, \gamma^*)}{\partial \beta^\top}\right)^{-1} \left(\frac{1}{\sqrt{n}}\breve{\ell}_n(\beta_0, \theta^*, \gamma^*) + \frac{1}{n}\frac{\partial \breve{\ell}_n(\beta_0, \theta^*, \gamma^*)}{\partial \theta^\top}\right. \\ &\quad \left. \times \sqrt{n}(\hat{\theta}_n - \theta^*) + \frac{1}{n}\frac{\partial \breve{\ell}_n(\beta_0, \theta^*, \gamma^*)}{\partial \gamma^\top}\sqrt{n}(\hat{\gamma}_n - \gamma^*)\right) + o_{\mathbb{P}}(1). \quad (7.11)\end{aligned}$$

Finally, combining (3.2), (5.1), (7.10) and (7.11) and using the limit central theorem yield the asymptotic distribution of $\sqrt{n}(\breve{\beta}_n - \beta_0)$ when either $m(\mathbf{W}_i, \theta)$ or $\pi(\mathbf{W}_i, \gamma)$ is correctly specified. Formulas for the asymptotic variance follow from easy albeit tedious calculations.

If both $m(\mathbf{W}_i, \theta)$ and $\pi(\mathbf{W}_i, \gamma)$ are correctly specified, then $\Sigma_7(\beta_0, \theta_0, \gamma_0) = \Sigma_8(\beta_0, \theta_0, \gamma_0) = \Sigma_1(\beta_0)$ and the asymptotic variance of $\breve{\beta}_n$ reduces to $\Sigma_1^{-1}(\beta_0)$, which concludes the proof. □

Acknowledgments

Authors are grateful to two referees and the Associate Editor for their comments and suggestions that led substantial improvements of this paper.

References

[1] Adamids, K., Loukas, S., 1994. Ml estimation in the bivariate poisson distribution in the presence of missing values via the EM algorithm. Journal of Statistical Computation and Simulation 50(12), 163-172. MR1418998

[2] Bakoyannis, G., Siannis, F., Touloumi, G., 2010. Modelling competing risks data with missing cause of failure. Biometrics 29(30), 3172-3185. MR2758711

[3] Bakoyannis, G., Zhang, Y., Yiannoutsos, C.T., 2020. Semiparametric regression and risk prediction with competing risks data under missing cause of failure. Lifetime Data Analysis 26(4), 659-684. MR4148442

[4] Bermúdez, L., Karlis, D., 2012. A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking. Computational Statistics & Data Analysis 56(12), 3988-3999. MR2957848

[5] Brownstein, N.C., Bunn, V., Castro, L.M., Sinha, D., 2021. Bayesian analysis of survival data with missing censoring indicators. Biometrics 77(1), 305-315. MR4229741

[6] Brunel, E., Comte, F., Guilloux, A., 2013. Nonparametric estimation for survival data with censoring indicators missing at random. Journal of Statistical Planning and Inference 143(10), 1653-1671. MR3082225

[7] Carroll, R.J., Ruppert, D., Stefanski, L.A.,Crainiceanu, C.M., 2006. Measurement Error in Nonlinear Models: A Modern Perspective (2nd ed.). Chapman and Hall/CRC. MR2243417

[8] Caudill, S. B., Mixon, F. G., 1995. Modeling household fertility decisions: Estimation and testing of censored regression models for count data. Empirical Economics 20(2), 183-196.

[9] Chen, X., Cai, J., 2018. Reweighted estimators for additive hazard model with censoring indicators missing at random. Lifetime Data Analysis 24(2), 224-249. MR3773858

[10] Dupuy, J.-F., Leconte, E., 2009. A study of regression calibration in a partially observed stratified Cox model. Journal of Statistical Planning and Inference 139(2), 317-328. MR2474008

[11] Famoye, F., Wang, W., 2004. Censored generalized Poisson regression model. Computational Statistics & Data Analysis 46(3), 547-560. MR2067036

[12] Faria, S., Soromenho, G., 2012. Comparison of EM and SEM Algorithms in Poisson Regression Models: A Simulation Study. Communications in Statistics – Simulation and Computation 41(4), 497-509. MR2869000

[13] Foutz, R.V., 1977. On the unique consistent solution to the likelihood equations. Journal of the American Statistical Association 72, 147-148. MR0445686

[14] Guo, X., Niu, C., Yang, Y., Xu, W., 2015. Empirical likelihood for single index model with missing covariates at random. Statistics 49(3), 588-601. MR3349080

[15] Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56(4), 1030-1039. MR1815581

[16] Hardin, J.W., Schmiediche, H., Carroll, R.J., 2003. The regression-calibration method for fitting generalized linear models with additive measurement error. The Stata Journal 3(4), 361-372.

[17] Horton, N.J., Lipsitz, S.R., 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician 55(3), 244-254. MR1963401

[18] Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47(260), 663-685. MR0053460

[19] Huang, S.Y.H., 2005. Regression calibration using response variables in linear models. Statistica Sinica 15, 685-696. MR2233906

[20] Hsu, C.H., Yu, M., 2019. Cox regression analysis with missing covariates via nonparametric multiple imputation. Statistical Methods in Medical Research 28(6), 1676-1688. MR3961958

[21] Ibrahim J.G., Chen M.-H., Lipsitz, S.R., Herring, A.H., 2005. Missing-data methods for generalized linear models: a comparative review. Journal of the American Statistical Association 100(469), 332-346. MR2166072

[22] Karlis, D., Papatla, P., Roy, S., 2016. Finite mixtures of censored Poisson regression models. Statistica Neerlandica 70(2), 100-122. MR3488660

[23] Kleinke, K., Reinecke, J., 2013. Multiple imputation of incomplete zero-inflated count data. Statistica Neerlandica 67(3), 311-336. MR3083213
[24] Van Der Laan, M.J., McKeague, I.W., 1998. Efficient estimation from right-censored data when failure indicators are missing at random. Annals of Statistics 26(1), 164-182. MR1611792
[25] Liao, X., Zucker, D.M., Li, Y., Spiegelman, D., 2011. Survival analysis with error-prone time-varying covariates: a risk set calibration approach. Biometrics 67(1), 50-58. MR2898816
[26] Mahmoud, M. M., Alderiny, M. M., 2010. On estimating parameters of censored generalized Poisson regression model. Applied Mathematical Sciences 4(13-16), 623-635. MR2595502
[27] McKeague, I.W., Subramanian, S., 1998. Product-limit estimators and Cox regression with missing censoring information. Scandinavian Journal of Statistics 25(4), 589-601. MR1666792
[28] Nevo, D., Nishihara, R., Ogino, S., Wang, M., 2018. The competing risks Cox model with auxiliary case covariates under weaker missing-at-random cause of failure. Lifetime Data Analysis 24(3), 425-442. MR3814704
[29] Nguyen, V. T., Dupuy, J.-F., 2021. Asymptotic results in censored zero-inflated Poisson regression. Communications in Statistics – Theory and Methods 50(12), 2759-2779. MR4268179
[30] Pollard, J., Kirk, S., Cade, J., 2002. Factors affecting food choice in relation to fruit and vegetable intake: A review. Nutrition Research Reviews 15(2), 373-387.
[31] R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria, https://www.R-project.org/.
[32] Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89(427), 846-866. MR1294730
[33] Saffari, S. E., Adnan, R., 2011. Zero-inflated Poisson regression models with right censored count data. Matematika 27(1), 21-29. MR2842673
[34] Seaman, S.R., White, I.R., 2013. Review of inverse probability weighting for dealing with missing data. Statistical methods in medical research 22(3), 278-295. MR3190658
[35] Steingrimsson, J.A., Strawderman, R.L., 2017. Estimation in the semiparametric accelerated failure time model with missing covariates: improving efficiency through augmentation. Journal of the American Statistical Association 112(519), 1221-1235. MR3735372
[36] Subramanian, S., 2006. Survival analysis for the missing censoring indicator model using kernel density estimation techniques. Statistical methodology 3(2), 125-136. MR2227416
[37] Subramanian, S., 2011. Multiple imputations and the missing censoring indicator model. Journal of Multivariate Analysis 102(1), 105-117. MR2729423
[38] Sun, Y., Qian, X., Shou, Q., Gilbert, P.B., 2017. Analysis of two-phase sampling data with semiparametric additive hazards models. Lifetime

Data Analysis 23(3), 377-399. MR3660317
[39] Terza, J.V., 1985. A Tobit-type estimator for the censored Poisson regression model. Economics Letters 18(4), 361-365.
[40] Tsiatis, A.A., 2007. Semiparametric Theory and Missing Data. Springer, New York. MR2233926
[41] van der Vaart, A.W., 2000. Asymptotic Statistics. Cambridge University Press. MR1652247
[42] Wang, C.Y., Chen, H.Y., 2001. Augmented inverse probability weighted estimator for Cox missing covariate regression. Biometrics 57(2), 414-419. MR1855674
[43] Wang, C.Y., Hsu, L., Feng, Z.D., Prentice, R.L., 1997. Regression calibration in failure time regression. Biometrics 53(1), 131-145. MR1450183
[44] Wang, Q., Dinse, G.E., 2011. Linear regression analysis of survival data with missing censoring indicators. Lifetime data analysis 17(2), 256-279. MR2777120
[45] Wang, Q., Shen, J., 2008. Estimation and confidence bands of a conditional survival function with censoring indicators missing at random. Journal of Multivariate Analysis 99(5), 928-948. MR2405099
[46] Wang, Q., Dinse, G.E., Liu, C., 2012. Hazard function estimation with cause-of-death data missing at random. Annals of the Institute of Statistical Mathematics 64(2), 415-438. MR2878913
[47] Weller E.A., Milton D.K., Eisen E.A., Spiegelman D., 2007. Regression calibration for logistic regression with multiple surrogates for one exposure. Journal of Statistical Planning and Inference 137(2), 449-461. MR2298949
[48] White, H., 1982. Maximum Likelihood Estimation of Misspecified Models. Econometrica 50(1), 1-25. MR0640163
[49] White, I., Royston, P., 2009. Imputing missing covariate values for the Cox model. Statistics in Medicine 28, 1982-1998. MR2750806
[50] Xie, F.-C., Wei, B.-C., 2007. Diagnostics analysis in censored generalized Poisson regression model. Journal of Statistical Computation and Simulation 77(8), 695-708. MR2407649
[51] Zheng, M., Lin, R., Yu, W., 2016. Competing risks data analysis under the accelerated failure time model with missing cause of failure. Annals of the Institute of Statistical Mathematics 68(4), 855-876. MR3520046
[52] Zou, Y., Fan, G., Zhang, R., 2020. Quantile regression and variable selection for partially linear single-index models with missing censoring indicators. Journal of Statistical Planning and Inference 204, 80-95. MR3961931