Copula-based regression models with data missing at random: A unified approach

Shigeyuki Hamori* Kaiji Motegi† Zheng Zhang‡

September 15, 2019

Abstract

The existing literature of copula-based regression models typically focuses on either conditional mean or quantile regression, and assumes complete data. This paper unifies the conditional mean and quantile regressions as well as other interesting regressions by formulating a general loss function which may not be continuously differentiable. Further, we relax the assumption of complete data by allowing the regressand and regressors to be missing at random (MAR). A semiparametric copula and the target regression curve are estimated via the calibration approach. The consistency and asymptotic normality of the estimated regression curve are proved. We show via Monte Carlo simulations that the proposed approach has sharp performance in finite samples, while a benchmark equal-weight approach fails with substantial bias under MAR. An empirical application on revenues and R&D expenses of U.S. manufacturing firms highlights a practical use of our approach.

AMS subject classifications: primary, 62G08; secondary, 62H12.

Keywords: calibration estimation; generalized regression model; missing at random; semiparametric copula.

*Graduate School of Economics, Kobe University. E-mail: hamori@econ.kobe-u.ac.jp

†Graduate School of Economics, Kobe University. E-mail: motegi@econ.kobe-u.ac.jp

‡*Corresponding author*. Institute of Statistics and Big Data, Renmin University of China. Address: Haidian District, Beijing 100872 China. E-mail: zhengzhang@ruc.edu.cn

1 Introduction

Regression is the most prevailing method for investigating the relationship between a regressand Y and regressors $\boldsymbol{W}$. Widely used regressions include conditional mean and quantile regressions. Noh, El Ghouch, and Bouezmarni (2013) proposed a novel approach to estimate a conditional mean regression function by exploiting copulas, where observations are assumed to be independently and identically distributed (*i.i.d.*) and completely observed. Their key insight is that the loss function expressed as a *conditional* expectation given regressors $\boldsymbol{W}$ can be rewritten as an *unconditional* expectation involving a parametric copula and nonparametric marginal distributions. The marginal distributions and the copula parameter are estimated via plug-in methods. The flexibility of the semiparametric copula modelling alleviates model specification issues such as how to transform regressors and which cross-products of regressors to include.[1]

The existing literature of the copula-based regression leaves two issues, and this paper resolves both of them. First, there is need for unified regression modelling. Each previous paper focuses on either conditional mean or quantile regression, leaving the theoretical relationship between them unclear. Other regressions such as asymmetric least squares of Newey and Powell (1987) should also be incorporated. This paper unifies all those regressions—with a particular emphasis on the mean regression of Noh, El Ghouch, and Bouezmarni (2013) and the quantile regression of Noh, El Ghouch, and Van Keilegom (2015)—by formulating a general loss function which may not be continuously differentiable. We derive asymptotic theory under the unified framework, a contribution which enhances the systematic interpretation of various regressions.

The second issue left in the literature is how to perform the copula-based regression analysis when some observations are missing. In Noh, El Ghouch, and Bouezmarni (2013) and Noh, El Ghouch, and Van Keilegom (2015), copula is simply a tool for estimating the regression curve flexibly. There also exists the vast literature where copula itself is a primary target of estimation (see, e.g., Genest, Ghoudi, and Rivest, 1995, Chen and Fan, 2005, 2006). In either case, most papers involving copula as-

[1] Noh, El Ghouch, and Van Keilegom (2015) applied the method of Noh, El Ghouch, and Bouezmarni (2013) to the quantile regression with *i.i.d.* or time series data that are completely observed. De Backer, El Ghouch, and Van Keilegom (2017) extended the method of Noh, El Ghouch, and Van Keilegom (2015) to the quantile regression with censored data. Kraus and Czado (2017) studied the quantile regression with complete data, using D-vine copulas. Rémillard, Nasri, and Bouezmarni (2017) discussed the asymptotic connection between the estimators of Noh, El Ghouch, and Van Keilegom (2015) and Kraus and Czado (2017).

sume complete data. The assumption of complete data is restrictive and unrealistic in many fields of research. In survey analysis, for example, respondents may refuse to report their personal information such as age and salary.

To relax the rather restrictive assumption of complete data, this paper allows both the regressand Y and the regressors $\boldsymbol{W}$ to be *missing at random (MAR)*, a key concept originally explored by Rubin (1976). Note that the MAR condition is more general than the *missing completely at random (MCAR)* condition. The MCAR condition requires that $\{Y, \boldsymbol{W}\}$ and their missing status $\boldsymbol{T}$ should be unconditionally independent of each other, whereas the MAR condition requires that they should be conditionally independent given covariates $\boldsymbol{X}$. The MAR condition has been popularly used in econometrics and statistics to identify the parameter of interest (see, e.g., Robins and Rotnitzky, 1995, Chen, Hong, and Tarozzi, 2008, Ding and Song, 2016, Hamori, Motegi, and Zhang, 2019, Delaigle, Huang, and Lei, 2019).

The most naïve way to handle missing data is *listwise deletion*, which discards individuals with incomplete data and assigns equal weights on individuals with complete data. The listwise deletion delivers consistent inference when data are MCAR, but may deliver inconsistent inference when data are MAR. Since the MCAR condition is restrictive in many applications, the listwise deletion is a risky approach that can cause serious bias.

Hamori, Motegi, and Zhang (2019) is one of few works to deal with data missing at random in copula modelling.[2] They use calibration weights proposed by Chan, Yam, and Zhang (2016) for both nonparametric marginal distributions and target copula parameters. The calibration estimation is a nonparametric method that balances the empirical moments of covariates between the observed and whole groups. It does not require an explicit specification of the missing mechanism, and delivers consistent inference under MAR.

Inspired by Hamori, Motegi, and Zhang (2019), this paper adopts the calibration estimation to perform the copula-based regression with $\{Y, \boldsymbol{W}\}$ missing at random. A semiparametric copula and the target regression curve are estimated via the calibration approach. The consistency and asymptotic normality of the estimated regression curve are proved. Our simulation study shows that the proposed approach performs well in finite samples, while a benchmark equal-weight approach fails with substan-

[2] Ding and Song (2016) proposed an EM algorithm for estimating the Gaussian copula under the MAR condition. Emura, Lin, and Wang (2010), Emura and Wang (2010) and Emura and Wang (2012) considered the copula inference with survival data.

tial bias under MAR. To illustrate a practical value of the proposed approach, an empirical application on revenues and R&D expenses of U.S. manufacturing firms is presented. The calibration approach detects a positive correlation between revenues and R&D for any firm size, while the equal-weight approach yields mixed results for large firms.

The remainder of this paper is organized as follows. Section 2 explains our basic framework and notation. Our estimator is proposed in Section 3, and its large sample properties are derived in Section 4. In Section 5, the simulation study is performed. In Section 6, the empirical application is presented. Brief conclusions are provided in Section 7. Mathematical details are collected in Technical Appendices.

2 Basic framework and notation

Let Y be a regressand, and let $\boldsymbol{W} = (W_1, \ldots, W_d)^\top$ be d-dimensional regressors. Consider the generalized regression problem:

$$a_0(\boldsymbol{w}) = \arg\min_{a \in \mathbb{R}} \mathrm{E}\left[L(g(Y) - a) | \boldsymbol{W} = \boldsymbol{w}\right], \tag{2.1}$$

where $\boldsymbol{w} = (w_1, \ldots, w_d)^\top$; $L(\cdot)$ is a pre-specified loss function whose derivative, denoted by $L'(\cdot)$, exists almost everywhere; $g(Y)$ is a known function of Y. We do not require $L(\cdot)$ to be continuously differentiable. The formulation (2.1) includes many prominent cases:

- $L(v) = v^2$ and $g(Y) = Y$, in which case $a_0(\boldsymbol{w}) = \mathrm{E}[Y|\boldsymbol{W} = \boldsymbol{w}]$ is the conditional mean regression studied by Noh, El Ghouch, and Bouezmarni (2013).
- $L(v) = v(\tau - \mathbf{1}(v < 0))$ and $g(Y) = Y$, in which case $a_0(\boldsymbol{w})$ is the τ^{th} conditional quantile studied by Noh, El Ghouch, and Van Keilegom (2015).
- $L(v) = v^2$ and $g(Y) = \mathbf{1}(Y \leq y)$, in which case $a_0(\boldsymbol{w}) = \mathrm{E}[\mathbf{1}(Y \leq y)|\boldsymbol{W} = \boldsymbol{w}] = \Pr(Y \leq y|\boldsymbol{W} = \boldsymbol{w})$ is the conditional distribution function.
- $L(v) = v^2|\tau - \mathbf{1}(v \leq 0)|$ and $g(Y) = Y$, in which case $a_0(\boldsymbol{w})$ corresponds to the asymmetric least squares studied by Newey and Powell (1987).

This paper considers the generalized regression problem (2.1) with both Y and $\boldsymbol{W}$ being possibly missing. Let $\boldsymbol{O}_{\text{obs}} \subset \boldsymbol{O} := \{Y, \boldsymbol{W}\}$ be a set of components which are observed with probability 1. Let $\boldsymbol{O}_{\text{mis}} := \boldsymbol{O} \setminus \boldsymbol{O}_{\text{obs}}$ be a set of components

which are missing with positive probability. Let d_{obs} (resp. d_{mis}) be the number of elements in $\boldsymbol{O}_{\text{obs}}$ (resp. $\boldsymbol{O}_{\text{mis}}$), then $d+1 = d_{\text{obs}} + d_{\text{mis}}$ by construction. Let $\boldsymbol{T}_i := (T_{1i}, \ldots, T_{d_{\text{mis}},i})^\top \in \{0,1\}^{d_{\text{mis}}}$ be binary indicators which represent the missing status of $\boldsymbol{O}_{i,\text{mis}} = (O_{1i,\text{mis}}, \ldots, O_{d_{\text{mis}},i,\text{mis}})^\top$, namely, $T_{ji} = 0$ (resp. $T_{ji} = 1$) if $O_{ji,\text{mis}}$ is missing (resp. observed) for $j \in \{1, \ldots, d_{\text{mis}}\}$ and $i \in \{1, \ldots, N\}$.

If $\boldsymbol{T}_i$ and $\boldsymbol{O}_{i,\text{mis}}$ are independent of each other, then the latter is called *missing completely at random (MCAR)*. Under the MCAR condition, an elementary approach of *listwise deletion*, which merely picks individuals with complete observations and puts equal weights on them, is well known to deliver consistent inference. The MCAR condition, however, is a strong assumption that is violated in many applications.

In this paper we impose a more realistic assumption called *missing at random (MAR)*. Let $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{ri})^\top$ be r-dimensional *covariates* that are observable for all individuals $i \in \{1, \ldots, N\}$, where $\boldsymbol{X}_i \supset \boldsymbol{O}_{i,\text{obs}}$ and hence $r \geq d_{\text{obs}}$. The MAR condition is that $\boldsymbol{T}_i$ and $\boldsymbol{O}_{i,\text{mis}}$ are conditionally independent of each other given covariates $\boldsymbol{X}_i$.

Assumption 1 (missing at random)**.** $\boldsymbol{T}_i \perp \boldsymbol{O}_{i,\text{mis}} \,|\, \boldsymbol{X}_i$.

The MAR condition is popularly used in econometrics and statistics to identify the parameter of interest. The MAR condition does not require the unconditional independence between $\boldsymbol{T}_i$ and $\boldsymbol{O}_{i,\text{mis}}$. In many applications, $\boldsymbol{T}_i$ and $\boldsymbol{O}_{i,\text{mis}}$ are correlated with each other through $\boldsymbol{X}_i$, and that violates MCAR but not MAR.

To simplify notation without losing generality, we assume hereafter that $\boldsymbol{O}_{i,\text{obs}} = \emptyset$ and $\boldsymbol{O}_{i,\text{mis}} = \boldsymbol{O}_i = \{Y_i, \boldsymbol{W}_i\}$ (i.e., any component of the regressand and regressors is missing with positive probability). Then $d_{\text{obs}} = 0$, $d_{\text{mis}} = d+1$, and $0 < \Pr(T_{ji} = 1) < 1$ for all $j \in \{0, 1, \ldots, d\}$, where T_{0i} indicates the missing status of Y_i and T_{ji} with $j \in \{1, \ldots, d\}$ indicates the missing status of W_{ji}. Assume further that the observations $\{\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{W}_i, Y_i\}_{i=1}^N$ are *i.i.d.*

Let $f_{Y,\boldsymbol{W}}(\cdot)$ and $f_{\boldsymbol{W}}(\cdot)$ be the joint density functions of $\{Y, \boldsymbol{W}\}$ and $\boldsymbol{W}$, respectively. Let $F_0(\cdot)$ and $F_j(\cdot)$ be the cumulative distribution functions of Y and W_j for $j = 1, \ldots, d$, respectively. Let $f_0(\cdot)$ and $f_j(\cdot)$ be the probability density functions of Y and W_j, respectively. Let $c(\cdot)$ and $c_{\boldsymbol{W}}(\cdot)$ be the copula densities of $\{Y, \boldsymbol{W}\}$ and $\boldsymbol{W}$, respectively. Sklar's Theorem ensures that $f_{Y,\boldsymbol{W}}(y, \boldsymbol{w}) = c(F_0(y), F_1(w_1), \ldots, F_d(w_d)) \cdot f_0(y) \cdot \prod_{j=1}^d f_j(w_j)$ and $f_{\boldsymbol{W}}(\boldsymbol{w}) = c_{\boldsymbol{W}}(F_1(w_1), \ldots, F_d(w_d)) \cdot \prod_{j=1}^d f_j(w_j)$. Define the *propensity score functions* as

$$\pi_j(\boldsymbol{x}) := \Pr(T_{ji} = 1 \,|\, \boldsymbol{X}_i = \boldsymbol{x}), \quad j \in \{0, 1, \ldots, d\}, \tag{2.2}$$

$$\eta(\boldsymbol{x}) := \Pr(T_{0i} = T_{1i} = \cdots = T_{di} = 1 \,|\, \boldsymbol{X}_i = \boldsymbol{x}). \tag{2.3}$$

Using Sklar's Theorem, the MAR condition, (2.2), and (2.3), $a_0(\boldsymbol{w})$ can be identified as follows:

$$\begin{aligned}
&a_0(\boldsymbol{w}) = \arg\min_{a\in\mathbb{R}} \mathrm{E}\left[L(g(Y)-a)|\boldsymbol{W}=\boldsymbol{w}\right] \\
&= \arg\min_{a\in\mathbb{R}} \int L(g(y)-a) f_{Y|\boldsymbol{W}}(y|\boldsymbol{w})dy = \arg\min_{a\in\mathbb{R}} \int L(g(y)-a)\frac{f_{Y,\boldsymbol{W}}(y,\boldsymbol{w})}{f_{\boldsymbol{W}}(\boldsymbol{w})}dy \\
&= \arg\min_{a\in\mathbb{R}} \int L(g(y)-a)\frac{c(F_0(y),F_1(w_1),\ldots,F_d(w_d))\cdot f_0(y)\cdot\prod_{j=1}^d f_j(w_j)}{c_{\boldsymbol{W}}(F_1(w_1),\ldots,F_d(w_d))\cdot\prod_{j=1}^d f_j(w_j)}dy \\
&= \arg\min_{a\in\mathbb{R}} \int L(g(y)-a)c(F_0(y),F_1(w_1),\ldots,F_d(w_d))\cdot f_0(y)dy \\
&= \arg\min_{a\in\mathbb{R}} \mathrm{E}\left[L(g(Y)-a)c(F_0(Y),F_1(w_1),...,F_d(w_d))\right] \qquad (2.4) \\
&= \arg\min_{a\in\mathbb{R}} \mathrm{E}\left[\mathrm{E}\left[L(g(Y)-a)c(F_0(Y),F_1(w_1),...,F_d(w_d))|\boldsymbol{X}\right]\right] \\
&= \arg\min_{a\in\mathbb{R}} \mathrm{E}\left[\mathrm{E}\left[\frac{T_0}{\pi_0(\boldsymbol{X})}\middle|\boldsymbol{X}\right]\cdot\mathrm{E}\left[L(g(Y)-a)c(F_0(Y),F_1(w_1),...,F_d(w_d))|\boldsymbol{X}\right]\right] \\
&= \arg\min_{a\in\mathbb{R}} \mathrm{E}\left[\mathrm{E}\left[\frac{T_0}{\pi_0(\boldsymbol{X})}\cdot L(g(Y)-a)c(F_0(Y),F_1(w_1),...,F_d(w_d))\middle|\boldsymbol{X}\right]\right] \\
&= \arg\min_{a\in\mathbb{R}} \mathrm{E}\left[\frac{T_0}{\pi_0(\boldsymbol{X})}\cdot L(g(Y)-a)c(F_0(Y),F_1(w_1),...,F_d(w_d))\right]. \qquad (2.5)
\end{aligned}$$

Note that $a_0(\boldsymbol{w})$ is expressed as the *conditional* expectation given $\boldsymbol{W}=\boldsymbol{w}$ in (2.1), while it is expressed as the *unconditional* expectation involving the copula density in (2.4). Compare (2.4) and (2.5) to see how missing data are handled. The objective function is associated with the *whole* group in (2.4), while it is associated with the *observed* group with respect to Y in (2.5). They coincide *if and only if* individuals in the observed group are weighted by $\pi_0(\boldsymbol{X})^{-1}$, a core insight of the inverse probability weighting (Horvitz and Thompson, 1952).

3 Calibration weighting estimation

Assume that the copula density of $\{Y, \boldsymbol{W}\}$ admits a parametric model $c(u_0,u_1,\ldots,u_d) = c(u_0,u_1,\ldots,u_d;\boldsymbol{\theta}_0)$. Assume further that $\boldsymbol{\theta}_0$ is identified as the maximizer of the log-likelihood: $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathrm{E}\left[\ln c(F_0(Y_i),F_1(W_{1i}),\ldots,F_d(W_{di});\boldsymbol{\theta})\right]$, where Θ is a compact subset of $\mathbb{R}^p$ which contains the true value $\boldsymbol{\theta}_0$. Using Assumption 1 and the

law of iterated expectations, $\boldsymbol{\theta}_0$ can be expressed as

$$\begin{aligned}
\boldsymbol{\theta}_0 &= \arg\max_{\boldsymbol{\theta}\in\Theta} \mathrm{E}\left[\ln c(F_0(Y_i), F_1(W_{1i}), \ldots, F_d(W_{di}); \boldsymbol{\theta})\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta} \mathrm{E}\left[\ln c(F_0(Y_i), F_1(W_{1i}), \ldots, F_d(W_{di}); \boldsymbol{\theta}) \cdot \mathrm{E}\left[\frac{\mathbf{1}(T_{0i} = \cdots = T_{di} = 1)}{\eta(\boldsymbol{X})}\middle| \boldsymbol{X}\right]\right] \\
&= \arg\max_{\boldsymbol{\theta}\in\Theta} \mathrm{E}\left[\frac{\mathbf{1}(T_{0i} = \cdots = T_{di} = 1)}{\eta(\boldsymbol{X}_i)} \ln c(F_0(Y_i), F_1(W_{1i}), \ldots, F_d(W_{di}); \boldsymbol{\theta})\right]. \qquad (3.1)
\end{aligned}$$

A sample counterpart to (3.1) is given by

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{N} \frac{\mathbf{1}(T_{0i} = \cdots = T_{di} = 1)}{N \cdot \eta(\boldsymbol{X}_i)} \ln c(F_0(Y_i), F_1(W_{1i}), \ldots, F_d(W_{di}); \boldsymbol{\theta}). \qquad (3.2)$$

$\tilde{\boldsymbol{\theta}}$ is an infeasible estimator of $\boldsymbol{\theta}_0$ since $\eta(\boldsymbol{X}_i)$ and $\{F_j\}_{j=0}^d$ are all unknown. We thus need to replace them with feasible estimators in order to estimate $\boldsymbol{\theta}_0$. First, rewrite $F_0(y)$ in the same way as (3.1):

$$\begin{aligned}
F_0(y) =& \mathrm{E}\left[\mathbf{1}(Y_i \le y)\right] = \mathrm{E}[\mathrm{E}\left[\mathbf{1}(Y_i \le y)|\boldsymbol{X}_i\right]] = \mathrm{E}\left[\mathrm{E}\left[\mathbf{1}(Y_i \le y)|\boldsymbol{X}_i\right] \cdot \mathrm{E}\left[\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}\middle|\boldsymbol{X}_i\right]\right] \\
=& \mathrm{E}\left[\mathrm{E}\left[\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}\mathbf{1}(Y_i \le y)\middle|\boldsymbol{X}_i\right]\right] = \mathrm{E}\left[\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}\mathbf{1}(Y_i \le y)\right]. \qquad (3.3)
\end{aligned}$$

Similarly, $F_j(w)$ with $j \in \{1, ..., d\}$ can be reritten as follows:

$$F_j(y) = \mathrm{E}\left[\mathbf{1}(W_{ji} \le y)\right] = \mathrm{E}\left[\frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)}\mathbf{1}(W_{ji} \le w)\right]. \qquad (3.4)$$

The sample counterparts to (3.3) and (3.4) are respectively given by

$$\tilde{F}_0(y) = \sum_{i=1}^{N} T_{0i}\left\{\frac{1}{N \cdot \pi_0(\boldsymbol{X}_i)}\right\}\mathbf{1}(Y_i \le y), \qquad (3.5)$$

$$\tilde{F}_j(w) = \sum_{i=1}^{N} T_{ji}\left\{\frac{1}{N \cdot \pi_j(\boldsymbol{X}_i)}\right\}\mathbf{1}(W_{ji} \le w),\ j \in \{1, ..., d\}. \qquad (3.6)$$

$\{\tilde{F}_j\}_{j=0}^d$ are infeasible estimators of $\{F_j\}_{j=0}^d$ since $\{\pi_j(\boldsymbol{X}_i)\}_{j=0}^d$ are unknown.

Equations (3.2), (3.5), and (3.6) motivate the estimation of $\{N \cdot \pi_j(\boldsymbol{X}_i)\}^{-1}$ and $\{N \cdot \eta(\boldsymbol{X}_i)\}^{-1}$. A naïve approach of directly estimating $\pi_j(\boldsymbol{X}_i)$ and $\eta(\boldsymbol{X}_i)$ and substituting $\{N \cdot \hat{\pi}_j(\boldsymbol{X}_i)\}^{-1}$ and $\{N \cdot \hat{\eta}(\boldsymbol{X}_i)\}^{-1}$ often perform poorly in practice. Parametric modelling of $\pi_j(\boldsymbol{X}_i)$ and $\eta(\boldsymbol{X}_i)$ may suffer from misspecification. Nonparametric esti-

mation such as kernel smoothing is unstable in finite samples; $\hat{\pi}_j(\boldsymbol{X}_i)$ and $\hat{\eta}(\boldsymbol{X}_i)$ can take extremely small values which destroy the entire computation. Indeed, the inverse probability weighting estimators are notoriously sensitive to the estimated propensity score (Kang and Schafer, 2007).

Covariate balancing is a useful approach which prevents the occurrence of extreme weights. This paper adopts the covariate balancing principle of Chan, Yam, and Zhang (2016) and Hamori, Motegi, and Zhang (2019) to perform a one-shot estimation of $\{N \cdot \pi_j(\boldsymbol{X}_i)\}^{-1}$ and $\{N \cdot \eta(\boldsymbol{X}_i)\}^{-1}$. Their key insight is that the following equation holds for any integrable function $u(\boldsymbol{X})$ and $j \in \{0, 1, ..., d\}$:

$$\mathrm{E}\left[T_{ji}\left\{\frac{1}{\pi_j(\boldsymbol{X}_i)}\right\}u(\boldsymbol{X}_i)\right] = \mathrm{E}[u(\boldsymbol{X}_i)], \tag{3.7}$$

$$\mathrm{E}\left[\mathbf{1}(T_{0i} = \cdots = T_{di} = 1)\left\{\frac{1}{\eta(\boldsymbol{X}_i)}\right\}u(\boldsymbol{X}_i)\right] = \mathrm{E}[u(\boldsymbol{X}_i)]. \tag{3.8}$$

Hence, the estimator of $\{N \cdot \pi_j(\boldsymbol{X})\}^{-1}$, denoted by $\hat{p}_{jK}(\boldsymbol{X})$, and the estimator of $\{N \cdot \eta(\boldsymbol{X})\}^{-1}$, denoted by $\hat{q}_K(\boldsymbol{X})$, should satisfy the sample counterparts of (3.7) and (3.8):

$$\sum_{i=1}^{N} T_{ji}\hat{p}_{jK}(\boldsymbol{X}_i)u_K(\boldsymbol{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i), \tag{3.9}$$

$$\sum_{i=1}^{N} \mathbf{1}(T_{0i} = \cdots = T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)u_K(\boldsymbol{X}_i) = \frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i), \tag{3.10}$$

where $u_K(\boldsymbol{X}) = (u_{K,1}(\boldsymbol{X}), ..., u_{K,K}(\boldsymbol{X}))^\top$ is a known sieve basis function that can approximate any suitable function $u(\boldsymbol{X})$ arbitrarily well, and $K \to \infty$ as $N \to \infty$. Common sieve basis functions include power series, splines, and wavelets. One can allow the dimension K to vary across $\{\hat{p}_{0,K_0}(\boldsymbol{X}), \ldots, \hat{p}_{d,K_d}(\boldsymbol{X}), \hat{q}_{K_q}(\boldsymbol{X})\}$ without any theoretical difficulty. For the sake of notational brevity, we use a single value K for all components hereafter.

Multiple values of $\hat{p}_{jK}(\boldsymbol{X}_i)$ and $\hat{q}_K(\boldsymbol{X}_i)$ satisfy (3.9) and (3.10), respectively. Among them, the calibration approach chooses the one closest to a uniform weight given some distance measure in order to enhance stable performance in finite samples. As shown in Hamori, Motegi, and Zhang (2019), the resulting estimator of $\{N \cdot \pi_j(\boldsymbol{X})\}^{-1}$ is

$$\hat{p}_{jK}(\boldsymbol{X}) = \frac{1}{N} \cdot \rho'\left\{\hat{\boldsymbol{\lambda}}_{jK}^\top u_K(\boldsymbol{X})\right\}, \tag{3.11}$$

where $\rho(\cdot)$ is a strictly concave function on $\mathbb{R}$, and $\hat{\boldsymbol{\lambda}}_{jK} \in \mathbb{R}^K$ maximizes the following concave objective function:

$$\hat{G}_{jK}(\boldsymbol{\lambda}) = \frac{1}{N}\sum_{i=1}^{N} T_{ji}\rho\{\boldsymbol{\lambda}^\top u_K(\boldsymbol{X}_i)\} - \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\lambda}^\top u_K(\boldsymbol{X}_i).$$

The first order condition implies (3.11) holds. Similarly, the estimator of $\{N\cdot\eta(\boldsymbol{X})\}^{-1}$ is defined by

$$\hat{q}_K(\boldsymbol{X}_i) = \frac{1}{N}\rho'\{\hat{\boldsymbol{\beta}}_K^\top u_K(\boldsymbol{X}_i)\} \tag{3.12}$$

for any i such that $T_{0i} = T_{1i} = \cdots = T_{di} = 1$, where $\hat{\boldsymbol{\beta}}_K$ maximizes the following concave objective function:

$$\hat{H}_K(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(T_{0i} = \cdots = T_{di} = 1)\rho\{\boldsymbol{\beta}^\top u_K(\boldsymbol{X}_i)\} - \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\beta}^\top u_K(\boldsymbol{X}_i).$$

$\hat{p}_{jK}(\boldsymbol{X})$ (resp. $\hat{q}_K(\boldsymbol{X})$) can be interpreted as a generalized empirical likelihood estimator of $\{N\cdot\pi_j(\boldsymbol{X})\}^{-1}$ (resp. $\{N\cdot\eta(\boldsymbol{X})\}^{-1}$). See Hamori, Motegi, and Zhang (2019) for a detailed discussion.

The $\rho(v)$ function can be any increasing and strictly concave function. Prominent examples include $\rho(v) = -\exp(-v)$ for the exponential tilting (Kitamura and Stutzer, 1997, Imbens, Spady, and Johnson, 1998); $\rho(v) = \ln(1+v)$ for the empirical likelihood; $\rho(v) = -(1-v)^2/2$ for the continuous updating of the generalized method of moments; $\rho(v) = v - \exp(-v)$ for the inverse logistic.

Use (3.11) in (3.5) and (3.6) to estimate the marginal distributions $\{F_j\}_{j=0}^d$:

$$\hat{F}_{0,K}(y) := \sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)\mathbf{1}(Y_i \le y), \tag{3.13}$$

$$\hat{F}_{j,K}(w) = \sum_{i=1}^{N} T_{ji}\hat{p}_{jK}(\boldsymbol{X}_i)\mathbf{1}(W_{ji} \le w), \quad j \in \{1,\ldots,d\}. \tag{3.14}$$

Use (3.12), (3.13), and (3.14) in (3.2) to estimate $\boldsymbol{\theta}_0$:

$$\hat{\boldsymbol{\theta}}_K := \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{N} \mathbf{1}(T_{0i} = \cdots = T_{di} = 1)\hat{q}_K(\boldsymbol{X}_i)\ln c\left(\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(W_{1i}), \ldots, \hat{F}_{d,K}(W_{di}); \boldsymbol{\theta}\right).$$

Finally, the target $a_0(\boldsymbol{w})$ in (2.5) is estimated via

$$\hat{a}(\boldsymbol{w}) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L(g(Y_i)-a)c\left(\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \ldots, \hat{F}_{d,K}(w_d); \hat{\boldsymbol{\theta}}_K\right). \tag{3.15}$$

Remark 1. *Under the conditional mean regression of Noh, El Ghouch, and Bouezmarni (2013), $L(g(Y_i)-a) = (Y_i-a)^2$ and hence $\hat{a}(\boldsymbol{w})$ has a closed form solution:*

$$\hat{a}(\boldsymbol{w}) = \frac{\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)Y_i c\left(\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \ldots, \hat{F}_{d,K}(w_d); \hat{\boldsymbol{\theta}}_K\right)}{\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)c\left(\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \ldots, \hat{F}_{d,K}(w_d); \hat{\boldsymbol{\theta}}_K\right)}. \tag{3.16}$$

4 Large sample properties

In this section, we derive large sample properties of the proposed estimator (3.15). Let $\|\cdot\|$ be the Frobenius norm defined by $\|\boldsymbol{A}\| := \mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^\top)^{1/2}$, where $\boldsymbol{A}$ is a real matrix. For any $K \in \mathbb{N}$, let $\zeta(K) := \sup_{\boldsymbol{x}\in\mathcal{X}} \|u_K(\boldsymbol{x})\|$ be the supremum norm of approximation sieves $u_K(\boldsymbol{x})$. In general, this bound depends on the array of basis used.[3] For any function $f(v_0, v_1, ..., v_d; \boldsymbol{\theta})$, define the following derivatives:

$$\begin{aligned}
&\partial_{\boldsymbol{\theta}} f(v_0, v_1, \ldots, v_d; \boldsymbol{\theta}) := (\partial/\partial\boldsymbol{\theta}) f(v_0, v_1, \ldots, v_d; \boldsymbol{\theta}),\\
&\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}} f(v_0, v_1, \ldots, v_d; \boldsymbol{\theta}) := (\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top) f(v_0, v_1, \ldots, v_d; \boldsymbol{\theta}),\\
&\partial_j f(v_0, v_1, \ldots, v_d; \boldsymbol{\theta}) := (\partial/\partial v_j) f(v_0, v_1, \ldots, v_d; \boldsymbol{\theta}) \quad \text{for } j \in \{0, 1, ..., d\},\\
&\partial^2_{\boldsymbol{\theta} j} f(v_0, v_1, \ldots, v_d; \boldsymbol{\theta}) := (\partial^2/\partial\boldsymbol{\theta}\partial v_j) f(v_0, v_1, \ldots, v_d; \boldsymbol{\theta}) \quad \text{for } j \in \{0, 1, ..., d\}.
\end{aligned}$$

Define $\ell(v_0, v_1, \ldots, v_d; \boldsymbol{\theta}) := \ln c(v_0, v_1, \ldots, v_d; \boldsymbol{\theta})$, $U_{0i} := F_0(Y_i)$, $U_{ji} := F_j(W_{ji})$ for $j \in \{1, \ldots, d\}$, and $\boldsymbol{U}_i := (U_{0i}, U_{1i}, \ldots, U_{di})^\top$.

The following conditions, which are also imposed in Hamori, Motegi, and Zhang (2019), are needed to establish the asymptotic normality of the estimated copula parameter $\hat{\boldsymbol{\theta}}_K$:

Assumption 2. *The support of the covariate $\boldsymbol{X}$, which is denoted by $\mathcal{X}$, is a Cartesian product of r compact intervals.*

[3] Newey (1997) shows that $\zeta(K) \leq CK$ for orthonormal polynomials, and $\zeta(K) \leq C\sqrt{K}$ for B-splines, where $C > 0$ is a universal positive constant.

Assumption 3. *The smallest eigenvalue of* $\mathrm{E}\left[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top\right]$ *is bounded away from zero uniformly in* K*.*

Assumption 4. *The inverse propensity scores* $\pi(\boldsymbol{x})^{-1}$ *and* $\eta^{-1}(\boldsymbol{X})$ *are bounded above, i.e., there exists some constant* $\delta < \infty$ *such that* $1 \leq \pi(\boldsymbol{x})^{-1} \leq \delta$ *and* $1 \leq \eta^{-1}(\boldsymbol{X}) \leq \delta$ *for any* $\boldsymbol{x} \in \mathcal{X}$*.*

Assumption 5. *There exist* $\boldsymbol{\lambda}_{jK}$*,* $\boldsymbol{\beta}_K \in \mathbb{R}^K$ *and* $\alpha > 0$ *such that* $\sup_{x\in\mathcal{X}} |(\rho')^{-1}(1/\pi_j(\boldsymbol{x})) - \boldsymbol{\lambda}_{jK}^\top u_K(\boldsymbol{x})| = O(K^{-\alpha})$ *and* $\sup_{x\in\mathcal{X}} |(\rho')^{-1}(1/\eta(\boldsymbol{x})) - \boldsymbol{\beta}_K^\top u_K(\boldsymbol{x})| = O(K^{-\alpha})$ *as* $K \to \infty$*.*

Assumption 6. $\zeta(K)^2K^4/N \to 0$ *and* $\sqrt{N}K^{-\alpha} \to 0$*.*

Assumption 7. $\rho(\cdot)$ *is a strictly concave function defined on* $\mathbb{R}$ *and three times continuously differentiable, and the range of* ρ' *contains* $[1, \delta]$*.*

Assumption 8. $\mathrm{E}[\partial_{\boldsymbol{\theta}}\ell(U_{0i}, U_{1i}, \ldots, U_{di}; \boldsymbol{\theta})|\boldsymbol{X}_i = \boldsymbol{x}]$ *is continuously differentiable in* $\boldsymbol{x}$*.*

Assumption 9. $\boldsymbol{B} := -\mathrm{E}\left[\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(U_{0i}, U_{1i}, \ldots, U_{di}; \boldsymbol{\theta}_0)\right]$ *and* $\boldsymbol{\Sigma} := \mathrm{Var}\{\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \boldsymbol{\theta}_0) + \sum_{j=0}^d R_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \boldsymbol{\theta}_0)\}$ *are finite and positive definite, where*

$$\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \boldsymbol{\theta}_0) := \frac{\mathbf{1}(T_{0i} = \cdots = T_{di} = 1)}{\eta(\boldsymbol{X}_i)}\partial_{\boldsymbol{\theta}}\ell(U_{0i}, U_{1i}, \ldots, U_{di}; \boldsymbol{\theta}_0) - \mathrm{E}[\partial_{\boldsymbol{\theta}}\ell(U_{0i}, U_{1i}, \ldots, U_{di}; \boldsymbol{\theta}_0)]$$
$$- \left\{\frac{\mathbf{1}(T_{0i} = \cdots = T_{di} = 1)}{\eta(\boldsymbol{X}_i)} - 1\right\}\mathrm{E}[\partial_{\boldsymbol{\theta}}\ell(U_{0i}, U_{1i}, \ldots, U_{di}; \boldsymbol{\theta}_0)|\boldsymbol{X}_i],$$

$$R_j(T_{ji}, \boldsymbol{X}_i, U_{0i}; \boldsymbol{\theta}_0) := \mathrm{E}\left[\partial^2_{\boldsymbol{\theta}\boldsymbol{j}}\ell(U_{0s}, U_{1s}, \ldots, U_{ds}; \boldsymbol{\theta}_0)\{\phi_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; U_{js}) - U_{js}\}|U_{ji}, \boldsymbol{X}_i, T_{ji}\right], \quad (s \neq i),$$

where $\phi_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; v) := \dfrac{T_{ji}}{\pi_j(\boldsymbol{X}_i)}\mathbf{1}(U_{ji} \leq v) - \left\{\dfrac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} - 1\right\}\mathrm{E}[\mathbf{1}(U_{ji} \leq v)|\boldsymbol{X}_i],\ v \in [0, 1].$

Assumption 10. *(i) For each* $(u_0, u_1, \ldots, u_d) \in (0, 1)^{d+1}$*,* $\partial^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(u_0, u_1, \ldots, u_d; \boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$ *in a neighborhood of* $\boldsymbol{\theta}_0$*. (ii)* $\mathrm{E}[\sup_{\boldsymbol{\theta}\in\Theta:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|=o(1)} \|\ell_{\boldsymbol{\theta}\boldsymbol{\theta}}(U_{0i}, U_{1i}, \ldots, U_{di}; \boldsymbol{\theta})\|] < \infty$*.*

Assumption 11. *For* $j \in \{0, 1, \ldots, d\}$*,* $\partial_{\boldsymbol{\theta}j}\ell(u_0, u_1, \ldots, u_d; \boldsymbol{\theta}_0)$ *is well defined and continuous in* $(u_0, u_1, \ldots, u_d) \in (0, 1)^{d+1}$*. Furthermore,*

(i) $\|\partial_{\boldsymbol{\theta}}\ell(u_0, u_1, \ldots, u_d; \boldsymbol{\theta}_0)\| \leq$ *constant* $\times \prod_{j=0}^d \{u_j(1-u_j)\}^{-a_j}$ *for some* $a_j \geq 0$ *such that* $\mathrm{E}[\prod_{j=0}^d \{U_{ji}(1-U_{ji})\}^{-2a_j}] < \infty$*;*

(ii) $\|\partial^2_{\boldsymbol{\theta}k}\ell(u_0, u_1, \ldots, u_d; \boldsymbol{\theta}_0)\| \leq$ *constant* $\times \{u_k(1-u_k)\}^{-b_k} \prod_{j=0, j\neq k}^d \{u_j(1-u_j)\}^{-a_j}$ *for some* $b_k > a_k$ *such that* $\mathrm{E}[\{U_{ki}(1-U_{ki})\}^{\xi_k - b_k} \prod_{j=0, j\neq k}^d \{U_{ji}(1-U_{ji})\}^{-a_j}] < \infty$ *for some* $\xi_k \in (0, 1/2)$*.*

Assumption 2 restricts the covariates to be bounded. This condition is restrictive but commonly imposed in the nonparametric regression literature, since it simplifies the derivation of the convergence rate under the L^∞ norm.[4] Assumption 3, which is also imposed in Newey (1997), essentially requires the sieve basis functions to be orthogonal. Assumption 4, a common condition in the missing data literature, ensures that a sufficient portion of marginal data are observed. Assumption 5 requires the sieve approximation errors of $\rho'^{-1}(\pi_j(\boldsymbol{x})^{-1})$ and $\rho'^{-1}(\eta(\boldsymbol{x})^{-1})$ to shrink at a polynomial rate. This condition is satisfied for a variety of sieve basis functions (Newey, 1997). If $\boldsymbol{X}$ is discrete, then the approximation error is zero for sufficiently large K, satisfying Assumption 5 with $\alpha = +\infty$. If $\boldsymbol{X}$ are continuous, the polynomial rate depends positively on the smoothness of $\rho'^{-1}(\pi_j(\boldsymbol{x})^{-1})$ and $\rho'^{-1}(\eta(\boldsymbol{x})^{-1})$ in the continuous components and negatively on the number of the continuous components; indeed, for power series and B-splines, $\alpha = -s/r$, where s is the smoothness of approximand and r is the dimension of $\boldsymbol{X}$. Hence, we admit that the proposed method suffers from the curse of dimensionality, a common challenge in nonparametric estimation. The extension to high dimensional covariates is beyond the scope of this paper and will be pursued in the future work. We will show that the convergence rate of the estimated weight function is bounded by this polynomial rate. Assumption 6, another common assumption in nonparametric regression, restricts the smoothing parameter to balance the bias and variance. Assumption 7 is a mild restriction on ρ which is satisfied in all important cases considered in the literature. Assumption 8 controls the approximation error. Assumption 9 guarantees the finiteness of the asymptotic variance. Assumption 10 guarantees the uniform convergence. Assumption 11 allows the score function and its partial derivatives with respect to the first d arguments to blow up at the boundaries, which occurs for many popular copula functions such as Gaussian, Clayton, and t-copulas.

Proposition 1. *Under Assumptions 1-11, we have that*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{\eta}_i + o_p(1),$$

where $\boldsymbol{\eta}_i = \boldsymbol{B}^{-1}\{\varphi(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{U}_i; \boldsymbol{\theta}_0) + \sum_{j=0}^{d} R_j(T_{ji}, \boldsymbol{X}_i, U_{ji}; \boldsymbol{\theta}_0)\}$.

Proposition 1 is a restatement of Hamori, Motegi, and Zhang (2019, Theorem 5).

The following conditions are needed to establish the consistency of $\hat{a}(\boldsymbol{w})$.

[4] Assumption 2 can be relaxed if we restrict the tail distribution of $\boldsymbol{X}$.

Assumption 12. *(i) The parameter space $\mathcal{A} \subset \mathbb{R}$ is a compact set and the true parameter $a_0(\boldsymbol{w})$ lies in the interior of $\mathcal{A}$; (ii)* $\mathrm{E}[\sup_{a\in\mathcal{A}} |L(g(Y)-a)|^2] < \infty$.

Condition (i) is often imposed in the regression literature. Condition (ii) is an envelope condition that is sufficient for the applicability of the uniform law of large numbers.

Theorem 1. *Under Assumptions 1-12, for each fixed $\boldsymbol{w}$, $\hat{a}(\boldsymbol{w}) \xrightarrow{p} a_0(\boldsymbol{w})$.*

A proof of Theorem 1 is presented in Appendix A.

We next establish the asymptotic normality of the proposed estimator. To handle a potentially non-smooth loss function, the following conditions are required.

Assumption 13. *(i) The loss function $L(v)$ is differentiable almost everywhere; (ii)* $\mathrm{E}\left[L'(g(Y)-a)c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right]$ *is differentiable with respect to a, and the derivative is nonzero; (iii)* $\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-\hat{a}(\boldsymbol{w}))c(\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \ldots, \hat{F}_{d,K}(w_d); \hat{\boldsymbol{\theta}}_K) = o_p(N^{-1/2})$.

Condition (i) is a mild condition since $L'(v)$ is not required to be continuous. All example loss functions presented in Section 2 satisfy Condition (i). Condition (ii) guarantees the asymptotic variance to be finite. Condition (iii) is essentially the asymptotic first order condition, similar to that used in Z-estimation. This first order condition is satisfied by popular non-smooth loss functions (see Pakes and Pollard, 1989).

The following theorem establishes the asymptotic normality of our proposed estimator.

Theorem 2. *Under Assumptions 1-13, we have*

$$\sqrt{N}\{\hat{a}(\boldsymbol{w}) - a_0(\boldsymbol{w})\} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} S(T_i, \boldsymbol{X}_i, Y_i; \boldsymbol{w}) + o_p(1),$$

which implies that $\sqrt{N}\{\hat{a}(\cdot) - a_0(\cdot)\}$ converges weakly to a Gaussian process with covariance function $\Omega(\boldsymbol{w}_1, \boldsymbol{w}_2) = \mathrm{E}\left[S(T, \boldsymbol{X}, Y; \boldsymbol{w}_1)S(T, \boldsymbol{X}, Y; \boldsymbol{w}_2)\right]$ for $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$, where

$$S(T_i, \boldsymbol{X}_i, Y_i; \boldsymbol{w}) = \frac{1}{b(\boldsymbol{w})} \times (A_{1i} + A_{2i} + A_{3i}),$$

$$b(\boldsymbol{w}) = -\partial_a \mathrm{E}\left[L'(g(Y)-a_0)c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right],$$

$$A_{1i} = \frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} L'(g(Y_i)-a_0)c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)$$

$$
\begin{aligned}
&\quad - \left(\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} - 1\right) \mathrm{E}\left[L'(g(Y_i) - a_0)c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) | \boldsymbol{X}_i\right], \\
A_{2i} &= \int L'(g(y) - a_0)\partial_0 c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad \cdot \left[\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}\mathbf{1}(Y_i \leq y) - \left(\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} - 1\right) F_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) - F_0(y)\right] dF_0(y), \\
A_{3i} &= \sum_{j=1}^{d} \mathrm{E}\left[L'(g(Y) - a_0)\partial_j c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right] \\
&\quad \cdot \left\{\frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)}\mathbf{1}(W_{ji} \leq w_j) - \left(\frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} - 1\right) F_{W_j|\boldsymbol{X}}(w_j|\boldsymbol{x}) - F_j(w_j)\right\} \\
&\quad + \boldsymbol{\eta}_i^{\top} \mathrm{E}\left[L'(g(Y) - a_0)\partial_{\boldsymbol{\theta}} c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right].
\end{aligned}
$$

A proof of Theorem 2 is presented in Appendix B.

Remark 2.

1. *Theorem 2 ensures that our proposed estimator satisfies $\sqrt{N}$-normality under the unified framework which covers non-smooth $L(\cdot)$.*

2. *Theorem 2 contains some important results in the existing literature as special cases. When the mean regression with complete data is considered, the influence function $S(T_i, \boldsymbol{X}_i, Y_i; \boldsymbol{w})$ reduces to that of Noh, El Ghouch, and Bouezmarni (2013). When the quantile regression with complete data is considered, $S(T_i, \boldsymbol{X}_i, Y_i; \boldsymbol{w})$ reduces to the influence function of Noh, El Ghouch, and Van Keilegom (2015). See Appendices C-D for derivations.*

5 Monte Carlo simulations

In this section, Monte Carlo simulations are conducted to evaluate the finite sample performance of the calibration estimation. Section 5.1 covers a benchmark scenario which has one regressor ($d = 1$) and one covariate ($r = 1$). Section 5.2 presents an extended scenario which has one regressor ($d = 1$) and two covariates ($r = 2$). Section 5.3 presents another extended scenario which has two regressors ($d = 2$) and one covariate ($r = 1$).

5.1 Benchmark scenario

Let $\boldsymbol{Z}_i = (Z_{1i}, Z_{2i}, Z_{3i})^\top$, and draw $\boldsymbol{Z}_i$ independently and identically from two well-known copulas:

- trivariate Clayton copula with parameter $\theta_0 = 1.333$;
- trivariate Gumbel copula with parameter $\theta_0 = 1.667$.

For both cases, the implied Kendall's tau is $\tau = 0.4$, a moderate level of association. Define the regressand $Y_i = \Phi^{-1}(Z_{1i})$, regressor $W_i = \Phi^{-1}(Z_{2i})$, and covariate $X_i = \Phi^{-1}(Z_{3i})$, where $\Phi^{-1}(\cdot)$ is the inverse distribution function of $N(0,1)$. Assume that W_i and X_i are observed for all individuals $i \in \{1, \ldots, N\}$, where the sample size is $N \in \{250, 500, 750\}$. Let T_i be a binary indicator which equals 1 if Y_i is observed and 0 if Y_i is missing. The regressand Y_i may be missing with the propensity score function:

$$\Pr(T_i = 1 \,|\, X_i = x) = \frac{1}{1 + \exp(b_0 + b_1 x)}. \tag{5.1}$$

The logistic function is commonly used to specify the propensity score function in the literature of missing data analysis (see, e.g., Qin, Leung, and Shao, 2002). Suppose that $(b_0, b_1) = (-0.57, 1.5)$, in which case $\mathrm{E}(T_i) = 0.6$ and Y_i is MAR.[5]

Following Noh, El Ghouch, and Bouezmarni (2013), the conditional mean regression is considered here.[6] In this case, the estimated regression curve $\hat{a}(w)$ has the closed form solution (3.16). Hence, the computation of $\hat{a}(w)$ is straightforward once the weights $\hat{p}_K(X_i)$ are computed.

We perform the calibration estimation with the exponential tilting $\rho(v) = -\exp(-v)$ and the following sieve basis function in order to compute $\hat{p}_K(X_i)$:

$$u_K(X) = (1, X, \ldots, X^{K-1})^\top, \qquad K \in \{2, 3, 4\}. \tag{5.2}$$

For comparison, the equal weight $\hat{p}_K(X_i) = 1/\sum_{i=1}^N T_i$ is also considered. The equal-weight approach should fail under the MAR mechanism (5.1), since by construction

[5] In an extra simulation not reported here, we also considered $(b_0, b_1) = (-0.405, 0)$, in which case $\mathrm{E}(T_i) = 0.6$ and Y_i is MCAR. Under MCAR, the calibration estimation and the benchmark approach of assigning equal weights on all individuals with complete data perform as well as each other. Since this result is not surprising, the present paper focuses on the MAR case for the sake of brevity.

[6] In an extra simulation not reported here, we also considered the conditional 25-percentile and median regressions in accordance with Noh, El Ghouch, and Van Keilegom (2015). Simulation results of the quantile regressions is qualitatively similar to those of the mean regression, and hence the present paper focuses on the mean regression for the sake of brevity.

it ignores the impact of X on the propensity score.

In the present simulation, a model misspecification is not discussed. When the underlying copula is Clayton (Gumbel), we fit the Clayton (Gumbel) copula to estimate θ_0 and then estimate the target function $a_0(w)$.

Draw $J = 1000$ Monte Carlo samples and compute integrated root mean squared errors (IRMSEs) as follows. First, the RMSE in the j^{th} sample is defined as

$$RMSE_j = \sqrt{\frac{1}{\#\mathcal{W}} \sum_{w \in \mathcal{W}} \{\hat{a}_j(w) - a_0(w)\}^2}, \tag{5.3}$$

where $\hat{a}_j(w)$ is the estimated target function; $\mathcal{W}$ is the set of w's considered; $\#\mathcal{W}$ is the number of w's considered. Since the marginal distribution of the regressor W_i is $N(0,1)$, the range $w \in [-3,3]$ should be covered in order to properly evaluate the performance of each estimator. Dividing this interval more finely would make the evaluation more accurate, but it would raise computational burden. To balance the evaluation accuracy and computational speed, we use $\mathcal{W} = \{-3.00, -2.95, \ldots, 2.95, 3.00\}$ and hence $\#\mathcal{W} = 121$. $RMSE_j$ in (5.3) contains the true value $a_0(w)$. Recall that

$$a_0(w) = \arg\min_{a \in \mathbb{R}} \mathrm{E}[L\{g(Y) - a\} c\{F_0(Y), F_1(w); \theta_0\}],$$

which can be approximated by

$$a_0(w) \simeq \arg\min_{a \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} L\{g(Y_i) - a\} c\{F_0(Y_i), F_1(w); \theta_0\}.$$

Since $L\{g(Y) - a\} = (Y - a)^2$ in the present study, $a_0(w)$ is simply given by

$$a_0(w) = \frac{\sum_{i=1}^{N} Y_i c\{F_0(Y_i), F_1(w); \theta_0\}}{\sum_{i=1}^{N} c\{F_0(Y_i), F_1(w); \theta_0\}}. \tag{5.4}$$

Substitute (5.4) into (5.3) to compute $RMSE_j$. Finally, the IRMSE is defined as

$$IRMSE = \frac{1}{J} \sum_{j=1}^{J} RMSE_j, \qquad J = 1000.$$

See Table 1 for the resulting IRMSEs. The calibration estimation performs well for both copulas considered. Besides, the selection of tuning parameter $K \in \{2,3,4\}$ has a relatively small impact on IRMSE, which is a practical advantage. As expected,

IRMSE shrinks as sample size N grows. The equal-weight approach always leads to larger IRMSEs than the calibration approach with any K considered. Under the Clayton copula with $N = 500$, for example, the IRMSE is $\{0.142, 0.124, 0.114\}$ for the calibration approach with $K \in \{2, 3, 4\}$ respectively, while it is 0.249 for the equal-weight approach. Thus, the calibration approach is a desired approach that dominates the equal-weight approach.

Table 1: IRMSE of $\hat{a}(w)$ under the benchmark scenario

	Clayton copula			Gumbel copula		
	$N = 250$	$N = 500$	$N = 750$	$N = 250$	$N = 500$	$N = 750$
Calibration ($K = 2$)	0.171	0.142	0.130	0.290	0.229	0.197
Calibration ($K = 3$)	0.156	0.124	0.107	0.256	0.178	0.145
Calibration ($K = 4$)	0.175	0.114	0.094	0.286	0.177	0.143
Equal weight	0.270	0.249	0.241	0.427	0.384	0.370

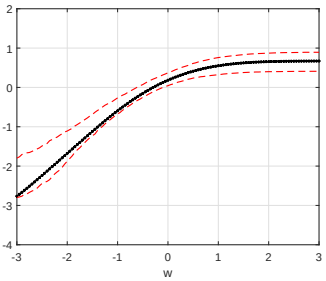
For the calibration estimator, the sieve basis function is constructed as $u_K(X) = (1, X, \ldots, X^{K-1})^\top$. "Equal weight" signifies the benchmark approach of assigning equal weights for all individuals. The regressand Y_i is missing at random (MAR). The integrated root mean squared error (IRMSE) is computed across the grid $w \in \{-3.00, -2.95, \ldots, 2.95, 3.00\}$ and $J = 1000$ Monte Carlo samples.

In Figures 1-2, we plot the true $a_0(w)$ with a black, solid line and point-wise 95% confidence bands $CB(w) = [\ell(w), u(w)]$ with red, dashed lines, where $\ell(w)$ and $u(w)$ are the lower and upper 2.5-percentiles of $\{\hat{a}_1(w), \ldots, \hat{a}_J(w)\}$, respectively. $N = 250$ in Figure 1, while $N = 750$ in Figure 2. Figures with $N = 500$ are omitted to conserve space, since their appearance is logically an intermediate case between Figures 1 and 2.

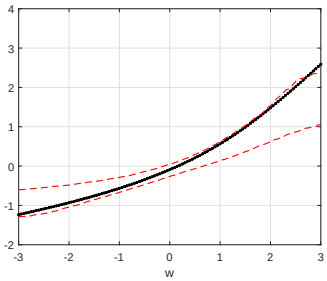
For the Clayton copula, the calibration estimation with any $K \in \{2, 3, 4\}$ leads to narrow enough confidence bands which contain $a_0(w)$ in the middle. Naturally, the larger sample size implies the even narrower confidence bands. When $N = 750$, $\ell(w)$ and $u(w)$ almost coincide with $a_0(w)$ for any $w > -2$, which highlights the strikingly sharp performance of the calibration estimation. The same implications hold for the Gumbel copula as far as $w < 1.5$ is concerned. There tends to be a negative bias for $w > 1.5$, which is not a surprising result since few observations are available in that region.

The equal-weight approach, by contrast, leads to substantial bias for almost any $w \in [-3, 3]$ (see the bottom panels of Figures 1-2). This is the source of the large IRMSEs of the equal-weight approach observed in Table 1. For the Clayton copula,

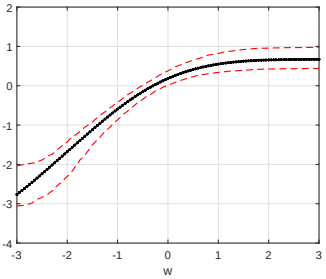
Figure 1: True $a_0(w)$ and 95% confidence bands (benchmark scenario; $N = 250$)



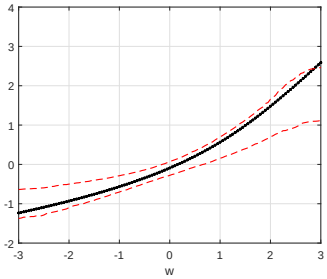
1. CAL ($K = 2$), Clayton



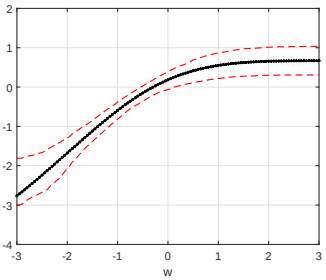
2. CAL ($K = 2$), Gumbel



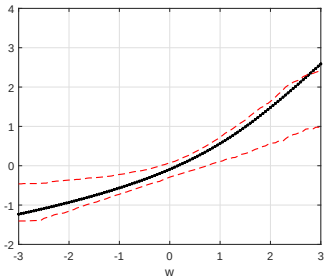
3. CAL ($K = 3$), Clayton



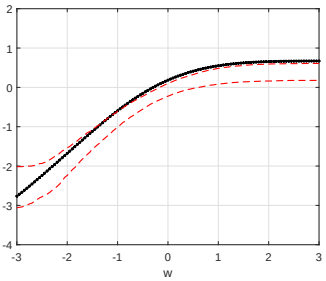
4. CAL ($K = 3$), Gumbel



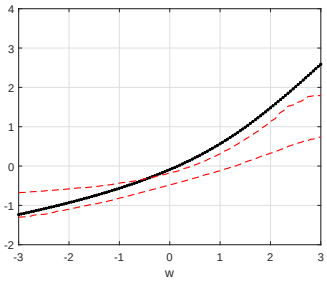
5. CAL ($K = 4$), Clayton



6. CAL ($K = 4$), Gumbel



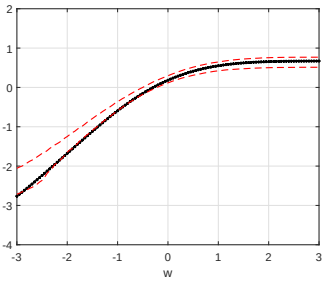
7. Equal, Clayton



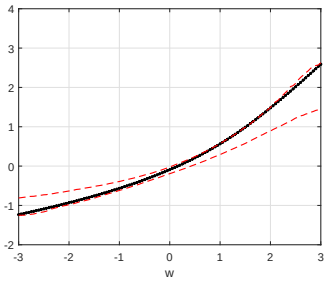
8. Equal, Gumbel

there is a clear downward bias for $w \in [-2, 3]$. For the Gumbel copula, there is a critical downward bias for $w > 0$. Those biases do not vanish as the sample size

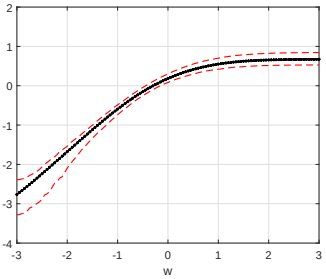
Figure 2: True $a_0(w)$ and 95% confidence bands (benchmark scenario; $N = 750$)



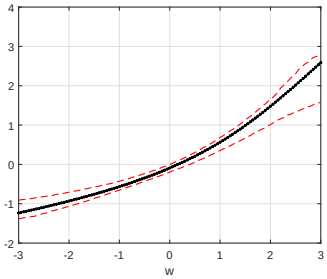
1. CAL ($K = 2$), Clayton



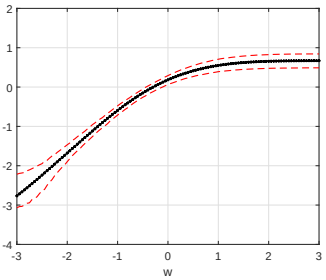
2. CAL ($K = 2$), Gumbel



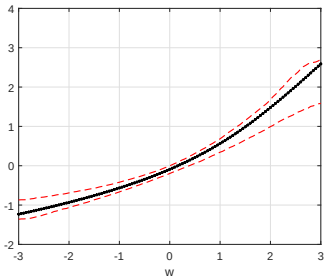
3. CAL ($K = 3$), Clayton



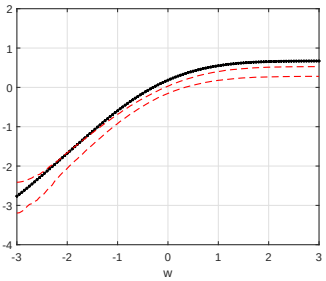
4. CAL ($K = 3$), Gumbel



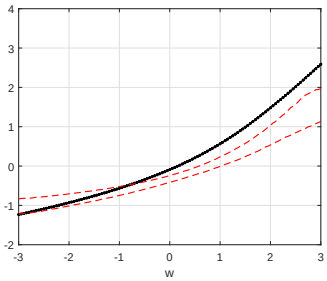
5. CAL ($K = 4$), Clayton



6. CAL ($K = 4$), Gumbel



7. Equal, Clayton



8. Equal, Gumbel

increases, a strong evidence that the equal-weight approach fails when the regressand is MAR.

In summary, the simulation results highlight the advantage of the calibration approach in finite samples. Under MAR, the calibration approach delivers consistent inference while the benchmark equal-weight approach delivers inconsistent inference.

5.2 Extended scenario I: Two covariates

In this section, we extend the benchmark scenario by adding another covariate ($r = 2$). Let $\boldsymbol{Z}_i = (Z_{1i}, Z_{2i}, Z_{3i}, Z_{4i})^\top$, and draw $\boldsymbol{Z}_i$ independently and identically from

- four-variable Clayton copula with parameter $\theta_0 = 1.442$;
- four-variable Gumbel copula with parameter $\theta_0 = 1.719$.

For both cases, the implied Kendall's tau is $\tau = 0.4$. The marginal distributions of $(Y_i, W_i, X_{1i}, X_{2i})$ are all $N(0,1)$ as in the benchmark scenario. Assume that W_i and $\boldsymbol{X}_i = (X_{1i}, X_{2i})^\top$ are observed for all individuals. The propensity score function is

$$\Pr(T_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}) = \frac{1}{1 + \exp(-0.53 + 0.75x_1 + 0.75x_2)}.$$

This specification implies that $\mathrm{E}(T_i) = 0.6$ and Y_i is MAR. For the calibration estimation, the following sieve basis functions are used:

$$\begin{aligned}
u_3(\boldsymbol{X}) &= (1, X_1, X_2)^\top, \\
u_6(\boldsymbol{X}) &= (1, X_1, X_2, X_1^2, X_2^2, X_1X_2)^\top, \\
u_{10}(\boldsymbol{X}) &= (1, X_1, X_2, X_1^2, X_2^2, X_1X_2, X_1^3, X_2^3, X_1^2X_2, X_1X_2^2)^\top.
\end{aligned}$$

See Table 2 for IRMSEs after $J = 1000$ Monte Carlo iterations. The calibration estimation with any $K \in \{3, 6, 10\}$ performs well for both copulas, and the performance becomes sharper as the sample size becomes larger. The equal-weight approach always leads to larger IRMSEs than the calibration approach with any $K \in \{3, 6, 10\}$. Under the Clayton copula with $N = 500$, for example, the IRMSE is $\{0.121, 0.120, 0.116\}$ for the calibration approach with $K \in \{3, 6, 10\}$ respectively, while it is 0.270 for the equal-weight approach. Thus, as in the benchmark scenario, the calibration approach dominates the equal-weight approach.[7]

Under the Gumbel copula, the performance of the calibration approach is more sensitive to K than in the benchmark scenario. When $N = 500$, the IRMSE is

[7] Figures of confidence bands are omitted, since they are qualitatively similar to Figures 1-2.

Table 2: IRMSE of $\hat{a}(w)$ under the extended scenario I (two covariates)

	Clayton copula			Gumbel copula		
	$N = 250$	$N = 500$	$N = 750$	$N = 250$	$N = 500$	$N = 750$
Calibration ($K = 3$)	0.153	0.121	0.106	0.272	0.204	0.181
Calibration ($K = 6$)	0.157	0.120	0.105	0.232	0.157	0.129
Calibration ($K = 10$)	0.162	0.116	0.100	0.296	0.235	0.222
Equal weight	0.288	0.270	0.264	0.433	0.392	0.371

For the calibration estimator, the sieve basis function is constructed as $u_3(\boldsymbol{X}) = (1, X_1, X_2)^\top$, $u_6(\boldsymbol{X}) = (1, X_1, X_2, X_1^2, X_2^2, X_1X_2)^\top$, or $u_{10}(\boldsymbol{X}) = (1, X_1, X_2, X_1^2, X_2^2, X_1X_2, X_1^3, X_2^3, X_1^2X_2, X_1X_2^2)^\top$. "Equal weight" signifies the benchmark approach of assigning equal weights for all individuals. The regressand Y_i is missing at random (MAR). The integrated root mean squared error (IRMSE) is computed across the grid $w \in \{-3.00, -2.95, \ldots, 2.95, 3.00\}$ and $J = 1000$ Monte Carlo samples.

$\{0.204, 0.157, 0.235\}$ for $K \in \{3, 6, 10\}$, respectively. It suggests that there is a greater need of a data-driven selection of K as the dimension of covariates increases.

5.3 Extended scenario II: Two regressors

In this section, we extend the benchmark scenario by adding another regressor ($d = 2$). The copulas used here are the same as those in the extended scenario I. The marginal distributions of $(Y_i, W_{1i}, W_{2i}, X_i)$ are all $N(0, 1)$ as in the previous scenarios. The propensity score is specified as in (5.1). For the calibration estimator, the sieve basis function is constructed as in (5.2).

See Table 3 for IRMSEs after $J = 1000$ Monte Carlo iterations. Since there are two regressors, the IRMSEs are computed over a two-dimensional grid $(w_1, w_2) \in \{-3.0, -2.5, \ldots, 2.5, 3.0\} \times \{-3.0, -2.5, \ldots, 2.5, 3.0\}$. Hence, the number of (w_1, w_2) considered is $\#\mathcal{W} = 169$. The results in Table 3 are all consistent with the previous results in Tables 1-2. The calibration estimation with any $K \in \{2, 3, 4\}$ performs well for both copulas, and their IRMSE is always smaller than the IRMSE of the equal-weight approach. Under the Clayton copula with $N = 500$, for example, the IRMSE is $\{0.197, 0.152, 0.142\}$ for the calibration approach with $K \in \{2, 3, 4\}$ respectively, while it is 0.247 for the equal-weight approach. The performance of the calibration approach is more sensitive to the choice of K than in the benchmark scenario. It suggests that there is a greater need of a data-driven selection of K as the dimension of regressors increases.

In summary, the calibration estimation performs well for any scenario considered,

Table 3: IRMSE of $\hat{a}(\boldsymbol{w})$ under the extended scenario II (two regressors)

	Clayton copula			Gumbel copula		
	$N = 250$	$N = 500$	$N = 750$	$N = 250$	$N = 500$	$N = 750$
Calibration ($K = 2$)	0.228	0.197	0.186	0.195	0.149	0.134
Calibration ($K = 3$)	0.202	0.152	0.136	0.165	0.121	0.101
Calibration ($K = 4$)	0.210	0.142	0.117	0.183	0.124	0.100
Equal weight	0.281	0.247	0.240	0.259	0.225	0.214

For the calibration estimator, the sieve basis function is constructed as $u_K(X) = (1, X, \ldots, X^{K-1})^\top$. "Equal weight" signifies the benchmark approach of assigning equal weights for all individuals. The regressand Y_i is missing at random (MAR). The integrated root mean squared error (IRMSE) is computed across the grid $(w_1, w_2) \in \{-3.0, -2.5, \ldots, 2.5, 3.0\} \times \{-3.0, -2.5, \ldots, 2.5, 3.0\}$ and $J = 1000$ Monte Carlo samples.

while the equal-weight approach fails with serious bias under the MAR mechanism.[8]

6 Empirical applications

In this section, we analyze the relationship between research and development (R&D) expenses and revenues of manufacturing firms in the U.S. R&D is a key element of manufacturing, and we use revenues as a proxy of the firm size. Intuitively, the larger manufacturing firm should have the larger R&D expense, hence the two variables should be positively correlated. A goal of this study is to check if that is indeed the case in the U.S.

6.1 Data and methodology

Let the regressor W_i be the operating revenue (turnover) of firm i measured in billions of U.S. dollars. Let the regressand Y_i be the R&D expense of firm i in billions of USD. In practice, many firms report their operating revenues but not R&D expenses, while few firms report their R&D expenses but not revenues. It is therefore reasonable to assume that the revenue W is always observed while the R&D expense Y is possibly missing.

Assume that the covariate is identical to the regressor: $X_i = W_i$. Intuitively, whether a firm reports its R&D expense should depend on the stringency of the

[8] As in the extended scenario I, figures of confidence bands are omitted since they are qualitatively similar to Figures 1-2.

accounting requirement that the firm is facing. The larger firm should meet the more stringent accounting rule in order to keep its social credibility. Hence, the magnitude of revenues is supposed to have a positive impact on the probability of reporting R&D.

All data used in this study are retrieved via Orbis maintained by Bureau van Dijk. There are 1914 firms whose (i) country ISO code is US, (ii) US SIC codes are 2000-3999 (Manufacturing), and (iii) operating revenues in 2017 are observed. The bottom 914 firms are discarded, since their revenues are negligibly small. The top 10 firms are also discarded, since their revenues and some of their R&D expenses are exceptionally large. The remaining $N = 990$ firms are retained for analysis. R&D expenses are observed for 701 of the 990 firms and missing for the other 289. The missing probability of R&D is as high as 29.2%, motivating the use of the calibration estimation.

As in the simulation study, the exponential tilting function $\rho(v) = -\exp(-v)$ is used to compute calibration weights. The sieve basis function is specified as $u_K(X) = (1, X, X^2)^\top$ with $K = 3$. Empirical implications do not alter when $K = 2$ or $K = 4$ is used. The equal-weight approach is also considered for comparison. For both approaches, the conditional mean regression of Noh, El Ghouch, and Bouezmarni (2013) with the Gumbel copula is performed. In this specific study, the Gumbel copula fits the data better than the Clayton copula.

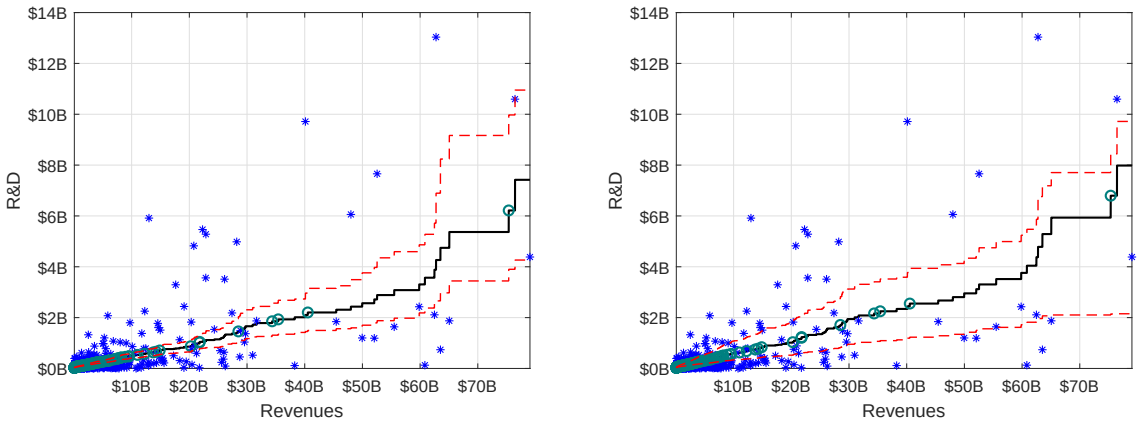
Since the conditional mean regression is considered, the regression curve $a_0(w)$ can be estimated via (3.16). The grid is specified as $w \in \{0.07, 0.08, \ldots, 79.04\}$, where 0.07 and 79.04 are the minimum and maximum of revenues, respectively. Confidence bands of $\hat{a}(w)$ are constructed via resampling, taking advantage of the *i.i.d.* assumption. Draw $\{i_1, \ldots, i_N\}$ independently and identically from the discrete uniform distribution on $\{1, \ldots, N\}$. Given the resampled dataset $\{Y_{i_j}, W_{i_j}, X_{i_j}, T_{i_j}\}_{j=1}^N$, estimate a regression curve and call it $\hat{a}_b(w)$. Repeat $B = 1000$ times to get $\{\hat{a}_b(w)\}_{b=1}^B$. For each $w \in \{0.07, \ldots, 79.04\}$, the 95% confidence band of $\hat{a}(w)$ is given by the lower and upper 2.5 percentiles of $\{\hat{a}_b(w)\}_{b=1}^B$.

6.2 Empirical results

See Figure 3 for empirical results. The blue asterisks represent the 701 firms whose revenues and R&D are both observed. The black, solid line represents the regression curve $\hat{a}(w)$. As expected, $\hat{a}(w)$ is positively sloped, suggesting a positive correlation between revenues and R&D. The green circles represent the 289 firms with unobserved R&D expenses, and those values are imputed with the conditional expectation given

W_i. By construction, all green circles lie on $\hat{a}(w)$. As expected, the vast majority of the 289 firms with unobserved R&D have relatively small revenues. The 95% confidence bands based on resampling are plotted as the red, dashed lines. The confidence bands become wider as the revenue w increases, reflecting the fact that there are fewer firms on the upper tail.

Figure 3: Empirical results



Calibration approach | Equal-weight approach

The calibration and equal-weight approaches produce similar regression curves $\hat{a}(w)$, but their confidence bands are clearly different. For $w < 40$, where most firms are present, the confidence bands for the equal-weight approach are roughly twice as wide as the confidence bands for the calibration approach. This result indicates that the calibration approach delivers the sharper inference than the equal-weight approach.

For $w > 40$, the two approaches yield very different confidence bands. For the calibration approach, both lower and upper bounds of the confidence bands increase as w grows. It implies that revenues and R&D are positively correlated for any $w \in [0.07, 79.04]$. For the equal-weight approach, the confidence bands have almost constant lower bounds and increasing upper bounds when $w > 40$. It indicates that the equal-weight approach faces a great deal of uncertainty on the relationship between revenues and R&D for large firms.

The reason why the calibration approach delivers the sharper inference than the equal-weight approach, especially for large firms, is that the former takes into account the revenues of the 289 firms with unobserved R&D while the latter does not. Most of the 289 firms are present at the lower tail, and hence the calibration approach assigns

large weights on small firms and small weights on large firms. Hence, the calibration approach asserts that the positive correlation between revenues and R&D observed for small firms should be preserved for large firms too, even though there is apparently a substantial dispersion across large firms. The equal-weight approach, by contrast, is directly affected by the dispersion at the upper tail and loses confidence on the positive correlation detected at the lower tail.

7 Conclusion

The existing literature of copula-based regression models typically focuses on either conditional mean or quantile regression, and assumes complete data. This paper has unified the conditional mean regression of Noh, El Ghouch, and Bouezmarni (2013), the conditional quantile regression of Noh, El Ghouch, and Van Keilegom (2015), and other regressions by formulating the general loss function. Furthermore, we relaxed the rather strong assumption of complete data by allowing the regressand and regressors to be missing at random. The calibration estimation of the regression curve is proposed, and its consistency and asymptotic normality are proved.

Our simulation results indicate that the proposed approach performs well in finite samples, while the benchmark equal-weight approach fails under the MAR mechanism. The empirical application on revenues and R&D expenses of U.S. manufacturing firms highlights a practical use of the calibration approach. The latter detects a positive correlation between the revenues and R&D for any firm size. The equal-weight approach, by contrast, detects a positive correlation for small firms but produces mixed results for large firms.

Acknowledgements

We thank Tatsuyoshi Okimoto and Naoya Sueishi, seminar participants at Kobe University, conference participants at the 2019 Japanese Joint Statistical Meeting for their helpful comments. The first author is grateful for the financial supports of JSPS KAKENHI Grant Number (A) 17H00983. The second author is grateful for the financial supports of JSPS KAKENHI Grant Number 19K13670 and Japan Center for Economic Research. The last author acknowledges the fund for building world-class universities (disciplines) of Renmin University of China.

References

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2016): "Globally Efficient Non-Parametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting," *Journal of the Royal Statistical Society, Series B*, 78, 673–700.

Chen, X., and Y. Fan (2005): "Pseudo-Likelihood Ratio Tests for Semiparametric Multivariate Copula Model Selection," *The Canadian Journal of Statistics*, 33, 389–414.

——— (2006): "Estimation of Copula-Based Semiparametric Time Series Models," *Journal of Econometrics*, 130, 307–335.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics*, 36, 808–843.

De Backer, M., A. El Ghouch, and I. Van Keilegom (2017): "Semiparametric copula quantile regression for complete or censored data," *Electronic Journal of Statistics*, 11, 1660–1698.

Delaigle, A., W. Huang, and S. Lei (2019): "Estimation of conditional prevalence from group testing data with missing covariates," *Journal of the American Statistical Association*, pp. 1–29.

Ding, W., and P. X.-K. Song (2016): "EM Algorithm in Gaussian Copula with Missing Data," *Computational Statistics and Data Analysis*, 101, 1–11.

Emura, T., C.-W. Lin, and W. Wang (2010): "A goodness-of-fit test for Archimedean copula models in the presence of right censoring," *Computational Statistics & Data Analysis*, 54(12), 3033–3043.

Emura, T., and W. Wang (2010): "Testing quasi-independence for truncation data," *Journal of Multivariate Analysis*, 101(1), 223–239.

——— (2012): "Nonparametric maximum likelihood estimation for dependent truncation data based on copulas," *Journal of Multivariate Analysis*, 110, 171–188.

Genest, C., K. Ghoudi, and L.-P. Rivest (1995): "A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions," *Biometrika*, 82, 543–552.

Hamori, S., K. Motegi, and Z. Zhang (2019): "Calibration estimation of semiparametric copula models with data missing at random," *Journal of Multivariate Analysis*, 173, 85–109.

Horvitz, D. G., and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Imbens, G. W., R. H. Spady, and P. Johnson (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357.

Kang, J. D. Y., and J. L. Schafer (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22(4), 523–539.

Kitamura, Y., and M. Stutzer (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861–874.

Kraus, D., and C. Czado (2017): "D-vine copula based quantile regression," *Computational Statistics and Data Analysis*, 110, 1–18.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of econometrics*, 79(1), 147–168.

Newey, W. K., and J. L. Powell (1987): "Asymmetric least squares estimation and testing," *Econometrica: Journal of the Econometric Society*, pp. 819–847.

Noh, H., A. El Ghouch, and T. Bouezmarni (2013): "Copula-based regression estimation and inference," *Journal of the American Statistical Association*, 108(502), 676–688.

Noh, H., A. El Ghouch, and I. Van Keilegom (2015): "Semiparametric conditional quantile estimation through copula-based multivariate models," *Journal of Business & Economic Statistics*, 33(2), 167–178.

Pakes, A., and D. Pollard (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica: Journal of the Econometric Society*, pp. 1027–1057.

Qin, J., D. Leung, and J. Shao (2002): "Estimation with Survey Data under Nonignorable Nonresponse or Informative Sampling," *Journal of the American Statistical Association*, 97, 193–200.

Rémillard, B., B. Nasri, and T. Bouezmarni (2017): "On copula-based conditional quantile estimators," *Statistics and Probability Letters*, 128, 14–20.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Rubin, D. B. (1976): "Inference and Missing Data," *Biometrika*, 63, 581–592.

Technical Appendices

A Proof of Theorem 1

Note that

$$
\begin{aligned}
&\sup_{a\in\mathcal{A}}\left|\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L(g(Y_i)-a)c\left(\hat{F}_{0,K}(Y_i),\hat{F}_{1,K}(w_1),\ldots,\hat{F}_{d,K}(w_d);\hat{\boldsymbol{\theta}}_K\right)\right.\\
&\quad\left.-\mathrm{E}\left[\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}L(g(Y_i)-a)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\right|\\
\leq&\sup_{a\in\mathcal{A}}\left|\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L(g(Y_i)-a)c\left(\hat{F}_{0,K}(Y_i),\hat{F}_{1,K}(w_1),\ldots,\hat{F}_{d,K}(w_d);\hat{\boldsymbol{\theta}}_K\right)\right.\\
&\quad\left.-\sum_{i=1}^{N}\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}L(g(Y_i)-a)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right| \qquad \text{(A.1)}\\
&+\sup_{a\in\mathcal{A}}\left|\sum_{i=1}^{N}\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}L(g(Y_i)-a)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right.\\
&\quad\left.-\mathrm{E}\left[\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}L(g(Y_i)-a)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\right|. \qquad \text{(A.2)}
\end{aligned}
$$

To show $\hat{a}(\boldsymbol{w})\xrightarrow{p}a_0(\boldsymbol{w})$, it is sufficient to show both (A.1) and (A.2) are of $o_p(1)$. By Proposition 1 and Assumption 12, and the result $\sup_{x\in\mathcal{X}}|N\hat{p}_{0K}(x)-\pi_0^{-1}(x)|=o_p(1)$ (Hamori, Motegi, and Zhang (2019, Theorem 1)), the term (A.1) is trivially of $o_p(1)$. For the term (A.2), note that for fixed $a\in\mathcal{A}$,

$$
\begin{aligned}
&\left|\sum_{i=1}^{N}\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}L(g(Y_i)-a)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right.\\
&\quad\left.-\mathrm{E}\left[\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}L(g(Y_i)-a)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\right|=o_p(1).
\end{aligned}
$$

Moreover, the dominating function is integrable:

$$
\begin{aligned}
&\mathrm{E}\left[\sup_{a\in\mathcal{A}}\left|\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}L(g(Y_i)-a)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right|\right]\\
=&\mathrm{E}\left[\sup_{a\in\mathcal{A}}\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}L(g(Y_i)-a)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\\
\leq&\delta\cdot\mathrm{E}\left[\sup_{a\in\mathcal{A}}L(g(Y_i)-a)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\\
=&\frac{\delta}{c_{\boldsymbol{W}}(F_1(\omega_1),\ldots,F_d(\omega_d))}\cdot\mathrm{E}\left[\sup_{a\in\mathcal{A}}L(g(Y_i)-a)|\boldsymbol{W}_i=\boldsymbol{w}\right]<\infty,
\end{aligned}
$$

where δ is defined in Assumption 4. Hence, by the uniform law of large numbers, the term (A.2) is of $o_p(1)$.

B Proof of Theorem 2

The following lemma will be used later to prove Theorem 2:

Lemma 1. *Under Assumptions 2-7, for all $j \in \{0, 1, ..., d\}$ and any square integrable function $\phi(Y, \boldsymbol{W}, \boldsymbol{X})$, we have*

$$\sum_{i=1}^{N} T_{ji}\hat{p}_{j,K}(\boldsymbol{X}_i)\phi(Y_i, \boldsymbol{W}_i, \boldsymbol{X}_i) = \frac{1}{N}\sum_{i=1}^{N} \frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)}\phi(Y_i, \boldsymbol{W}_i, \boldsymbol{X}_i) - \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} - 1\right\}\mathrm{E}[\phi(Y_i, \boldsymbol{W}_i, \boldsymbol{X}_i)|\boldsymbol{X}_i] + o_p(N^{-1/2}).$$

A proof of Lemma 1 is omitted since it is similar to the proof of Hamori, Motegi, and Zhang (2019, Theorem 2).

By Assumption 13 (iii), we have $\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-\hat{a})c(\hat{F}_{0,K}(Y_i), \hat{F}_{1,K}(w_1), \ldots, \hat{F}_{d,K}(w_d); \hat{\boldsymbol{\theta}}_K) = o_p(N^{-1/2})$. Since the loss function $L(\cdot)$ may not be twice differentiable, we cannot directly apply the Taylor's expansion to obtain the expression for $\sqrt{N}(\hat{a} - a_0)$. Define the function

$$f(a) := \mathrm{E}\left[\frac{T_0}{\pi_0(\boldsymbol{X})} \cdot L'(g(Y) - a)c(F_0(Y), F_1(w_1), ..., F_d(w_d); \boldsymbol{\theta}_0)\right],$$

which is differentiable with respect to a. Define

$$\nu_N(a) := \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{T_{0i}N\hat{p}_{0,K}(\boldsymbol{X}_i) \cdot L'(g(Y_i) - a)c(F_0(Y), F_1(w_1), ..., F_d(w_d); \boldsymbol{\theta}_0) - f(a)\right\},$$

which is an empirical process indexed by a. Note that $f(a_0) = 0$ holds by definition. Using the Mean Value Theorem, we have $0 = \sqrt{N}f(a_0) = \sqrt{N}f(\hat{a}) - f'(\tilde{a}) \cdot \sqrt{N}(\hat{a} - a_0)$, where $\tilde{a}$ lies between a_0 and $\hat{a}$. Since $f'(a)$ is a continuous function of a and $\hat{a} \xrightarrow{p} a_0$, we have

$$\begin{aligned}
&\sqrt{N}(\hat{a} - a_0) = f'(a_0)^{-1} \cdot \sqrt{N}f(\hat{a}) + o_p(1) \\
=&f'(a_0)^{-1} \cdot \left\{\sqrt{N}f(\hat{a}) - \sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i) \cdot L'(g(Y_i) - \hat{a})c(F_0(Y_i), F_1(w_1), ..., F_d(w_d); \boldsymbol{\theta}_0)\right. \\
&\qquad \left. + \sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i) \cdot L'(g(Y_i) - \hat{a})c(F_0(Y_i), F_1(w_1), ..., F_d(w_d); \boldsymbol{\theta}_0)\right\} + o_p(1) \\
=&- f'(a_0)^{-1}\nu_N(\hat{a}) + f'(a_0)^{-1} \cdot \sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i) \cdot L'(g(Y_i) - \hat{a})c(F_0(Y_i), F_1(w_1), ..., F_d(w_d); \boldsymbol{\theta}_0) + o_p(1)
\end{aligned}$$

$$
\begin{aligned}
=& - f'(a_0)^{-1}\nu_N(\hat{a}) + f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)\cdot L'(g(Y_i)-\hat{a})c(F_0(Y_i),F_1(w_1),...,F_d(w_d);\boldsymbol{\theta}_0)\\
&- f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-\hat{a})c\left(\hat{F}_{0,K}(Y_i),\hat{F}_{1,K}(w_1),\ldots,\hat{F}_{d,K}(w_d);\hat{\boldsymbol{\theta}}_K\right)+o_p(1)\\
=& - f'(a_0)^{-1}\cdot[\nu_N(\hat{a})-\nu_N(a_0)] - f'(a_0)^{-1}\cdot\nu_N(a_0)\\
&+ f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)\cdot L'(g(Y_i)-\hat{a})c(F_0(Y_i),F_1(w_1),...,F_d(w_d);\boldsymbol{\theta}_0)\\
&- f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-\hat{a})c\left(\hat{F}_{0,K}(Y_i),\hat{F}_{1,K}(w_1),\ldots,\hat{F}_{d,K}(w_d);\hat{\boldsymbol{\theta}}_K\right)+o_p(1)\\
=& - f'(a_0)^{-1}\cdot\nu_N(a_0) + f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)\cdot L'(g(Y)-\hat{a})c(F_0(Y),F_1(w_1),...,F_d(w_d);\boldsymbol{\theta}_0)\\
&- f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-\hat{a})c\left(\hat{F}_{0,K}(Y_i),\hat{F}_{1,K}(w_1),\ldots,\hat{F}_{d,K}(w_d);\hat{\boldsymbol{\theta}}_K\right)+o_p(1)\\
=& - f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)\cdot L'(g(Y_i)-a_0)c(F_0(Y_i),F_1(w_1),...,F_d(w_d);\boldsymbol{\theta}_0)\\
&+ f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)\cdot L'(g(Y_i)-\hat{a})c(F_0(Y_i),F_1(w_1),...,F_d(w_d);\boldsymbol{\theta}_0)\\
&- f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-\hat{a})c\left(\hat{F}_{0,K}(Y_i),\hat{F}_{1,K}(w_1),\ldots,\hat{F}_{d,K}(w_d);\hat{\boldsymbol{\theta}}_K\right)\\
=& - f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)\cdot L'(g(Y_i)-a_0)c(F_0(Y),F_1(w_1),...,F_d(w_d);\boldsymbol{\theta}_0)\\
&- f'(a_0)^{-1}\cdot\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-\hat{a})\cdot\partial_0 c\left(\tilde{F}_{0,K}(Y_i),\tilde{F}_{1,K}(w_1),\ldots,\tilde{F}_{d,K}(w_d);\tilde{\boldsymbol{\theta}}_K\right)\sqrt{N}\{\hat{F}_{0,K}(Y_i)-F_0(Y_i)\}\\
&- f'(a_0)^{-1}\cdot\sum_{i=1}^{N}\sum_{j=1}^{d} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-\hat{a})\cdot\partial_j c\left(\tilde{F}_{0,K}(Y_i),\tilde{F}_{1,K}(w_1),\ldots,\tilde{F}_{d,K}(w_d);\tilde{\boldsymbol{\theta}}_K\right)\sqrt{N}\{\hat{F}_{j,K}(w_j)-F_j(w_j)\}\\
&- f'(a_0)^{-1}\cdot\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-\hat{a})\cdot\partial_{\boldsymbol{\theta}} c\left(\tilde{F}_{0,K}(Y_i),\tilde{F}_{1}(w_1),\ldots,\tilde{F}_{d}(w_d);\tilde{\boldsymbol{\theta}}_K\right)\sqrt{N}(\hat{\boldsymbol{\theta}}_K-\boldsymbol{\theta}_0),
\end{aligned}
\tag{B.1}
$$

where $(\tilde{F}_{0,K}(Y_i),\tilde{F}_{1,K}(w_1),\ldots,\tilde{F}_{d,K}(w_d))$ lies on the line joining from $(\hat{F}_{0,K}(Y_i),\hat{F}_{1,K}(w_1),\ldots,\hat{F}_{d,K}(w_d))$ to $(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d))$, and $\tilde{\boldsymbol{\theta}}_K$ lies on the line joining $\hat{\boldsymbol{\theta}}_K$ and $\boldsymbol{\theta}_0$. By Lemma 1, we can deduce the following identities.

$$
\begin{aligned}
&\sqrt{N}\sum_{i=1}^{N} T_{0i}\hat{p}_{0K}(\boldsymbol{X}_i)L'(g(Y_i)-a_0)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\\
=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)}L'(g(Y_i)-a_0)c\left(F_0(Y_i),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right.
\end{aligned}
$$

$$- \left(\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} - 1 \right) \mathrm{E}\left[L'(g(Y_i) - a_0) c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) | \boldsymbol{X}_i \right] \Bigg\}, \quad \text{(B.2)}$$

and

$$\begin{aligned}
&\sqrt{N} \sum_{i=1}^{N} T_{0i} \hat{p}_{0K}(\boldsymbol{X}_i) L'(g(Y_i) - \hat{a}) \partial_0 c\left(\tilde{F}_{0,K}(Y_i), \tilde{F}_1(w_1), \ldots, \tilde{F}_d(w_d); \tilde{\boldsymbol{\theta}}_K \right) \{ \hat{F}_{0,K}(Y_i) - F_0(Y_i) \} \\
=&\sqrt{N} \sum_{i=1}^{N} T_{0i} \hat{p}_{0K}(\boldsymbol{X}_i) \Bigg\{ L'(g(Y_i) - \hat{a}) \partial_0 c\left(\tilde{F}_{0,K}(Y_i), \tilde{F}_{1,K}(w_1), \ldots, \tilde{F}_{d,K}(w_d); \tilde{\boldsymbol{\theta}}_K \right) \\
&\qquad - L'(g(Y_i) - a_0) \partial_0 c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \Bigg\} \{ \hat{F}_{0,K}(Y_i) - F_0(Y_i) \} \\
&+ \sqrt{N} \sum_{i=1}^{N} T_{0i} \hat{p}_{0K}(\boldsymbol{X}_i) L'(g(Y_i) - a_0) \partial_0 c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \{ \hat{F}_{0,K}(Y_i) - F_0(Y_i) \} \\
=&o_p(1) + \int L'(g(y) - a_0) \partial_0 c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \cdot \sqrt{N} \{ \hat{F}_{0,K}(y) - F_0(y) \} dF_0(y) \\
=&o_p(1) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \int L'(g(y) - a_0) \partial_0 c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\qquad \cdot \left\{ \frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} \mathbf{1}(Y_i \le y) - \left(\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} - 1 \right) F_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) - F_0(y) \right\} dF_0(y), \quad \text{(B.3)}
\end{aligned}$$

and

$$\begin{aligned}
&\sqrt{N} \sum_{i=1}^{N} T_{0i} \hat{p}_{0K}(\boldsymbol{X}_i) L'(g(Y_i) - \hat{a}) \sum_{j=1}^{d} \partial_j c\left(\tilde{F}_{0,K}(Y_i), \tilde{F}_{1,K}(w_1), \ldots, \tilde{F}_{d,K}(w_d); \tilde{\boldsymbol{\theta}}_K \right) \{ \hat{F}_j(w_j) - F_j(w_j) \} \\
=&o_p(1) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{j=1}^{d} \mathrm{E}\left[L'(g(Y) - a_0) \partial_j c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right] \\
&\qquad \cdot \left\{ \frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} \mathbf{1}(W_{ji} \le w_j) - \left(\frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} - 1 \right) F_{W_j|\boldsymbol{X}}(w_j|\boldsymbol{x}) - F_j(w_j) \right\}. \quad \text{(B.4)}
\end{aligned}$$

and

$$\begin{aligned}
&\sqrt{N} \sum_{i=1}^{N} T_{0i} \hat{p}_{0K}(\boldsymbol{X}_i) L'(g(Y_i) - \hat{a}) \partial_{\boldsymbol{\theta}} c\left(\tilde{F}_{0,K}(Y_i), \tilde{F}_{1,K}(w_1), \ldots, \tilde{F}_{d,K}(w_d); \tilde{\boldsymbol{\theta}}_K \right) (\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}_0) \\
=&o_p(1) + \mathrm{E}\left[L'(g(Y) - a_0) \partial_{\boldsymbol{\theta}} c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right] \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{\eta}_i. \quad \text{(B.5)}
\end{aligned}$$

Combine (B.1)-(B.5) to obtain

$$\begin{aligned}
&\sqrt{N}(\hat{a}_K(\boldsymbol{w}) - a_0(\boldsymbol{w})) \\
&= -\partial_a \mathrm{E}\left[L'(g(Y) - a_0) c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right]^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\times \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} L'(g(Y_i) - a_0) c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right. \\
&\quad - \left(\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} - 1 \right) \mathrm{E}\left[L'(g(Y_i) - a_0) c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) | \boldsymbol{X}_i \right] \\
&\quad + \int L'(g(y) - a_0) \partial_0 c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\qquad \times \left[\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} \mathbf{1}(Y_i \leq y) - \left(\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} - 1 \right) F_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) - F_0(y) \right] dF_0(y) \\
&\quad + \sum_{j=1}^{d} \mathrm{E}\left[L'(g(Y) - a_0) \partial_j c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right] \\
&\qquad \times \left\{ \frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} \mathbf{1}(W_{ji} \leq w_j) - \left(\frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} - 1 \right) F_{W_j|\boldsymbol{X}}(w_j|\boldsymbol{x}) - F_j(w_j) \right\} \\
&\quad \left. + \boldsymbol{\eta}_i^{\top} \mathrm{E}\left[L'(g(Y) - a_0) \partial_{\boldsymbol{\theta}} c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right] \right\} \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} S(T_i, \boldsymbol{X}_i, Y_i; \boldsymbol{w}) + o_p(1),
\end{aligned}
$$

where

$$
\begin{aligned}
&S(T_i, \boldsymbol{X}_i, Y_i; \boldsymbol{w}) \\
&\quad = -\partial_a \mathrm{E}\left[L'(g(Y) - a_0) c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right]^{-1} \\
&\qquad \times \left\{ \frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} L'(g(Y_i) - a_0) c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right. \\
&\qquad\quad - \left(\frac{T_{0i}}{\pi_0(\boldsymbol{X}_i)} - 1 \right) \mathrm{E}\left[L'(g(Y_i) - a_0) c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) | \boldsymbol{X}_i \right] \\
&\qquad\quad + \int L'(g(y) - a_0) \partial_0 c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\qquad\qquad \times \left[\frac{T_i}{\pi(\boldsymbol{X}_i)} \mathbf{1}(Y_i \leq y) - \left(\frac{T_i}{\pi(\boldsymbol{X}_i)} - 1 \right) F_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) - F_0(y) \right] dF_0(y) \\
&\qquad\quad + \sum_{j=1}^{d} \mathrm{E}\left[L'(g(Y) - a_0) \partial_j c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right] \\
&\qquad\qquad \times \left\{ \frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} \mathbf{1}(W_{ji} \leq w_j) - \left(\frac{T_{ji}}{\pi_j(\boldsymbol{X}_i)} - 1 \right) F_{W_j|\boldsymbol{X}}(w_j|\boldsymbol{x}) - F_j(w_j) \right\} \\
&\qquad\quad \left. + \boldsymbol{\eta}_i^{\top} \mathrm{E}\left[L'(g(Y) - a_0) \partial_{\boldsymbol{\theta}} c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right] \right\} \\
&\quad = \frac{1}{b(\boldsymbol{w})} \times (A_{1i} + A_{2i} + A_{3i}).
\end{aligned}
$$

C Special case I: Mean regression with complete data

In this section, we show that Theorem 2 contains a key result of the conditional mean regression derived by Noh, El Ghouch, and Bouezmarni (2013) as a special case. Specifically, we show that the influence function $S(T_i, \boldsymbol{X}_i, Y_i; \boldsymbol{w})$ in Theorem 2 reduces to that of Noh, El Ghouch, and Bouezmarni (2013) when the conditional mean regression with complete data is considered. Suppose that $T_{0i} = T_{1i} = \cdots = T_{di} \equiv 1$ for $i = 1, \ldots, N$ with probability 1, $L(v) = v^2$, and $a_0(\boldsymbol{w}) = \mathrm{E}[Y|\boldsymbol{W} = \boldsymbol{w}]$. Then, key quantities in Theorem 2 are simplified as follows.

$$
\begin{aligned}
b(\boldsymbol{w}) &= 2 \cdot \mathrm{E}\left[c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right] = 2 \cdot c_{\boldsymbol{W}}(F_1(w_1), ..., F_d(w_d); \boldsymbol{\theta}_0), \\
A_{1i} &= 2 \cdot (Y_i - a_0(\boldsymbol{w})) \cdot c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right), \\
A_{2i} &= 2 \int (y - a_0(\boldsymbol{w})) \partial_0 c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \cdot \left[\mathbf{1}(Y_i \leq y) - F_0(y)\right] dF_0(y) \\
&= 2 \int (y - a_0(\boldsymbol{w}))\left[\mathbf{1}(Y_i \leq y) - F_0(y)\right] dc\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&= 2 \int_{Y_i}^{\infty} (y - a_0(\boldsymbol{w})) dc\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad - 2 \int_{-\infty}^{\infty} (y - a_0(\boldsymbol{w})) \cdot F_0(y) dc\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&= -2 \cdot (Y_i - a_0(\boldsymbol{w})) \cdot c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad - 2 \int_{Y_i}^{\infty} c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) dy \\
&\quad + 2 \int_{-\infty}^{\infty} c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) F_0(y) dy \\
&\quad + 2 \int_{-\infty}^{\infty} c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) (y - a_0(\boldsymbol{w})) dF_0(y) \\
&= 2 \int_{-\infty}^{\infty} c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) (y - a_0(\boldsymbol{w})) dF_0(y) \\
&\quad - 2 \cdot (Y_i - a_0(\boldsymbol{w})) \cdot c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad - 2 \int (\mathbf{1}(Y_i \leq y) - F_0(y)) c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) dy, \\
A_{3i} &= 2 \cdot \sum_{j=1}^{d} \mathrm{E}\left[(Y - a_0(\boldsymbol{w})) \partial_j c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right] \cdot \left\{\mathbf{1}(W_{ji} \leq w_j) - F_j(w_j)\right\} \\
&\quad + 2 \cdot \boldsymbol{\eta}_i^{\top} \mathrm{E}\left[(Y - a_0(\boldsymbol{w})) \partial_{\boldsymbol{\theta}} c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right].
\end{aligned}
$$

Hence, the influence function $S(T_i, \boldsymbol{X}_i, Y_i; \boldsymbol{w})$ in Theorem 2 is simplified as follows.

$$
S_{mean} = \frac{1}{b(\boldsymbol{w})} \times (A_{1i} + A_{2i} + A_{3i})
$$

$$
\begin{aligned}
=&\frac{1}{c_{\boldsymbol{W}}(F_1(w_1),...,F_d(w_d);\boldsymbol{\theta}_0)}\times\Bigg\{\int_{-\infty}^{\infty} c\left(F_0(y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)(y-a_0(\boldsymbol{w}))dF_0(y)\\
&-\int(\mathbf{1}(Y_i\le y)-F_0(y))c\left(F_0(y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)dy\\
&+\sum_{j=1}^{d}\mathrm{E}\left[(Y-a_0(\boldsymbol{w}))\partial_j c\left(F_0(Y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\cdot\{\mathbf{1}(W_{ji}\le w_j)-F_j(w_j)\}\\
&+\boldsymbol{\eta}_i^{\top}\mathrm{E}\left[(Y-a_0(\boldsymbol{w}))\partial_{\boldsymbol{\theta}} c\left(F_0(Y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\Bigg\}\\
=&\frac{1}{c_{\boldsymbol{W}}(F_1(w_1),...,F_d(w_d);\boldsymbol{\theta}_0)}\times\Bigg\{-\int(\mathbf{1}(Y_i\le y)-F_0(y))c\left(F_0(y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)dy\\
&+\sum_{j=1}^{d}\mathrm{E}\left[(Y-a_0(\boldsymbol{w}))\partial_j c\left(F_0(Y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\cdot\{\mathbf{1}(W_{ji}\le w_j)-F_j(w_j)\}\\
&+\boldsymbol{\eta}_i^{\top}\mathrm{E}\left[(Y-a_0(\boldsymbol{w}))\partial_{\boldsymbol{\theta}} c\left(F_0(Y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\Bigg\},
\end{aligned}
$$

where the third equality holds because

$$
\frac{1}{c_{\boldsymbol{W}}(F_1(w_1),...,F_d(w_d);\boldsymbol{\theta}_0)}\left\{\int_{-\infty}^{\infty} c\left(F_0(y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)(y-a_0(\boldsymbol{w}))dF_0(y)\right\}=0.
$$

Hence, our influence function reduces to that of Noh, El Ghouch, and Bouezmarni (2013).

D Special case II: Quantile regression with complete data

In this section, we show that Theorem 2 contains a key result of the conditional quantile regression derived by Noh, El Ghouch, and Van Keilegom (2015) as a special case. Specifically, we show that the influence function $S(T_i,\boldsymbol{X}_i,Y_i;\boldsymbol{w})$ in Theorem 2 reduces to that of Noh, El Ghouch, and Van Keilegom (2015) when the conditional τ^{th}-quantile regression with complete data is considered. Suppose that $T_{0i}=T_{1i}=\cdots=T_{di}\equiv 1$ for $i=1,\ldots,N$ with probability 1 and $L(v)=v(\tau-\mathbf{1}(v\le 0))$, which implies that $L'(v)=\tau-\mathbf{1}(v\le 0)$. Then, key quantities in Theorem 2 are simplified as follows.

$$
\begin{aligned}
b(\boldsymbol{w})&=-\partial_a\mathrm{E}\left[(\tau-\mathbf{1}(Y\le a_0))c\left(F_0(Y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\\
&=\partial_a\int_{-\infty}^{a} c\left(F_0(y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)f_0(y)dy\Big|_{a=a_0(\boldsymbol{w})}\\
&=c\left(F_0(a_0(\boldsymbol{w})),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)f_0(a_0(\boldsymbol{w})),
\end{aligned}
$$

$$
\begin{aligned}
A_{1i} &= \{\tau - \mathbf{1}(Y_i \le a_0(\boldsymbol{w}))\} \cdot c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right), \\
A_{2i} &= \int \{\tau - \mathbf{1}(y \le a_0)\} \partial_0 c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \cdot [\mathbf{1}(Y_i \le y) - F_0(y)]\, dF_0(y) \\
&= \int \{\tau - \mathbf{1}(y \le a_0)\} \cdot [\mathbf{1}(Y_i \le y) - F_0(y)]\, dc\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&= \tau \cdot \left[\int_{Y_i}^{\infty} dc\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) - \int_{-\infty}^{\infty} F_0(y) dc\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \right] \\
&\quad - \int_{Y_i}^{a_0} dc\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) + \int_{-\infty}^{a_0} F_0(y) dc\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&= -\tau \cdot c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) + \tau \cdot c_{\boldsymbol{W}}(F_1(w_1), ..., F_d(w_d); \boldsymbol{\theta}_0) \\
&\quad - \mathbf{1}(Y_i \le a_0) \cdot c\left(F_0(a_0(\boldsymbol{w})), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad + \mathbf{1}(Y_i \le a_0) \cdot c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad + F_0(a_0(\boldsymbol{w})) \cdot c\left(F_0(a_0(\boldsymbol{w})), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad - \int_{-\infty}^{a_0} c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) dF_0(y) \\
&= -\tau \cdot c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad - \mathbf{1}(Y_i \le a_0(\boldsymbol{w})) \cdot c\left(F_0(a_0(\boldsymbol{w})), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad + \mathbf{1}(Y_i \le a_0(\boldsymbol{w})) \cdot c\left(F_0(Y_i), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) \\
&\quad + F_0(a_0(\boldsymbol{w})) \cdot c\left(F_0(a_0(\boldsymbol{w})), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right),
\end{aligned}
$$

where the last equality holds because

$$
\begin{aligned}
&\tau \cdot c_{\boldsymbol{W}}(F_1(w_1), ..., F_d(w_d); \boldsymbol{\theta}_0) - \int_{-\infty}^{a_0} c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) dF_0(y) \\
=&\mathrm{E}\left[L'(Y - a_0) c\left(F_0(y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right] = o_p(N^{-1/2}).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
A_{3i} =& \sum_{j=1}^{d} \mathrm{E}\left[\{\tau - \mathbf{1}(Y \le a_0)\} \partial_j c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right] \cdot \{\mathbf{1}(W_{ji} \le w_j) - F_j(w_j)\} \\
&+ \boldsymbol{\eta}_i^{\top} \mathrm{E}\left[\{\tau - \mathbf{1}(Y \le a_0)\} \partial_{\boldsymbol{\theta}} c\left(F_0(Y), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)\right].
\end{aligned}
$$

Hence, the influence function $S(T_i, \boldsymbol{X}_i, Y_i; \boldsymbol{w})$ in Theorem 2 is simplified as follows.

$$
\begin{aligned}
&S_{quantile} = \frac{1}{b(\boldsymbol{w})} \times (A_{1i} + A_{2i} + A_{3i}) \\
=&\frac{1}{c\left(F_0(a_0(\boldsymbol{w})), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right) f_0(a_0(\boldsymbol{w}))} \\
&\times \Big[- \{\mathbf{1}(Y_i \le a_0(\boldsymbol{w})) - F_0(a_0(\boldsymbol{w}))\} \cdot c\left(F_0(a_0(\boldsymbol{w})), F_1(w_1), \ldots, F_d(w_d); \boldsymbol{\theta}_0\right)
\end{aligned}
$$

$$+\sum_{j=1}^{d}\mathrm{E}\left[\{\tau-\mathbf{1}(Y\leq a_0)\}\partial_j c\left(F_0(Y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\cdot\{\mathbf{1}(W_{ji}\leq w_j)-F_j(w_j)\}$$

$$+\boldsymbol{\eta}_i^{\top}\mathrm{E}\left[\{\tau-\mathbf{1}(Y\leq a_0)\}\partial_{\boldsymbol{\theta}} c\left(F_0(Y),F_1(w_1),\ldots,F_d(w_d);\boldsymbol{\theta}_0\right)\right]\Bigg],$$

which coincides with the influence function of Noh, El Ghouch, and Van Keilegom (2015).