Statistica Sinica: Supplement

ON STRATIFIED DENSITY-RATIO MODELS

Moming Li[1] and Guoqing Diao[2]

[1] *University of California San Francisco and* [2] *George Washington University*

Supplementary Material

S1 Second Real-Data Example

We now turn our attention to the analysis of the German health data. This five-year-period (1984–1988) dataset is a subset from the well-known German national health registry database (SOEP Group, 2001; Hilbe, 2011). A sample of 19609 observations on 17 variables are available from the R package **COUNT**. A sub-sample of size=1000 with 15 variables is randomly drawn from the sample, constituting a dataset for our data analysis. We pick the variable *docvis* (in log-scale), number of doctor visits during a year, as the response variable, and other 14 socio-economic variables as covariates. Summary of these variables and transformations used in our models are briefed in Table 1.

The categorical covariate A we consider here is the binary variable

Table 1: Summary of German Health 1984-1988 data (a random sample of size=1000).

Abbreviation	Description	Summary Statistics	Transformation
docvis	number of doctor visits during year	range:0-49, mean=3.346	$x \mapsto \log(1+x)$
outwork	1=out of work, 0=working	#1=361, #0=639	-
hospvis	number of days in hospital during year	range:0-11, mean=0.132	$x \mapsto \log(1+x)$
age	age in years	range:25-64, mean=44.07	-
income	household yearly income in marks (DM/1000)	range:0.4-12, mean=3.357	$x \mapsto \log(1+x)$
female	1=female, 0=male	#1=487, #0=513	-
married	1=married, 0=not married	#1=778, #0=222	-
kids	1=have children, 0=no children	#1=405, #0=595	-
self	1=self-employed, 0=not self-employed	#1=67, #0=933	-
edlevel1	reference level, not high school graduate	#0=796	-
edlevel2	1=high school graduate	#1=52	-
edlevel3	1=university/college	#1=78	-
edlevel4	1=graduate school	#1=74	-
year.84	reference level, year 1984	#0=206	-
year.85	1=year 1985	#1=206	-
year.86	1=year 1986	#1=181	-
year.87	1=year 1987	#1=167	-
year.88	1=year 1988	#1=240	-

outwork, which takes value 1 if the selected person is out of work, and 0 otherwise. Table 2 presents regression coefficient estimates using both the DRM and the SDRM, where in the later case the regression coefficient estimate of *outwork* is replaced by that of the dispersion parameter ϕ. Estimated baseline CDFs are plotted in Figure 1.

Table 2: Estimated coefficients for German health 1984-1988 data (a random sample of size=1000).

SDRM					DRM				
Var.	Coef.	Std. Err.	t	$P > \|t\|$	Var.	Coef.	Std. Err.	t	$P > \|t\|$
hospvis	0.240	0.041	5.810	< 0.001	hospvis	0.299	0.037	8.110	< 0.001
age	0.125	0.039	3.252	0.001	age	0.168	0.041	4.047	< 0.001
income	-0.002	0.031	-0.053	0.958	income	0.010	0.039	0.263	0.793
female	0.153	0.067	2.270	0.023	female	0.186	0.079	2.358	0.018
married	-0.026	0.074	-0.359	0.720	married	-0.045	0.092	-0.491	0.623
kids	-0.184	0.069	-2.652	0.008	kids	-0.200	0.084	-2.398	0.016
self	-0.201	0.149	-1.355	0.175	self	-0.291	0.156	-1.863	0.062
edlevel2	0.195	0.120	1.620	0.105	edlevel2	0.235	0.153	1.538	0.124
edlevel3	-0.081	0.117	-0.696	0.486	edlevel3	-0.121	0.142	-0.851	0.395
edlevel4	-0.191	0.132	-1.443	0.149	edlevel4	-0.211	0.150	-1.407	0.159
year.85	-0.156	0.090	-1.732	0.083	year.85	-0.198	0.111	-1.786	0.074
year.86	-0.037	0.088	-0.424	0.672	year.86	-0.049	0.112	-0.442	0.658
year.87	-0.032	0.090	-0.356	0.722	year.87	-0.016	0.114	-0.142	0.887
year.88	-0.056	0.083	-0.671	0.502	year.88	-0.073	0.105	-0.692	0.489
ϕ	0.481	0.205	2.352	0.019	outwork	0.191	0.086	2.237	0.025

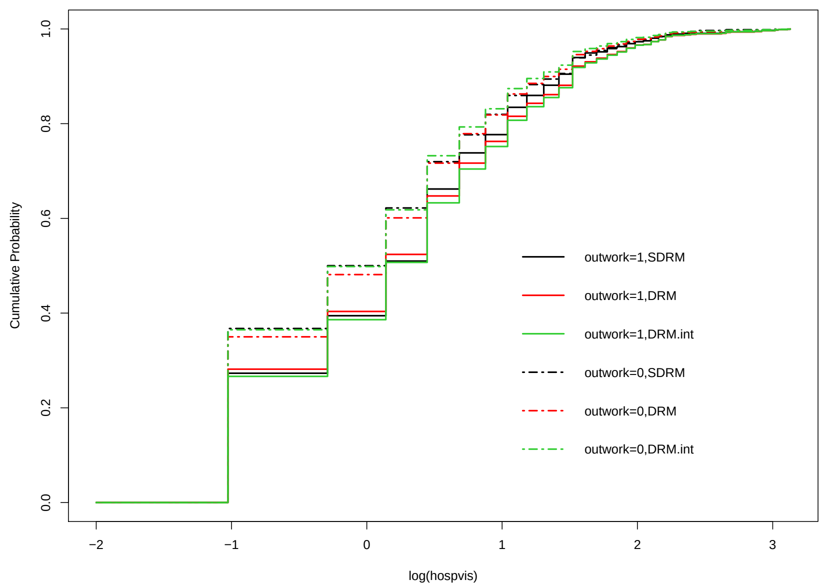

Figure 1: Estimated baseline CDFs for German health 1984-1988 data (a random sample of size=1000).

From Figure 1, we do not see much difference in the baseline CDF estimates between the two models for each *outwork* group. The goodness-of-fit test described in Section 3 is not significant with a p-value of 0.269 based on 2,000 bootstrap samples. However, the dispersion parameter ϕ has a significant estimate of 0.481 (p-value $= 0.019$). All these results together suggest that *outwork* impacts the effects of other covariates on the response variable, that is, there may exist interaction effect between *outwork* and other covariates.

For this application, results from the SDRM and the DRM are mainly in agreement, possibly because the departure from the density-ratio assumption is not severe. Labeled by DRM.int in Figure 1, we also include the estimated baseline CDFs based on the DRM with *outwork* interacting with all other covariates. All methods give very close baseline CDF estimates, however, a direct diagnostic procedure is still required to evaluate the adequacy.

S2 Proofs

The following notations, which are consistent with the notations used in previous sections, will be used throughout the entire Appendix section. Let $\boldsymbol{\eta} = (\boldsymbol{\theta}, \mathbf{F})$ be the parameters under consideration, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi})$ are finite-dimensional, and $\mathbf{F}$ is infinite-dimensional; their regularity conditions are postulated in Section 2 of the main article. Let $\mathbb{P}_n$ and $\mathbb{P}$ be the empirical measure and the expectation of n i.i.d. observations $\mathcal{O}_1, ..., \mathcal{O}_n$. That is, for any measurable function $g(\cdot)$,

$$\mathbb{P}_n[g(\mathcal{O})] = \frac{1}{n}\sum_{i=1}^{n} g(\mathcal{O}_i), \quad \mathbb{P}[g(\mathcal{O})] = \mathrm{E}_{\boldsymbol{\eta}_0}[g(\mathcal{O})],$$

where $\boldsymbol{\eta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \mathbf{F}_0)$ are the true parameter values.

S2.1 Proof of Theorem 1

We first prove that under conditions (C1)–(C5), the parameters $\boldsymbol{\eta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F})$ are identifiable. Recall from Section 2 of the main article that $\boldsymbol{\phi} = (\phi_1, ..., \phi_{K-1})$ and $\mathbf{F} = (F_1, ..., F_K)$. The likelihood function about $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F})$ based on a size n sample of i.i.d. observations $\{\mathcal{O}_i = (Y_i, \mathbf{X}_i, A_i), i = 1, ..., n\}$ is given by

$$\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\frac{dF_k(Y_i) \exp\{Y_i \boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_i V(\phi_k)\}}{\int_{\mathcal{Y}_k} \exp\{s \boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_i V(\phi_k)\} dF_k(s)} \right]^{I\{A_i = k\}},$$

where $dF_k(\cdot)$ $(k = 1, ..., K)$ are probability density functions assumed with respect to some dominating measure. Note that $\boldsymbol{\beta}$ is common to all strata while ϕ_k and F_k are stratum-specific. Suppose that two sets of parameter values $\bar{\boldsymbol{\eta}}$ and $\widetilde{\boldsymbol{\eta}}$ give the same likelihood function for a single observation $\mathcal{O} = (Y, \mathbf{X}, A)$. Then, for $A = k$, we have

$$\frac{d\bar{F}_k(Y) \exp\{Y \bar{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\bar{\phi}_k)\}}{\int_{\mathcal{Y}} \exp\{s \bar{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\bar{\phi}_k)\} d\bar{F}_k(s)} = \frac{d\widetilde{F}_k(Y) \exp\{Y \widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\widetilde{\phi}_k)\}}{\int_{\mathcal{Y}_k} \exp\{s \widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\widetilde{\phi}_k)\} d\widetilde{F}_k(s)}. \tag{S2.1}$$

Since (S2.1) holds for all $\mathbf{X}$, by letting $\mathbf{X} = \mathbf{0}$ we have $\bar{F}_k(y) = \widetilde{F}_k(y)$, for any $y \in \mathcal{Y}_k$. It follows that

$$\frac{\exp\{Y \bar{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\bar{\phi}_k)\}}{\int_{\mathcal{Y}_k} \exp\{s \bar{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\bar{\phi}_k)\} d\bar{F}_k(s)} = \frac{\exp\{Y \widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\widetilde{\phi}_k)\}}{\int_{\mathcal{Y}_k} \exp\{s \widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\widetilde{\phi}_k)\} d\widetilde{F}_k(s)}. \tag{S2.2}$$

Substitute $y_1 \neq y_2 \in \mathcal{Y}_k$ for Y in (S2.2), we have

$$\frac{\exp\{y_1 \bar{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\bar{\phi}_k)\}}{\int_{\mathcal{Y}_k} \exp\{s \bar{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\bar{\phi}_k)\} d\bar{F}_k(s)} = \frac{\exp\{y_1 \widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\widetilde{\phi}_k)\}}{\int_{\mathcal{Y}_k} \exp\{s \widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X} V(\widetilde{\phi}_k)\} d\widetilde{F}_k(s)},$$

and

$$\frac{\exp\{y_2\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}V(\bar{\phi}_k)\}}{\int_{\mathcal{Y}_k}\exp\{s\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}V(\bar{\phi}_k)\}d\bar{F}_k(s)} = \frac{\exp\{y_2\widetilde{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}V(\widetilde{\phi}_k)\}}{\int_{\mathcal{Y}_k}\exp\{s\widetilde{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}V(\widetilde{\phi}_k)\}d\widetilde{F}_k(s)},$$

respectively. It follows that

$$(y_1 - y_2)\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}V(\bar{\phi}_k) = (y_1 - y_2)\widetilde{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}V(\widetilde{\phi}_k).$$

For $A = K$, note that $\bar{\phi}_K = \widetilde{\phi}_K = 0$ and $V(0) = 1$. From condition (C1) in Section 2 of the main article, we obtain $\bar{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}$. It follows immediately that $\bar{\phi}_k = \widetilde{\phi}_k$ $(k = 1, ..., K-1)$. This establishes the identifiability of the parameters $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F})$. With this result, we next prove the consistency of the NPMLEs.

Recall from Section 2 of the main article that the sample size is n_k in stratum k while the number of distinct observations is m_k, and $\widetilde{F}_{n,k}(t) = \sum_{j=1}^{m_k}\widetilde{p}_{kj}I\{Y_{k(j)} \leqslant t\}$ is the NPMLE of $F_{k0}(t)$, $\widetilde{p}_{kj} = \widetilde{F}_{n,k}\{Y_{k(j)}\}$. Indexed by $\{n\}_{n\in\mathbb{N}}$, let $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n) \in \boldsymbol{\Theta}$ be a sequence of estimators of $(\boldsymbol{\beta}, \boldsymbol{\phi})$. Since $\boldsymbol{\Theta}$ is compact, there exists a subsequence $\{n_l\}_{l\in\mathbb{N}}$ such that $(\widetilde{\boldsymbol{\beta}}_{n_l}, \widetilde{\boldsymbol{\phi}}_{n_l}) \to (\boldsymbol{\beta}^*, \boldsymbol{\phi}^*)$, for some point $(\boldsymbol{\beta}^*, \boldsymbol{\phi}^*) \in \boldsymbol{\Theta}$. Since $\widetilde{F}_{n,k}$ is uniformly bounded over $\mathcal{Y}_k$, Helly's Selection Theorem implies that, for any subsequence, we can always choose a further subsequence such that $\widetilde{F}_{n_l,k}$ converges pointwise to some

For the ease of notation and presentation, let the sub-subsequence be still indexed by $\{n_l\}_{l\in\mathbb{N}}$.

distribution function F_k^* in $\mathcal{Y}_k$. Recall from (2.8) in the main article indexed by the subsequence $\{n_l\}_{l\in\mathbb{N}}$, $\widetilde{F}_{n_l,k}$ satisfies

$$\widetilde{F}_{n_l,k}\{Y_{k(j)}\} = \frac{\lambda_{kj}}{n\mathbb{P}_n[Q(\mathcal{O};\widetilde{\boldsymbol{\beta}}_n, \widetilde{\phi}_{n,k}, \widetilde{F}_{n,k})I\{A=k\}]}\bigg|_{Y=Y_{k(j)}},$$

where

$$Q(\mathcal{O};\boldsymbol{\beta},\phi_k,F_k) = \frac{\exp\{Y\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}V(\phi_k)\}}{\int_{\mathcal{Y}_k}\exp\{s\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}V(\phi_k)\}dF_k(s)}.$$

Next, we construct another step function $\breve{F}_{n,k}(t)$ by imitating $\widetilde{F}_{n,k}(t)$ as

$$\breve{F}_{n,k}\{Y_{k(j)}\} = \frac{\lambda_{kj}}{n\mathbb{P}_n[Q(\mathcal{O};\boldsymbol{\beta}_0, \phi_{k0}, F_{k0})I\{A=k\}]}\bigg|_{Y=Y_{k(j)}}.$$

Since both

$$\mathcal{F}_1 = \{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}V(\phi_k) : (\boldsymbol{\beta},\boldsymbol{\phi})\in\boldsymbol{\Theta}\},$$

and

$$\mathcal{F}_2 = \{F_k(y) : F_k \text{ is a distribution function on } \mathcal{Y}_k\}$$

are $\mathbb{P}$-Donsker classes, and Q is bounded away from 0, the preservation of the Donsker property (van der Vaart and Wellner, 1996) implies that the following class

$$\mathcal{Q} = \{Q^{-1}(\mathcal{O};\boldsymbol{\beta},\phi_k,F_k) : y\in\mathcal{Y}_k, (\boldsymbol{\beta},\boldsymbol{\phi})\in\boldsymbol{\Theta}, F_k \text{ is a distribution function on } \mathcal{Y}_k\}$$

is a bounded $\mathbb{P}$-Donsker class, and hence is also a $\mathbb{P}$-Glivenko-Cantelli class. By the Glivenko-Cantelli theorem, uniformly in $t \in \mathcal{Y}_k$, the followings hold almost surely:

$$\begin{aligned}
\breve{F}_{n,k}(t) &= \sum_{j=1}^{m_k} \breve{F}_{n,k}\{Y_{k(j)}\} I\{Y_{k(j)} \leqslant t\} \\
&= \sum_{j=1}^{m_k} \frac{\lambda_{kj} I\{Y_{k(j)} \leqslant t\}}{n\mathbb{P}_n[Q(\mathcal{O}; \boldsymbol{\beta}_0, \phi_{k0}, F_{k0}) I\{A = k\}]} \\
&\to \mathrm{E}_{\boldsymbol{\eta}_0}\left[\frac{I\{Y \leqslant t\}}{\mu(Y)}\middle| A = k\right],
\end{aligned}$$

where $\mu(Y|A = k) = \mathrm{E}_{\boldsymbol{\eta}_0}[Q(\mathcal{O}; \boldsymbol{\beta}_0, \phi_{k0}, F_{k0})] = Q(\mathcal{O}; \boldsymbol{\beta}_0, \phi_{k0}, F_{k0})$. Direct calculation gives

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\eta}_0}\left[\frac{I\{Y \leqslant t\}}{\mu(Y)}\middle| A = k\right] = \int_{\mathcal{Y}_k} \frac{I\{y \leqslant t\} \exp\{y\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X} V(\phi_{k0})\}}{\mu(y) \int_{\mathcal{Y}_k} \exp\{s\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X} V(\phi_{k0})\} dF_k(s)} dF_{k0}(y) \\
&= \int_{\mathcal{Y}_k} \left(\frac{I\{y \leqslant t\} \exp\{y\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X} V(\phi_{k0})\}}{\left[\frac{\exp\{y\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X} V(\phi_{k0})\}}{\int_{\mathcal{Y}_k} \exp\{s\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)}\right] \int_{\mathcal{Y}_k} \exp\{s\boldsymbol{\beta}_0^{\mathsf{T}} \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)} \right) dF_{k0}(y) \\
&= \int_{\mathcal{Y}_k} I\{y \leqslant t\} dF_{k0}(y) \\
&= F_{k0}(t).
\end{aligned}$$

Consequently, we conclude that $\breve{F}_{n,k}(t)$ converges uniformly to $F_{k0}(t)$ on $\mathcal{Y}_k$ almost surely.

Since $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n)$ maximizes $\ell_n(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F})$, we have $\ell_n(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n) \geqslant$

$\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \breve{\mathbf{F}})$. Let $n \to \infty$, we have

$$0 \leqslant \frac{1}{n}\ell_n(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n) - \frac{1}{n}\ell_n(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \breve{\mathbf{F}}_n) \to \mathrm{E}_{\boldsymbol{\eta}_0}\left[\log \frac{\prod_{k=1}^K \left[\frac{dF_k(Y_i)\exp\{Y_i\boldsymbol{\beta}^{*\intercal}\mathbf{X}_iV(\phi_k^*)\}}{\int_{\mathcal{Y}_k}\exp\{s\boldsymbol{\beta}^{*\intercal}\mathbf{X}_iV(\phi_k^*)\}dF_k^*(s)}\right]^{I\{A_i=k\}}}{\prod_{k=1}^K \left[\frac{dF_{k0}(Y_i)\exp\{Y_i\boldsymbol{\beta}_0^\intercal\mathbf{X}_iV(\phi_{k0})\}}{\int_{\mathcal{Y}_k}\exp\{s\boldsymbol{\beta}_0^\intercal\mathbf{X}_iV(\phi_{k0})\}dF_{k0}(s)}\right]^{I\{A_i=k\}}}\right],$$

which is the negative Kullback-Leibler information in $(\boldsymbol{\beta}^*, \boldsymbol{\phi}^*, \mathbf{F}^*)$. Together with the identifiability results proved at the beginning of Section S2.1, we conclude that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$, and $\mathbf{F}^* = \mathbf{F}_0$.

S2.2 Proof of Theorem 2

Consider the set of indices

$$\mathcal{A} = \{\mathbf{H} \equiv (\mathbf{b}, \mathbf{c}, \mathbf{h}) : \mathbf{b} \in \mathbb{R}^d, \mathbf{c} \in \mathbb{R}^{K-1}, \mathbf{h} \in \mathcal{H}^K;$$
$$\|\mathbf{b}\| \leqslant 1, \|\mathbf{c}\| \leqslant 1, |h_k|_V \leqslant 1, k = 1, ..., K\},$$

where $|h_k|_V$ denotes the total variation of $h_k(\cdot)$ on $\mathcal{Y}_k$.

Define a neighborhood of the true parameters $\boldsymbol{\eta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \mathbf{F}_0)$ as follows:

$$\mathcal{U} = \{\boldsymbol{\eta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F}) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| + \|\boldsymbol{\phi} - \boldsymbol{\phi}_0\| + \sum_{k=1}^K \sup_{t \in \mathcal{Y}_k} |F_k(t) - F_{k0}(t)| < \epsilon_0\}, \tag{S2.3}$$

where $\epsilon_0 > 0$ is a small constant. If n_k, for all $k = 1, ..., K$, is large enough, then $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n)$ belong to $\mathcal{U}$ with probability approaching one.

Denote by $\ell(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F}) = \sum_{k=1}^K I\{A = k\}\ell_k(\boldsymbol{\beta}, \phi_k, F_k)$ the log-likelihood function about $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F})$ based on a single observation $(Y, \mathbf{X}, A)$. Recall that $\boldsymbol{\beta}$ is common to all strata, while ϕ_k and F_k are stratum-specific. Let $\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\eta})$ and $\dot{\ell}_{\phi_k}(\boldsymbol{\eta})$ denote the derivatives of $\ell(\boldsymbol{\eta})$ with respect to $\boldsymbol{\beta}$ and ϕ_k $(k = 1, ..., K-1)$, respectively. Then, $\mathbf{b}^\mathsf{T}\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\eta})$ is the score function for $\boldsymbol{\beta}$ corresponding to a one-dimensional submodel $\mathrm{P}(\boldsymbol{\beta} + \epsilon\mathbf{b}, \boldsymbol{\phi}, \mathbf{F})$, for a small enough $\epsilon > 0$. Likewise, $c_k\dot{\ell}_{\phi_k}(\boldsymbol{\eta})$ is the score function for ϕ_k corresponding to a one-dimensional submodel $\mathrm{P}(\boldsymbol{\beta}, \phi_k + \epsilon c_k, F_k)$. For $k = 1, ..., K$, let $\dot{\ell}_{F_k}(\boldsymbol{\eta})[h_k]$ denote the path-wise derivative of $\ell(\boldsymbol{\eta})$ with respect to F_k along the path $F_k(y) + \epsilon\int_{\mathcal{Y}_k} Q_{F_k}[h_k](y)dF_k(y)$, where $Q_{F_k}[h_k](y) = h_k(y) - \int_{\mathcal{Y}_k} h_k(y)dF_k(y)$.

We calculate each derivative as follows:

$$
\begin{aligned}
\mathbf{b}^\mathsf{T}\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\eta}) &= \sum_{k=1}^{K}\left[I\{A=k\}\frac{d\ell_k(\boldsymbol{\beta}+\epsilon\mathbf{b}, \phi_k, F_k)}{d\epsilon}\bigg|_{\epsilon=0}\right] \\
&= \sum_{k=1}^{K} I\{A=k\}\left[Y - \frac{\int_{\mathcal{Y}} s\exp\{s\boldsymbol{\beta}^\mathsf{T}\mathbf{X}V(\phi_k)\}dF_k(s)}{\int_{\mathcal{Y}}\exp\{s\boldsymbol{\beta}^\mathsf{T}\mathbf{X}V(\phi_k)\}dF_k(s)}\right]\mathbf{b}^\mathsf{T}\mathbf{X}V(\phi_k) \\
&= \sum_{k=1}^{K} I\{A=k\}\mathbf{b}^\mathsf{T}\left[\mathbf{X}V(\phi_k)\left\{Y - \mathrm{E}(Y|\mathbf{X})\right\}\right],
\end{aligned}
$$

$$
\begin{aligned}
c_k\dot{\ell}_{\phi_k}(\boldsymbol{\eta}) &= I\{A=k\}\frac{d\ell_k(\boldsymbol{\beta}, \phi_k+\epsilon c_k, F_k)}{d\epsilon}\bigg|_{\epsilon=0} \\
&= I\{A=k\}\left[Y - \frac{\int_{\mathcal{Y}_k} s\exp\{s\boldsymbol{\beta}^\mathsf{T}\mathbf{X}V(\phi_k)\}dF_k(s)}{\int_{\mathcal{Y}}\exp\{s\boldsymbol{\beta}^\mathsf{T}\mathbf{X}V(\phi_k)\}dF_k(s)}\right]\boldsymbol{\beta}^\mathsf{T}\mathbf{X}V'(\phi_k)c_k \\
&= I\{A=k\}c_k\left[\boldsymbol{\beta}^\mathsf{T}\mathbf{X}V'(\phi_k)\left\{Y - \mathrm{E}(Y|\mathbf{X})\right\}\right],
\end{aligned}
\tag{S2.4}
$$

$$
\begin{aligned}
\dot{\ell}_{F_k}(\boldsymbol{\eta})[h_k] &= I\{A=k\}\frac{d\ell_k(\boldsymbol{\beta},\phi_k,F_k+\epsilon\int_{\mathcal{Y}}Q_{F_k}[h_k]dF_k)}{d\epsilon}\bigg|_{\epsilon=0}\\
&= I\{A=k\}\left(Q_{F_k}[h_k](Y)-\left[\frac{\int_{\mathcal{Y}}Q_{F_k}[h_k](s)\exp\{s\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}V(\phi_k)\}dF_k(s)}{\int_{\mathcal{Y}_k}\exp\{s\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}V(\phi_k)\}dF_k(s)}\right]\right)\\
&= I\{A=k\}\left(Q_{F_k}[h_k](Y)-\mathrm{E}\big[Q_{F_k}[h_k](Y)|\mathbf{X}\big]\right).
\end{aligned}
$$

Then, the score operator indexed by $\mathbf{H}\in\mathcal{A}$ is defined as

$$
\psi(\boldsymbol{\eta})[\mathbf{H}] = \mathbf{b}^{\mathsf{T}}\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\eta}) + \sum_{k=1}^{K-1} c_k\dot{\ell}_{\phi_k}(\boldsymbol{\eta}) + \sum_{k=1}^{K}\dot{\ell}_{F_k}(\boldsymbol{\eta})[h_k]. \tag{S2.5}
$$

We define a sequence of maps $\Psi_n:\mathcal{U}\to l^\infty(\mathcal{A})$ as follows:

$$
\begin{aligned}
\Psi_n(\boldsymbol{\eta})[\mathbf{H}] &= \mathbb{P}_n\left[\psi(\boldsymbol{\eta})[\mathbf{H}]\right]\\
&= \mathbb{P}_n\left[\mathbf{b}^{\mathsf{T}}\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\eta}) + \sum_{k=1}^{K-1} c_k\dot{\ell}_{\phi_k}(\boldsymbol{\eta}) + \sum_{k=1}^{K}\dot{\ell}_{F_k}(\boldsymbol{\eta})[h_k]\right]\\
&= \mathbb{P}_n\left[\mathbf{b}^{\mathsf{T}}\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\eta})\right] + \sum_{k=1}^{K-1}\mathbb{P}_n\left[c_k\dot{\ell}_{\phi_k}(\boldsymbol{\eta})\right] + \sum_{k=1}^{K}\mathbb{P}_n\left[\dot{\ell}_{F_k}(\boldsymbol{\eta})[h_k]\right]\\
&\equiv A_n^{(1)}[\mathbf{b}] + \sum_{k=1}^{K-1}A_n^{(2)}[c_k] + \sum_{k=1}^{K}A_n^{(3)}[h_k],
\end{aligned}
$$

where $A_n^{(1)}$, $A_n^{(2)}$, and $A_n^{(3)}$ can be viewed as linear functionals defined on $\mathbb{R}^d$, $\mathbb{R}$, and $BV(\mathcal{Y}_k)$, working on indices $\mathbf{b}$, c_k, and h_k, respectively, and $BV(\mathcal{Y}_k)$ denotes the space of functions defined on $\mathcal{Y}_k$ with bounded variation.

Correspondingly, we can define the limiting map $\Psi:\mathcal{U}\to l^\infty(\mathcal{A})$ as

$$
\Psi(\boldsymbol{\eta})[\mathbf{H}] = A^{(1)}[\mathbf{b}] + \sum_{k=1}^{K-1}A^{(2)}[c_k] + \sum_{k=1}^{K}A^{(3)}[h_k],
$$

where the linear functionals $A^{(1)}$, $A^{(2)}$, and $A^{(3)}$ are obtained by replacing the empirical measures by the corresponding expectations. Clearly, $\Psi_n(\widetilde{\boldsymbol{\eta}}_n) = 0$, and $\Psi(\boldsymbol{\eta}_0) = 0$. Then, $\sqrt{n}(\Psi_n - \Psi)(\boldsymbol{\eta}) = \{\mathbb{G}_n\psi(\boldsymbol{\eta})[\mathbf{H}] : \mathbf{H} \in \mathcal{A}\}$ is an empirical process in the space $l^\infty(\mathcal{A})$ indexed by the class of score functions $\{\psi(\boldsymbol{\eta})[\mathbf{H}] : \mathbf{H} \in \mathcal{A}\}$. To prove the asymptotic normality of the NPMLEs $\widetilde{\boldsymbol{\eta}}_n = (\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n)$, we shall verify the conditions stated in Theorem 3.3.1 of van der Vaart and Wellner (1996). From the definition and the consistency result we have established, $\Psi(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \mathbf{F}_0) = 0$ and $\Psi_n(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n) = o_P(n^{-1/2})$ hold. It remains to verify the following four conditions:

(VW1)[approximation condition]

$$\begin{aligned}&\sqrt{n}(\Psi_n - \Psi)(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n) - \sqrt{n}(\Psi_n - \Psi)(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \mathbf{F}_0)\\ &= o_P\left(1 + \sqrt{n}\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \sqrt{n}\|\widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0\| + \sqrt{n}\sum_{k=1}^{K}\sup_{t\in\mathcal{Y}_k}|\widetilde{F}_{n,k}(t) - F_{k0}(t)|\right).\end{aligned}$$

(VW2)[asymptotic distribution of score function]

$\sqrt{n}(\Psi_n - \Psi)(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \mathbf{F}_0) \rightsquigarrow \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a tight Gaussian process on $l^\infty(\mathcal{A})$.

(VW3)[Fréchet-differentiability]

The map $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F}) \mapsto \Psi(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F})$ is Fréchet differentiable at $(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \mathbf{F}_0)$.

(VW4)[invertibility]

The derivative of $\Psi(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F})$ at $(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \mathbf{F}_0)$, denoted by $\dot{\Psi}(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \mathbf{F}_0)$, is continuously invertible.

Recall the neighborhood $\mathcal{U}$ around the true parameter value $\boldsymbol{\eta}_0$ from (S2.3), the class

$$\{\psi(\boldsymbol{\eta})[\mathbf{H}] - \psi(\boldsymbol{\eta}_0)[\mathbf{H}] : \boldsymbol{\eta} \in \mathcal{U}, \mathbf{H} \in \mathcal{A}\}$$

is a Donsker class, and $A^{(1)}$, $A^{(2)}$, and $A^{(3)}$ are bounded Lipschitz functionals with respect to $\mathcal{A}$. Therefore, as $\boldsymbol{\eta} \to \boldsymbol{\eta}_0$,

$$\sup_{\mathbf{H}\in\mathcal{A}} \mathrm{E}_{\boldsymbol{\eta}_0}\left[\left\{\psi(\boldsymbol{\eta})[\mathbf{H}] - \psi(\boldsymbol{\eta}_0)[\mathbf{H}]\right\}^2\right] \to 0.$$

According to Lemma 3.3.5 in van der Vaart and Wellner (1996), the approximation condition is satisfied. Since $A_n^{(1)}$, $A_{nk}^{(2)}$, and $A_{nk}^{(3)}$ are bounded Lipschitz functionals with respect to $\mathcal{A}$, and the class of score functions $\{\psi(\boldsymbol{\eta})[\mathbf{H}] : \mathbf{H} \in \mathcal{A}\}$ is $\mathbb{P}$-Donsker, by the Donsker Theorem, $\sqrt{n}(\Psi_n - \Psi)(\boldsymbol{\eta}_0)$ weakly converges to a tight zero-mean Gaussian process $\boldsymbol{\xi}$ in $l^\infty(\mathcal{A})$ indexed by $\mathbf{H}$. The covariance function between $\boldsymbol{\xi}(\mathbf{H}_1)$ and $\boldsymbol{\xi}(\mathbf{H}_2)$ is given by

$$\mathrm{E}_{\boldsymbol{\eta}_0}\big[\psi(\boldsymbol{\eta}_0)[\mathbf{H}_1] \times \psi(\boldsymbol{\eta}_0)[\mathbf{H}_2]\big].$$

Therefore, the asymptotic distribution of score function condition is satisfied. By the smoothness of $\Psi(\boldsymbol{\eta})$, the Fréchet differentiability condition holds and the derivative of $\Psi(\boldsymbol{\eta})$ at $\boldsymbol{\eta}_0$, denoted by $\dot{\Psi}(\boldsymbol{\eta}_0)$, is a map from the space $\{\boldsymbol{\eta}_0 - \boldsymbol{\eta} : \boldsymbol{\eta} \in \mathcal{U}\}$ to $l^\infty(\mathcal{A})$. To verify the invertibility condition, we follow the arguments in Zeng and Lin (2007, 2010). It suffices to prove that for any one-dimensional submodel

$$\mathrm{P}\left(\boldsymbol{\beta}_0 + \epsilon \mathbf{b}, \boldsymbol{\phi}_0 + \epsilon \mathbf{c}, F_{10} + \epsilon \int_{\mathcal{Y}_1} Q_{F_{10}}[h_k] dF_{10}, ..., F_{K0} + \epsilon \int_{\mathcal{Y}_K} Q_{F_{K0}}[h_k] dF_{K0}\right),$$

the Fisher information along this submodel is non-singular. If the Fisher information along this submodel is singular, then the score function for this submodel is 0 almost surely. This is similar to prove the identifiability of the model parameters. Recall the definition of the score operator ψ indexed by $\mathbf{H}$ in (S2.5) with components defined in (S2.4), we will show that $\psi(\boldsymbol{\eta}_0)[\mathbf{H}] = 0$ implies $\mathbf{H} = (\mathbf{b}, \mathbf{c}, h_1, ..., h_K) = \mathbf{0}$. For a single observation $(Y, \mathbf{X}, A)$, when $A = k$, $\mathbf{X} = \mathbf{0}$ implies $Q_{F_{k0}}[h_k](Y) - \int_{\mathcal{Y}_k} Q_{F_{k0}}[h_k](y) dF_{k0}(y) = 0$. Thus, $Q_{F_{k0}}[h_k](y) = 0$, for all $y \in \mathcal{Y}_k$. Let $y_1 \neq y_2 \in \mathcal{Y}_k$. We have

$$\begin{aligned}&\left[y_1 - \frac{\int_{\mathcal{Y}_k} s \exp\{s\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)}{\int_{\mathcal{Y}_k} \exp\{s\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)}\right] \mathbf{b}^\top \mathbf{X} V(\phi_{k0}) \\ &\quad + \left[y_1 - \frac{\int_{\mathcal{Y}_k} s \exp\{s\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)}{\int_{\mathcal{Y}_k} \exp\{s\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)}\right] \boldsymbol{\beta}_0^\top \mathbf{X} V'(\phi_{k0}) c_k = 0,\end{aligned} \tag{S2.6}$$

and

$$\begin{aligned}&\left[y_2 - \frac{\int_{\mathcal{Y}_k} s \exp\{s\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)}{\int_{\mathcal{Y}_k} \exp\{s\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)}\right] \mathbf{b}^\top \mathbf{X} V(\phi_{k0}) \\ &\quad + \left[y_2 - \frac{\int_{\mathcal{Y}_k} s \exp\{s\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)}{\int_{\mathcal{Y}_k} \exp\{s\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\} dF_{k0}(s)}\right] \boldsymbol{\beta}_0^\top \mathbf{X} V'(\phi_{k0}) c_k = 0.\end{aligned} \tag{S2.7}$$

Subtracting (S2.7) from (S2.6), we have

$$(y_1 - y_2)\mathbf{b}^{\mathsf{T}}\mathbf{X}V(\phi_{k0}) + (y_1 - y_2)\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}V'(\phi_{k0})c_k = 0.$$

Condition (C1) in Section 2 of the main article and $\phi_{K0} = 0$ imply $\mathbf{b} = \mathbf{0}$ and $c_k = 0$. Thus, all four conditions are verified, and hence we can conclude that $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0, \widetilde{\mathbf{F}}_n - \mathbf{F}_0) \rightsquigarrow -\dot{\Psi}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \mathbf{F}_0)\boldsymbol{\xi}$. Moreover, it can be shown that

$$\begin{aligned}
&\sqrt{n}\left\{\mathbf{b}^{\mathsf{T}}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \mathbf{c}^{\mathsf{T}}(\widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) + \sum_{k=1}^{K}\int_{\mathcal{Y}_k} Q_{F_k}[h_k]d(\widetilde{F}_{n,k} - F_{k0})\right\} \\
&= -\sqrt{n}(\mathbb{P}_n - \mathbb{P})\left\{\widetilde{\mathbf{b}}^{\mathsf{T}}\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\eta}_0) + \widetilde{\mathbf{c}}^{\mathsf{T}}\dot{\ell}_{\boldsymbol{\phi}}(\boldsymbol{\eta}_0) + \sum_{k=1}^{K}\dot{\ell}_{F_k}(\boldsymbol{\eta}_0)[\widetilde{h}_k]\right\} + o_P(1),
\end{aligned}$$

where $\widetilde{\mathbf{b}}$, $\widetilde{\mathbf{c}}$, and $\widetilde{h}_k$ involve the inverse of a Fredholm operator used to verify condition (VW4). From the joint asymptotic normality of $\widetilde{\boldsymbol{\eta}}_n = (\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n)$, by choosing $h_k = 0$ $(k = 1, ..., K)$, we see that $\mathbf{b}^{\mathsf{T}}\widetilde{\boldsymbol{\beta}}_n + \mathbf{c}^{\mathsf{T}}\widetilde{\boldsymbol{\phi}}_n$ is an asymptotically linear estimator of $\mathbf{b}^{\mathsf{T}}\boldsymbol{\beta}_0 + \mathbf{c}^{\mathsf{T}}\boldsymbol{\phi}_0$ with influence function $\widetilde{\mathbf{b}}^{\mathsf{T}}\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\eta}_0) + \widetilde{\mathbf{c}}^{\mathsf{T}}\dot{\ell}_{\boldsymbol{\phi}}(\boldsymbol{\eta}_0)$ lying in the space spanned by the score functions. It follows that $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n)$ are semiparametrically efficient (Bickel et al., 1993).

S2.3 Proof of Theorem 3

Theorem 3 can be considered as a direct consequence of Theorem 2. We only outline the heuristics here; detailed argument parallels that of Parner

(1998). The key point is that the variance can be uniformly approximated by its empirical counterpart under the regularity conditions.

The operator $\dot{\Psi}(\boldsymbol{\eta})[\mathbf{H}]$ maps $\boldsymbol{\eta} - \boldsymbol{\eta}_0$ to a bounded functional in $l^{\infty}(\mathcal{A})$. Specifically, $\dot{\Psi}_{\boldsymbol{\eta}_0}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0, \widetilde{\mathbf{F}}_n - \mathbf{F}_0)[\mathbf{b}, \mathbf{c}, \mathbf{h}]$ is equal to the expectation (with respect to the true parameter $\boldsymbol{\eta}_0$) of the second derivative of the log-likelihood function along the directions of $(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0, \widetilde{\mathbf{F}}_n - \mathbf{F}_0)$ and $(\mathbf{b}, \mathbf{c}, \int_{\mathcal{Y}_1} h_1 dF_{10}, ..., \int_{\mathcal{Y}_K} h_K dF_{K0})$. For any direction $\mathbf{h}_n = (\mathbf{b}, \mathbf{c}, \vec{h}_1, ..., \vec{h}_K)$, where $\vec{h}_k = (h_k(Y_{k(1)}) - h_k(Y_{k(m_k)}), ..., h_k(Y_{k(m_k-1)}) - h_k(Y_{k(m_k)}))$ and $(\mathbf{b}, \mathbf{c}, \mathbf{h})$ $= (\mathbf{b}, \mathbf{c}, h_1(\cdot), ..., h_K(\cdot)) \in \mathcal{A}$. With direction $\mathbf{h}_n$, the second derivative can be approximated uniformly in $(\mathbf{b}, \mathbf{c}, \mathbf{h}) \in \mathcal{A}$ by

$$(\mathbf{b}^{\intercal}, \mathbf{c}^{\intercal}, \vec{h}_1^{\intercal}, ..., \vec{h}_K^{\intercal})(\mathbf{J}_n/n) \begin{pmatrix} \widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \\ \widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0 \\ \widetilde{F}_{n,1}(Y_{1(1)}) - F_{10}(Y_{1(1)}) \\ \vdots \\ \widetilde{F}_{n,k}(Y_{k(j)}) - F_{k0}(Y_{k(j)}) \\ \vdots \\ \widetilde{F}_{n,K}(Y_{K(m_K-1)}) - F_{K0}(Y_{K(m_K-1)}) \end{pmatrix}_{j=1,...,m_k-1;k=1,...,K},$$

where $\mathbf{J}_n$ is the negative Hessian matrix of (2.7) in the main article with respect to $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n)$. From the joint asymptotic normality of $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n)$, we have

$$\sqrt{n}\begin{pmatrix} \widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \\ \widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0 \\ \widetilde{F}_{n,1}(Y_{1(1)}) - F_{10}(Y_{1(1)}) \\ \vdots \\ \widetilde{F}_{n,k}(Y_{k(j)}) - F_{k0}(Y_{k(j)}) \\ \vdots \\ \widetilde{F}_{n,K}(Y_{K(m_K-1)}) - F_{K0}(Y_{K(m_K-1)}) \end{pmatrix}_{j=1,\ldots,m_k-1;k=1,\ldots,K}$$
$$\overset{d}{\approx} (\mathbf{J}_n/n)^{-1/2}\mathbf{G},$$

where $\mathbf{G}$ is a standard multivariate Gaussian vector. Thus, we have

$$\sqrt{n}(\mathbf{b}^\top, \mathbf{c}^\top, \vec{h}_1^\top, ..., \vec{h}_K^\top)\begin{pmatrix} \widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \\ \widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0 \\ \widetilde{F}_{n,1}(Y_{1(1)}) - F_{10}(Y_{1(1)}) \\ \vdots \\ \widetilde{F}_{n,k}(Y_{k(j)}) - F_{k0}(Y_{k(j)}) \\ \vdots \\ \widetilde{F}_{n,K}(Y_{K(m_K-1)}) - F_{K0}(Y_{K(m_K-1)}) \end{pmatrix}_{j=1,\ldots,m_k-1;k=1,\ldots,K}$$
$$\overset{d}{\approx} (\mathbf{b}^\top, \mathbf{c}^\top, \vec{h}_1^\top, ..., \vec{h}_K^\top)(\mathbf{J}_n/n)^{-1/2}\mathbf{G}.$$

It follows that $\sqrt{n}[\mathbf{b}^\top(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \mathbf{c}^\top(\widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) + \sum_{k=1}^K \int_{\mathcal{Y}_k} h_k(t)d\{\widetilde{F}_{n,k}(t) -$

$F_{k0}(t)\}]$ converges to a zero-mean Gaussian distribution with variance V, where

$$V = \lim_{n\to\infty} n(\mathbf{b}^\top, \mathbf{c}^\top, \vec{h}_1^\top, ..., \vec{h}_K^\top)\mathbf{J}_n^{-1}(\mathbf{b}^\top, \mathbf{c}^\top, \vec{h}_1^\top, ..., \vec{h}_K^\top)^\top.$$

S2.4 Proof of Theorem 4

Theorem 4 is also a consequence of Theorem 2, hence we keep it brief. If the density-ratio assumption holds, then both the DRM and the SDRM can yield consistent estimators of the baseline CDFs, although the former one is more efficient. It suffices to notice that $\sqrt{n}(\widehat{\mathbf{F}}_n - \widetilde{\mathbf{F}}_n) = \sqrt{n}\{(\widehat{\mathbf{F}}_n - \mathbf{F}_0) - (\widetilde{\mathbf{F}}_n - \mathbf{F}_0)\} \equiv \boldsymbol{\xi}_1 - \boldsymbol{\xi}_2$, where $\boldsymbol{\xi}_1 = \sqrt{n}(\widehat{\mathbf{F}}_n - \mathbf{F}_0)$ and $\boldsymbol{\xi}_2 = \sqrt{n}(\widetilde{\mathbf{F}}_n - \mathbf{F}_0)$ both have limiting Gaussian processes with mean zeros and covariance functions obtained from the inverse of the observed Fisher information matrix.

References

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Hilbe, J. M. (2011). *Negative Binomial Regression, Second Edition*. Cambridge University Press.

Parner, E. (1998). Asymptotic theory for the correlated Gamma-frailty model. *The Annals of Statistics 26*, 183–214.

SOEP Group (2001). The german socio-economic panel (gsoep) after more than 15 years - overview. *Vierteljahrshefte zur Wirtschaftsforschung 70*, 7–14.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

Zeng, D. and D. Y. Lin (2007). Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). *Journal of the Royal Statistical Society: Series B 69*, 507–564.

Zeng, D. and D. Y. Lin (2010). A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statistica Sinica 20*, 871–910.