Statistica Sinica **32** (2022), 153-177
doi:https://doi.org/10.5705/ss.202019.0412

ON STRATIFIED DENSITY-RATIO MODELS

Moming Li and Guoqing Diao

University of California San Francisco and George Washington University

Abstract: Density-ratio models are receiving increasing attention, particularly because of their relationship with generalized linear models and their applications in missing-data analyses. The density-ratio assumption, however, may not be true in some applications, and an important limitation is that the standard density-ratio model does not accommodate heterogeneity within the underlying population. To address these issues, we propose a new density-ratio model that incorporates a stratification procedure and dispersion parameters. The resulting stratified density-ratio model 1) retains attractive properties of the standard density-ratio model, while allowing the density-ratio assumption to be violated for some covariate, and 2) provides a validation tool, using a Kolmogorov–Smirnov-type statistic, to check the modeling assumption. We estimate the finite-dimensional and infinite-dimensional parameters simultaneously using an efficient nonparametric maximum likelihood approach. The resulting estimators are shown to be consistent and asymptotically normal. The asymptotic covariance matrix of the estimators for the finite-dimensional parameters attains the semiparametric efficiency bound.

Key words and phrases: Bootstrap test, density ratio models, generalized linear models, Kolmogorov-Smirnov test, nonparametric maximum likelihood estimation, semiparametric efficiency.

1. Introduction

The exponential family is a rich and flexible parametric family of distributions possessing nice theoretical properties and wide practical applicability. The generalized linear model (GLM) (McCullagh and Nelder (1989)) relates the response variable Y and θ, which contains covariate information, as

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{1.1}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions, and ϕ is called the dispersion parameter; refer to Jørgensen (1997) for a comprehensive treatment and generalization of the error distributions considered by Nelder and Wedderburn (1972).

A critical assumption in the classical GLM is that $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are

Corresponding author: Guoqing Diao, Department of Biostatistics and Bioinformatics, George Washington University, Washington, DC 20052, USA. E-mail: gdiao@gwu.edu.

known functions. In addition, model (1.1) is an exponential family density function with canonical parameter θ only if ϕ is known; otherwise, it is an exponential dispersion model. If $c(\cdot)$ is unspecified in model (1.1), then neither is $b(\cdot)$. Using the canonical link function with a linear predictor $\theta = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}$, where $\mathbf{X}$ is a set of covariates, we have $f(y|\phi) = \text{constant} \times \exp\{c(y, \phi)\}$ when $\mathbf{X} = \mathbf{0}$ and ϕ is a known quantity. Therefore, a properly normalized function of $c(\cdot)$ serves as a so-called baseline density function. The preceding argument heuristically introduces a semiparametric specification of the GLM when $c(\cdot)$ is left unspecified. In this study, we formalize this idea in a density-ratio modeling framework with unspecified $c(\cdot)$ and unknown ϕ.

The semiparametric density-ratio model (DRM) (Diao, Ning and Qin (2012)), also called the proportional likelihood ratio model (Luo and Tsai (2012)), has been studied extensively in recent years, with its early history dating back to Anderson (1972). Several statistical models in the literature are closely related to the DRM, such as the Cox proportional hazards model (Cox (1972, 1975)), generalized linear models (Nelder and Wedderburn (1972)), DRMs for categorical covariates (Qin and Zhang (1997); Qin (1998); Fokianos et al. (2001); Zhang (2000, 2002)), biased sampling models (Vardi (1985); Gill, Vardi and Wellner (1988); Gilbert, Lele and Vardi (1999); Chen (2001)), semiparametric single-index models (Ichimura (1993)), generalized odds ratio models (Liang and Qin (2000)), and semiparametric generalized linear models (Rathouz and Gao (2009); Huang and Rathouz (2012); Huang (2014).) While obtaining some desirable properties, such as the efficiency of the estimators (within a suitable class) and robustness to model mis-specification, studies based on the standard DRM have been extended to accommodate different types of data, such as missing and truncated data (Chan (2012)), right-censored data (Zhu (2014)), longitudinal data (Luo and Tsai (2014)), time-series data (Kedem et al. (2008); Fung and Huang (2016)), multivariate extreme-value data in risk assessment (De Carvalho and Davison (2014)), survival data from prevalent cohort studies (Zhu et al. (2017)), and correlated data with multivariate outcomes (Marchese and Diao (2017)).

The key to modeling the distribution of a response variable Y conditional on a given covariate vector $\mathbf{X}$ using the density-ratio technique is to use a properly normalized product of an unspecified baseline probability density function $f(\cdot)$ and an exponential function of the linear predictor containing the covariate information:

$$f(y|\mathbf{X}) = \frac{f(y)\exp(y\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})}{b(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}, f)}, \quad b(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}, f) = \int_{\mathcal{Y}} f(s)\exp(s\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})ds, \tag{1.2}$$

where $b(\cdot)$ is a normalizing constant dependent on $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}$ and $f(\cdot)$, and $\mathcal{Y}$ is the support of the response variable. Closely related to the DRM, note that the semiparametric generalized linear models (SPGLMs) of Rathouz and Gao (2009), Huang and Rathouz (2012), and Huang (2014) explicitly model the mean structure $\mathrm{E}(Y|\mathbf{X}) = \eta(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})$ using a user-specified inverse link function $\eta(\cdot)$ in addition to the error distribution $f(y) = \exp(b+y\theta)f_0(y)$ in density-ratio form, where $f_0(\cdot)$ is some reference density function and b is a normalizing constant. The merit of the SPGLMs is the reverse specification of the canonical parameter $\theta \equiv \theta(\mathbf{X};\boldsymbol{\beta}, f)$ as an implicit solution of the conditional mean and the error distribution, which gives $\boldsymbol{\beta}$ the usual mean contrast interpretation.

Robustness is a major advantage of the DRM and its variants. However, the performance still relies on the density-ratio assumption, by which, we mean that the logarithm of two probability density functions are related linearly in y:

$$\log \frac{f(y|\mathbf{X})}{f(y)} = y\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X} + \widetilde{b}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}, f), \tag{1.3}$$

where $\widetilde{b}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}, f) = -\log b(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}, f)$. If the functional form of the baseline density function is known, then (1.3) can be used to check the density-ratio assumption. However, the baseline density function $f(\cdot)$ is left unspecified. Therefore, validating the assumption by directly checking the linearity of the functional forms is practically infeasible under the foregoing semiparametric specification. To the best of our knowledge, very few works have justified the validity of the density-ratio assumption. Most existing studies examining the density-ratio modeling technique seek to enhance model flexibility using the robust nature and some invariant properties of the density-ratio form. Violating the density-ratio assumption can yield inconsistent estimators of the unknown parameters, both finite-dimensional and infinite-dimensional. Other statistical inferences may not be reliable either.

Under the DRM (1.2), the population is typically assumed to be homogeneous with a common baseline distribution, based upon which the linear predictor quantifies the covariate effects. If this is untrue for some covariate, especially for a discrete covariate, then including all the covariates in a linear predictor based upon a single baseline distribution is inappropriate. The following two examples offer insight into violations due to heterogeneity, motivating us to consider the stratified density-ratio model in this study.

Example 1. (ANCOVA with unequal covariances). Figure 1 plots the post- vs. pre-treatment blood lead concentration levels from a clinical trial data set

(Fitzmaurice, Laird and Ware (2012)). The open and closed circles show the succimer and placebo groups, respectively, and their corresponding fitted regression lines are also displayed. A typical feature of this type of data is that the pretreatment scores have roughly the same means and variances between the two treatment groups. However, this is not the case for the post-treatment scores, because individual responses to different treatments are unlikely to be the same. Here, the analysis of covariance (ANCOVA) model with random effects can be useful in modeling this type of data. The underlying heterogeneity is group specific. Hence, the density-ratio assumption is clearly violated if we do not consider separate baseline density functions.

The Poisson distribution (and its variants) is another commonly used parametric family of distributions when modeling count data. The density-ratio assumption is also violated in the following example of a heterogeneous negative-binomial regression model, which offers extra flexibility in modeling over-dispersed count data.

Example 2. (Negative-binomial regression with heterogeneous dispersion). The probability mass function is

$$\mathrm{P}(Y=y|A)=\frac{\Gamma(y+(1/\phi))}{\Gamma(y+1)\Gamma((1/\phi))}\left(\frac{1}{1+\phi\mu}\right)^{(1/\phi)}\left(\frac{\phi\mu}{1+\phi\mu}\right)^{y},$$

where

$$\phi\equiv\phi(A)=\begin{cases}0, & A=0,\\ \phi_1\in(0,1], & A=1.\end{cases}$$

Under the above model,

$$\mathrm{E}(Y|A)=\mu,\quad \mathrm{Var}(Y|A)=\mu+\phi(A)\mu^2.$$

Note that as $\phi\to 0$, the negative-binomial distribution $\mathrm{NB}(\mu,\mu+\phi\mu^2)$ converges to a Poisson distribution with mean μ. Here, μ can be treated as a baseline quantity or further modeled using other covariates. Clearly, the (conditional) variance here is covariate dependent.

The remainder of this paper is organized as follows. In Section 2, we present a new density-ratio model based on stratification that incorporates dispersion parameters. To avoid a fully stratified model, a common regression vector (direction) is assumed, while the magnitudes are allowed to vary across strata. We develop likelihood-based inference procedures and establish the asymptotic prop-

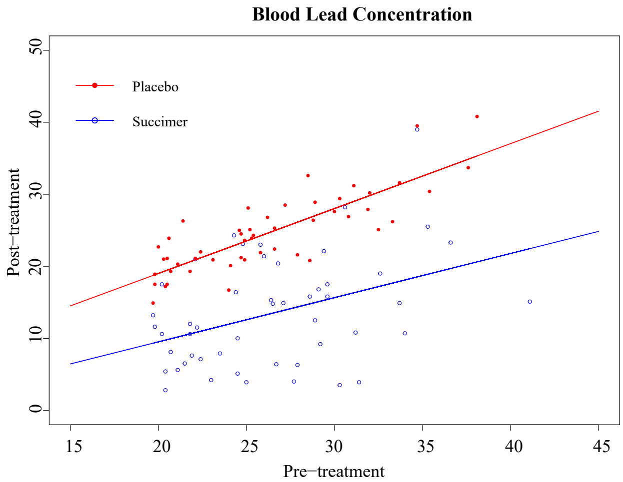

Figure 1. Treatment-specific blood lead concentration.

erties of the proposed estimators. In Section 3, we illustrate how to use the proposed model to validate the density-ratio assumption using a Kolmogorov–Smirnov-type goodness-of-fit test. In addition, we propose a bootstrap procedure to approximate the p-value of the goodness-of-fit test statistic. In Section 4, we conduct simulation studies to assess the finite-sample performance of the proposed model and the testing procedure. In Section 5, we illustrate the proposed methodology by analyzing two data sets: blood lead concentration data (Fitzmaurice, Laird and Ware (2012)), and German health data for 1984–1988 (SOEP Group (2001); Hilbe (2011)). All technical details are provided in the Supplementary Material.

2. Methods

2.1. The models

We first define some notation. Let Y be a general univariate response variable supported on $\mathcal{Y} \subseteq \mathbb{R}$. We consider a K-level categorical covariate, namely $A \in \{1, 2, \ldots, K\}$. Denote by $\mathbf{A} = (A_1, \ldots, A_{K-1})^\top$ the $(K-1) \times 1$ vector of dummy variables associated with A, where $A_k = I\{A = k\}$ $(k = 1, \ldots, K-1)$ correspond to the first $K-1$ levels of A, and the Kth variable is set as the reference level. Let $\mathbf{X}$ be a $d \times 1$ vector of other available covariates, and denote by $\mathbf{Z} = (\mathbf{X}^\top, \mathbf{A}^\top)^\top$ the totality of covariates defined on $\mathcal{Z} \subseteq \mathbb{R}^{d+K-1}$. Let $F(\cdot|\mathbf{Z})$ and $F_k(\cdot|\mathbf{X})$ be the distribution functions of Y conditional on $\mathbf{Z}$ and $(\mathbf{X}, A = k)$ $(k = 1, \ldots, K)$, respectively. Assume that the aforementioned conditional distributions all possess density functions (with respect to some proper dominating measures suppressed without ambiguity), denoted by $f(\cdot|\mathbf{Z})$ and $f_k(\cdot|\mathbf{X})$ $(k = 1, \dots, K)$, respectively. Let $f(\cdot) = f(\cdot|\mathbf{Z} = \mathbf{0})$ and $f_k(\cdot) = f_k(\cdot|\mathbf{X} = \mathbf{0})$ $(k = 1, \dots, K)$ be their corresponding baseline density functions. It is clear that $f(\cdot) = f_K(\cdot)$.

Recall that under the density-ratio assumption, $Y|(\mathbf{X}, \mathbf{A}) \sim f(y|\mathbf{X}, \mathbf{A})$ satisfies

$$f(y|\mathbf{X}, \mathbf{A}) = \frac{f(y)\exp\{y(\boldsymbol{\alpha}^\mathsf{T}\mathbf{A} + \boldsymbol{\beta}^\mathsf{T}\mathbf{X})\}}{b(\boldsymbol{\alpha}^\mathsf{T}\mathbf{A} + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}, f)},$$
$$b(\boldsymbol{\alpha}^\mathsf{T}\mathbf{A} + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}, f) = \int_{\mathcal{Y}} f(s)\exp\{s(\boldsymbol{\alpha}^\mathsf{T}\mathbf{A} + \boldsymbol{\beta}^\mathsf{T}\mathbf{X})\}ds.$$

Similarly, if the density-ratio assumption is postulated within each stratum of A, that is, $Y|(\mathbf{X}, A = k) \sim f_k(y|\mathbf{X})$ $(k = 1, \dots, K)$, then

$$f_k(y|\mathbf{X}) = \frac{f_k(y)\exp(y\boldsymbol{\beta}_k^\mathsf{T}\mathbf{X})}{b_k(\boldsymbol{\beta}_k^\mathsf{T}\mathbf{X}, f_k)}, \quad b_k(\boldsymbol{\beta}_k^\mathsf{T}\mathbf{X}, f_k) = \int_{\mathcal{Y}_k} f_k(s)\exp(s\boldsymbol{\beta}_k^\mathsf{T}\mathbf{X})ds.$$

To avoid a fully stratified model with different baseline density functions and different regression coefficients, we consider the following "parallel-slope" model:

$$\begin{aligned} f_k(y|\mathbf{X}) &= \frac{f_k(y)\exp\{y\boldsymbol{\beta}^\mathsf{T}\mathbf{X}V(\phi_k)\}}{b_k(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}, \phi_k, f_k)}, \\ b_k(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}, \phi_k, f_k) &= \int_{\mathcal{Y}_k} f_k(s)\exp\{s\boldsymbol{\beta}^\mathsf{T}\mathbf{X}V(\phi_k)\}ds, \end{aligned} \tag{2.1}$$

where $\mathcal{Y}_k$ is the support of $Y|(\mathbf{X}, A = k)$ $(k = 1, \dots, K)$, ϕ_k is the unknown dispersion parameter corresponding to stratum k $(k = 1, \dots, K)$, and $V(\cdot)$ is a known positive function subject to $V(0) = 1$. To ensure model identifiability, we set ϕ_K to zero. This semiparametric specification mimics the formulation of the classical parametric generalized linear models with dispersion parameters. A typical example is the normal regression model with mean parameter $\boldsymbol{\beta}$ and error variance σ^2, where $\boldsymbol{\beta}^* \equiv \boldsymbol{\beta}/\sigma^2$ is identified as the true parameter in the DRM. We refer to (2.1) as the stratified density-ratio model (SDRM).

Remark 1. The conditional mean/variance function is useful for prediction and model diagnostics. Based on the DRM, the conditional mean function is given by

$$\mu(\mathbf{z}) \equiv \mathrm{E}(Y|\mathbf{Z} = \mathbf{z}) = \frac{\int_{\mathcal{Y}} yf(y)\exp\{y(\alpha_k + \boldsymbol{\beta}^\mathsf{T}\mathbf{x})\}dy}{\int_{\mathcal{Y}} f(y)\exp\{y(\alpha_k + \boldsymbol{\beta}^\mathsf{T}\mathbf{x})\}dy}, \tag{2.2}$$

where $\mathbf{z} = (\mathbf{x}, k)$, and the conditional variance function is given by

$$\text{Var}(Y|\mathbf{Z} = \mathbf{z}) = \frac{\int_{\mathcal{Y}} \{y - \mu(\mathbf{z})\}^2 f(y) \exp\{y(\alpha_k + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x})\} dy}{\int_{\mathcal{Y}} f(y) \exp\{y(\alpha_k + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x})\} dy}. \tag{2.3}$$

These conditional mean and conditional variance functions are understood in the same way as any other model in a regression analysis. It is clear that (2.2) and (2.3) depend on the linear predictor $(\alpha_k + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x})$ and the baseline density $f(\cdot)$. If $f_k(\cdot) \equiv f(\cdot|A = k)$ are heterogeneous for different k, then the heterogeneity contained in the baseline density functions plays a part in the overall conditional mean and conditional variance.

Remark 2. As mentioned previously, when the true model is a normal linear regression model, the DRM can only estimate the quotient $\boldsymbol{\beta}/\sigma^2$, where σ^2 is the variance of the residual error. In general, for a K-level categorical covariate A, let $\sigma_k^2 = \sigma^2(A = k)$ $(k = 1, \ldots, K)$, and the regression parameter in the SDRM be $\boldsymbol{\beta}^*$. Then,

$$\frac{\boldsymbol{\beta}}{\sigma_k^2} = \frac{\boldsymbol{\beta}}{\sigma_K^2} \times \frac{\sigma_K^2}{\sigma_k^2} \equiv \boldsymbol{\beta}^* \times V(\phi_k). \tag{2.4}$$

We set the Kth stratum to be the reference level. Therefore, $V(\phi_k)$ is the variance ratio of the reference stratum of the baseline distribution to that of the kth stratum $(k = 1, \ldots, K - 1)$. We set $V(\cdot)$ to be the exponential function $\exp(\cdot)$. In practice, because the true variances of the distributions are unknown, the SDRM essentially estimates the log odds ratio parameter, for all strata, in the same direction $\|\boldsymbol{\beta}^*\|^{-1}\boldsymbol{\beta}^*$, the magnitude of which is controlled by the dispersion parameter in the corresponding stratum. A similar technique was used in a genetic study by Schifano et al. (2013).

Remark 3. The positive function $V(\cdot)$ should not be confused with the variance function in the GLM literature; it is not a variance function that depends on the conditional mean (e.g., $V_{\text{GLM}}(\mu) = \mu(1-\mu)$ for the binomial family). As in (2.4), $V(\phi_k)$ represents the baseline variance ratio. Any positive function subject to $V(0) = 1$ (identifiability) is a potential candidate, though the actual numerical performance may differ. Therefore, the exponential function is a natural choice.

Remark 4. From (2.4), we can also see that if the heterogeneity indeed exists, then not all the ratios σ_K^2/σ_k^2 $(k = 1, \ldots, K - 1)$ are equal to one. Therefore, without the stratification, the estimator of $\boldsymbol{\beta}^*$ is no longer consistent. This can be generalized to arbitrary responses beyond the normal data using the proposed SDRM.

Remark 5. The estimator of $\boldsymbol{\beta}^*$ can be converted back to the original scale $\boldsymbol{\beta}$ by adjusting the error variance σ_K^2 of the reference distribution. Following Marchese and Diao (2017), we estimate σ_K^2 from the residuals using the observations from the reference stratum after obtaining the regression coefficient estimates.

2.2. Nonparametric maximum likelihood estimation

Let $\{(Y_i, \mathbf{X}_i, A_i), i = 1, \ldots, n\}$ be an independent and identically distributed (i.i.d.) sample of observations of size n. The sample size in the kth group of A is $n_k = \sum_{i=1}^n I\{A_i = k\}$, for $k = 1, \ldots, K$. If the response variable Y follows a discrete or mixed underlying distribution, tied outcomes may be observed. Let m and m_k $(k = 1, \ldots, K)$ be the numbers of distinct observations in the whole sample and in the stratified sample corresponding to stratum k, respectively. Note that $n = \sum_{k=1}^K n_k$, but $m \leqslant \sum_{k=1}^K m_k$, in general. Based on the definition in (2.1) and an i.i.d. sample of size n, the likelihood function about the unknown parameter $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F})$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{K-1})$ and $\mathbf{F} = (F_1, \ldots, F_K)$, is given by

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

$$= \prod_{k=1}^K \prod_{r=1}^{n_k} \frac{dF_k(Y_{kr}) \exp\{Y_{kr} \boldsymbol{\beta}^\top \mathbf{X}_{kr} V(\phi_k)\}}{\int_{\mathcal{Y}_k} \exp\{s \boldsymbol{\beta}^\top \mathbf{X}_{kr} V(\phi_k)\} dF_k(s)}, \tag{2.5}$$

where $dF_k(\cdot) = f_k(\cdot)$ $(k = 1, \ldots, K)$ are the baseline density functions with respect to some dominating measure, and Y_{kr} and $\mathbf{X}_{kr}$ $(k = 1, \ldots, K; r = 1, \ldots, n_k)$ are the response and covariates for the rth subject in the kth group of A, respectively.

However, the likelihood function (2.5) can be maximized, without exploding to infinity, only when the baseline distribution functions are discretized at the distinct observations, and the corresponding jump sizes are considered as unknown parameters to be estimated. Let $p_{kj} = F_k\{Y_{k(j)} | \mathbf{X} = \mathbf{0}\}$ $(j = 1, \ldots, m_k; k = 1, \ldots, K)$ be the probability masses that the discretized conditional distributions assign to the ordered distinct observations when the covariate $\mathbf{X}$ takes $\mathbf{0}$, where $Y_{k(j)}$ denotes the jth-order statistic in stratum k. Then, the nonparametric likelihood function is given by

$$\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{p}) = \prod_{k=1}^K \prod_{r=1}^{n_k} \frac{p_{kj} \exp\{Y_{k(j)} \boldsymbol{\beta}^\top \mathbf{X}_{kr} V(\phi_k)\}}{\sum_{l=1}^{m_k} p_{kl} \exp\{Y_{k(l)} \boldsymbol{\beta}^\top \mathbf{X}_{kr} V(\phi_k)\}}, \tag{2.6}$$

where $\boldsymbol{p}_k = (p_{k1}, \ldots, p_{k,m_k-1})^\mathsf{T}$, for $k = 1, \ldots, K$, and $\boldsymbol{p} = (\boldsymbol{p}_1^\mathsf{T}, \ldots, \boldsymbol{p}_K^\mathsf{T})^\mathsf{T}$. We introduce an intermediate index j in (2.6) to account for possible tied values of Y, where j depends on r via $\{j : Y_{k(j)} = Y_{kr}, 1 \leqslant j \leqslant m_k, 1 \leqslant r \leqslant n_k\} \equiv \mathcal{J}_k$. For stratum k, the multiplicity of j, denoted by λ_{kj}, is defined as the cardinality of the set $\mathcal{J}_k$. It follows that $\sum_{j=1}^{m_k} \lambda_{kj} = n_k$ $(k = 1, \ldots, K)$. The corresponding nonparametric log-likelihood function can be written as

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{p}) = \sum_{k=1}^{K} \left\{ \sum_{j=1}^{m_k} \lambda_{kj} \log(p_{kj}) + \sum_{r=1}^{n_k} Y_{kr} \boldsymbol{\beta}^\mathsf{T} \mathbf{X}_{kr} V(\phi_k) - \sum_{i=1}^{n_k} \log \left[\sum_{l=1}^{m_k} p_{kl} \exp\{Y_{k(l)} \boldsymbol{\beta}^\mathsf{T} \mathbf{X}_{kr} V(\phi_k)\} \right] \right\}. \quad (2.7)$$

To maximize (2.7) with respect to $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{p})$, a commonly used approach is the re-normalizing iterative procedure; that is, repeatedly updating the estimates of the jump sizes $\boldsymbol{p}$ and the regression parameters $(\boldsymbol{\beta}, \boldsymbol{\phi})$ until both converge. It can be shown that the resulting nonparametric maximum likelihood estimators (NPMLEs) of $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{p})$, denoted by $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\boldsymbol{p}}_n)$, satisfy

$$\widetilde{p}_{kj} = \lambda_{kj} \left[\sum_{r=1}^{n_k} \frac{\exp\{Y_{k(j)} \widetilde{\boldsymbol{\beta}}^\mathsf{T} \mathbf{X}_{kr} V(\widetilde{\phi}_k)\}}{\sum_{l=1}^{m_k} \widetilde{p}_{kl} \exp\{Y_{k(l)} \widetilde{\boldsymbol{\beta}}^\mathsf{T} \mathbf{X}_{kr} V(\widetilde{\phi}_k)\}} \right]^{-1}. \quad (2.8)$$

Alternatively, Diao, Ning and Qin (2012) and Marchese and Diao (2017) proposed using the quasi-Newton algorithm (Press et al. (1992)) to directly optimize the negative nonparametric likelihood function (2.6) with respect to the regression parameters $(\boldsymbol{\beta}, \boldsymbol{\phi})$ and the re-parametrized jump sizes $\boldsymbol{p}$ via the softmax transformation

$$p_{kj} = \frac{\exp(\zeta_{kj})}{\sum_{l=1}^{m_k} \exp(\zeta_{kl})}, \ \zeta_{km_k} \equiv 0 \ \ (j = 1, \ldots, m_k - 1; k = 1, \ldots, K). \quad (2.9)$$

The softmax transformation (2.9) implicitly transforms the constrained optimization problem into an unconstrained one and stabilizes the jump sizes, thus facilitating the numerical computation. In general, the iterative procedure and the direct optimization procedure yield almost identical solutions. However, the direct optimization is considerably faster than the iterative procedure, and is less prone to convergence problems. In our numerical studies, we adopt the direct optimization approach using the softmax transformation.

Using this transformation, the resulting effective nonparametric log-likelihood function about $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\zeta})$ is given by

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\zeta}) = \sum_{k=1}^{K}\sum_{r=1}^{n_k} \log\left[\frac{\exp\{\zeta_{kj} + Y_{k(j)}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{kr}V(\phi_k)\}}{\sum_{l=1}^{m_k}\exp\{\zeta_{kl} + Y_{k(l)}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{kr}V(\phi_k)\}}\right],$$

where $\boldsymbol{\zeta}_k = (\zeta_{k1}, \ldots, \zeta_{k,m_k-1})^{\mathsf{T}}$, for $k = 1, \ldots, K$, and $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\zeta}_K^{\mathsf{T}})^{\mathsf{T}}$ is of dimension $\sum_{k=1}^{K}(m_k - 1)$. The intermediate index j plays the same role as that in (2.6). The totality of the unknown parameters is $d + K - 1 + \sum_{k=1}^{K}(m_k - 1) = d - 1 + \sum_{k=1}^{K} m_k$.

Then, the NPMLE of $F_k(t)$ $(k = 1, \ldots, K)$ is given by

$$\widetilde{F}_{n,k}(t) = \int_{\mathcal{Y}_k} I\{y \leqslant t\} d\widetilde{F}_k(y) = \sum_{j=1}^{m_k} \widetilde{p}_{kj} I\{Y_{k(j)} \leqslant t\},$$

where

$$\widetilde{p}_{kj} = \begin{cases} \dfrac{\exp(\widetilde{\zeta}_{kj})}{1 + \sum_{l=1}^{m_k-1}\exp(\widetilde{\zeta}_{kl})}, & j \neq m_k \\ 1 - \displaystyle\sum_{l=1}^{m_k-1} \widetilde{p}_{kl}, & j = m_k, \end{cases}$$

and $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\boldsymbol{\zeta}}_n)$ are obtained by solving $\nabla\ell_n(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\boldsymbol{\zeta}}_n) = \mathbf{0}$. Here, $\nabla\ell_n(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\zeta})$ denote the first derivatives of $\ell_n(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\zeta})$ with respect to the unknown parameters $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\zeta})$.

2.3. Asymptotic theory

In this subsection, we establish the asymptotic properties of the proposed NPMLEs. We first impose the following regularity conditions:

(C1) The covariate vector $\mathbf{Z} = (\mathbf{X}^{\mathsf{T}}, \mathbf{A}^{\mathsf{T}})^{\mathsf{T}}$ is bounded almost surely, and $\mathbf{a}^{\mathsf{T}}\mathbf{Z} = \mathbf{0}$ almost surely if and only if $\mathbf{a} = \mathbf{0}$.

(C2) For some fixed limit $\rho_k \in (0, 1)$, $n_k/n \to \rho_k$, as $n \to \infty$ $(k = 1, \ldots, K)$.

(C3) The true parameter values of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, denoted by $\boldsymbol{\beta}_0$ and $\boldsymbol{\phi}_0$, belong to the interior of a known compact set,

$$\boldsymbol{\Theta} = \{(\boldsymbol{\beta}, \boldsymbol{\phi}) : \|\boldsymbol{\beta}\| \leqslant B_0, \|\boldsymbol{\phi}\| \leqslant B_0, \text{ for some positive constant } B_0\}.$$

In addition, $\phi_k \mapsto V(\phi_k)$ is a known positive function with $V(0) \equiv 1$, such that it is bounded away from zero and ∞ on $\boldsymbol{\Theta}$ almost surely, and continuously differentiable in a neighborhood of ϕ_{k0} $(k = 1, \ldots, K-1)$.

(C4) The true baseline cumulative distribution function of $F_k(t)$, denoted by $F_{k0}(t) = \int_{\mathcal{Y}_k} I\{y \leqslant t\} dF_{k0}(y)$ $(k = 1, \ldots, K)$, is a class of distribution functions defined on $\mathcal{Y}_k \subseteq \mathbb{R}$ with finite first and second moments. The integral is understood in the usual Lebesgue–Stieltjes sense, where the probability density function is assumed with respect to some proper dominating measure, suppressed without ambiguity.

(C5) There exist positive constants $B_1 \leqslant B_2$ such that the following inequalities hold almost surely:

$$B_1 \leqslant \mathrm{E}_{\boldsymbol{\eta}_0}\left[\exp\{Y\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\}\right] \leqslant B_2,$$

$$\mathrm{E}_{\boldsymbol{\eta}_0}\left\|\frac{\partial}{\partial \boldsymbol{\theta}_0}\exp\{Y\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\}\right\| \leqslant B_2,$$

$$\mathrm{E}_{\boldsymbol{\eta}_0}\left\|\frac{\partial^2}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0^\top}\exp\{Y\boldsymbol{\beta}_0^\top \mathbf{X} V(\phi_{k0})\}\right\| \leqslant B_2,$$

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

We now establish the consistency and asymptotic normality of the proposed NPMLEs.

Theorem 1. *Under conditions (*C1*)–(*C5*),* $\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \to 0$, $\|\widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0\| \to 0$, *and* $\sup_{t \in \mathcal{Y}_k} |\widetilde{F}_{n,k}(t) - F_{k0}(t)| \to 0$ $(k = 1, \ldots, K)$, *almost surely, where* $\|\cdot\|$ *is the Euclidean norm.*

Theorem 2. *Under conditions (*C1*)–(*C5*), the random element* $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0, \widetilde{\mathbf{F}}_n - \mathbf{F}_0)$ *converges weakly to a tight, zero-mean Gaussian process in the metric space* $l^\infty(\mathbb{R}^d \times \mathbb{R}^{K-1} \times \mathcal{H}^K)$, *where* $l^\infty(\mathcal{H})$ *is a linear space, equipped with the supremum norm, consisting of all bounded functions.*

Remark 6. Although it is not required that the support of $\mathbf{Z}$ contain $\mathbf{0}$ to prove Theorems 1 and 2, we impose this additional assumption such that $F_k(\cdot)$, for $k = 1, \ldots, K$, have meaningful interpretations. In practice, if a covariate takes positive values (for example, age), one can center the covariate at its sample mean. In this case, $F_k(\cdot)$ has the interpretation of the conditional CDF of the response variable for the kth group, given that the covariates take values at their means.

In addition to estimating the baseline distribution function in each stratum, we can estimate the asymptotic covariance matrix of the baseline distribution function estimator, together with that of the finite-dimensional parameter $\boldsymbol{\theta}$. We can regard the likelihood function (2.7) as a function of $(\boldsymbol{\beta}, \boldsymbol{\phi})$, and the parameters that represent the jump sizes of $F_k(\cdot)$ $(k = 1, \ldots, K)$ at distinct observed values. From the classical Fisher information theory of parametric models, the asymptotic covariance matrix in Theorem 2 can be estimated using the inverse of the observed Fisher information matrix jointly in all parameters $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{F})$. Let $(\mathbf{b}, \mathbf{c}) \in \mathbb{R}^d \times \mathbb{R}^{K-1}$ be any constant vector, and $\mathbf{h} = (h_1, \ldots, h_K) \in \mathcal{H}^K$ be any bounded K-function. The asymptotic variance of the random element $\mathbf{b}^{\top}\widetilde{\boldsymbol{\beta}}_n + \mathbf{c}^{\top}\widetilde{\boldsymbol{\phi}}_n + \sum_{k=1}^{K}\int_{\mathcal{Y}_k} h_k(t)d\widetilde{F}_{n,k}(t)$ is equal to that of $\mathbf{b}^{\top}\widetilde{\boldsymbol{\beta}}_n + \mathbf{c}^{\top}\widetilde{\boldsymbol{\phi}}_n + \sum_{k=1}^{K}\sum_{j=1}^{m_k} h_k(Y_{k(j)})\widetilde{p}_{kj}$. Therefore, it can be consistently estimated using $\mathbf{h}_n^{\top}\mathbf{J}_n^{-1}\mathbf{h}_n$, where $\mathbf{h}_n$ is the column vector consisting of $\mathbf{b}$, $\mathbf{c}$, and $h_k(Y_{k(j)}) - h_k(Y_{k(m_k)})$ $(j = 1, \ldots, m_k - 1; k = 1, \ldots, K)$, and $\mathbf{J}_n$ is the negative Hessian matrix of $\ell_n(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n)$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\phi})$ and the jump sizes $F_k\{Y_{k(j)}\}$ $(j = 1, \ldots, m_k - 1; k = 1, \ldots, K)$. The next theorem provides a theoretical justification for this result.

Theorem 3. *Let $V(\mathbf{b}, \mathbf{c}, \mathbf{h})$ be the asymptotic variance of the random element $\sqrt{n}[\mathbf{b}^{\top}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \mathbf{c}^{\top}(\widetilde{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) + \sum_{k=1}^{K}\int_{\mathcal{Y}_k} h_k(t)d\{\widetilde{F}_{n,k}(t) - F_{k0}(t)\}]$. Under conditions (C1)–(C5), the estimator $n\mathbf{h}_n^{\top}\mathbf{J}_n^{-1}\mathbf{h}_n \to V(\mathbf{b}, \mathbf{c}, \mathbf{h})$ uniformly in $(\mathbf{b}, \mathbf{c}, \mathbf{h})$ in probability.*

The proofs of Theorems 1–3 are provided in the Supplementary Material.

3. A Goodness-of-fit Test

An appealing feature of the proposed SDRM is that it can be used to check the density-ratio assumption. Recall that under the DRM, the nonparametric likelihood function of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{q})$ is given by

$$\mathcal{L}_n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{q}) = \prod_{i=1}^{n} \frac{q_j \exp\{Y_{(j)}(\boldsymbol{\alpha}^{\top}\mathbf{A}_i + \boldsymbol{\beta}^{\top}\mathbf{X}_i)\}}{\sum_{l=1}^{m} q_l \exp\{Y_{(l)}(\boldsymbol{\alpha}^{\top}\mathbf{A}_i + \boldsymbol{\beta}^{\top}\mathbf{X}_i)\}}, \tag{3.1}$$

where $Y_{(j)}$ is the jth-order statistic in the entire sample, and $q_j = F\{Y_{(j)}|\mathbf{Z} = \mathbf{0}\}$ $(j = 1, \ldots, m)$. Note that, similar to (2.6), the intermediate index j in (3.1) is related to i via $Y_i = Y_{(j)}$ $(j = 1, \ldots, m; i = 1, \ldots, n)$.

The NPMLEs of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{q})$, denoted by $(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{q}}_n)$, are obtained by solving $\nabla \log \mathcal{L}_n(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{q}}_n) = \mathbf{0}$ subject to the constraint $\sum_{j=1}^{m} q_j = 1$, and the NPMLE of the baseline distribution function $F(\cdot)$ is given by

$$\widehat{F}_n(t) = \sum_{j=1}^{m} \widehat{q}_j I\{Y_{(j)} \leqslant t\}.$$

Recall from Section 2 that $F_k(\cdot) = F(\cdot|\mathbf{X} = \mathbf{0}, A = k)$ is the distribution function conditional on $(\mathbf{X} = \mathbf{0}, A = k)$ and $\mathbf{F} = (F_1, \ldots, F_K)$. Under the DRM, the NPMLEs of $F_k(\cdot)$, denoted by $\widehat{F}_{n,k}(\cdot)$, for $k = 1, \ldots, K$, are given by

$$\widehat{F}_{n,k}(t) = \begin{cases} \sum_{j=1}^{m} \widehat{q}_j I\{Y_{(j)} \leq t\}, & k = K, \\ \sum_{j=1}^{m} \frac{\widehat{q}_j \exp(Y_{(j)}\widehat{\alpha}_k) I\{Y_{(j)} \leqslant t\}}{\sum_{l=1}^{m} \widehat{q}_l \exp(Y_{(l)}\widehat{\alpha}_k)}, & k = 1, \ldots, K-1. \end{cases}$$

Intuitively, if the density-ratio assumption is valid, we expect that $\widetilde{F}_{n,k}(\cdot)$ and $\widehat{F}_{n,k}(\cdot)$ to be close for all $k = 1, \ldots, K$. We then propose a Kolmogorov–Smirnov-type (KS) statistic,

$$\Delta_n = \sum_{k=1}^{K} \frac{n_k}{n} \Delta_{n,k}, \tag{3.2}$$

to test the density-ratio assumption, where

$$\Delta_{n,k} = \sup_{t \in \mathcal{Y}_k} \sqrt{n}\, |\Delta_{n,k}(t)| = \sup_{t \in \mathcal{Y}_k} \sqrt{n} \left| \widehat{F}_{n,k}(t) - \widetilde{F}_{n,k}(t) \right|$$

measures the maximum discrepancy between the estimated baseline distributions for the kth group based on the DRM and SDRM. A large value of Δ_n indicates a departure from the density-ratio assumption. The validity of the test is based on the following theorem.

Theorem 4. *Under the density-ratio assumption and regularity conditions (*C1*)–(*C5*), the stochastic process* $\sqrt{n}(\widehat{\mathbf{F}} - \widetilde{\mathbf{F}}) \rightsquigarrow \mathbf{W}$*, where* $\mathbf{W} = (W_1, \ldots, W_K)$ *is a zero-mean* K*-variate Gaussian process in the metric space* $l^{\infty}(\mathcal{H}^K)$*.*

Theorem 4 serves as the basis for justifying the proposed goodness-of-fit test. Let δ_p be the pth quantile of the asymptotic null distribution of the test statistic Δ_n defined in (3.2), that is, δ_p satisfies

$$\mathrm{P}\left(\sum_{k=1}^{K} \rho_k \left\{ \sup_{t \in \mathcal{Y}_k} |W_k(t)| \right\} \leqslant \delta_p \right) = p.$$

Because the supremum map is uniformly continuous in $l^{\infty}(\mathcal{H})$, according to the continuous mapping theorem (van der Vaart and Wellner (1996)), we have

$$\begin{aligned}\lim_{n\to\infty} \mathrm{P}(\Delta_n \geqslant \delta_{1-p}) &= \lim_{n\to\infty} \mathrm{P}\left(\sum_{k=1}^{K} \frac{n_k}{n} \left\{\sup_{t\in\mathcal{Y}_k} \sqrt{n}\left|\widehat{F}_{n,k}(t) - \widetilde{F}_{n,k}(t)\right|\right\} \geqslant \delta_{1-p}\right) \\ &= \mathrm{P}\left(\sum_{k=1}^{K} \rho_k \left\{\sup_{t\in\mathcal{Y}_k} |W_k(t)|\right\} \geqslant \delta_{1-p}\right) \\ &= p.\end{aligned}$$

Because there is no explicit analytic expression for the weighted suprema of the Gaussian processes, we propose a bootstrap procedure to approximate the p-value of the goodness-of-fit test. The algorithm proceeds as follows.

Step 1. Obtain the NPMLEs $(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n, \widehat{\mathbf{F}}_n)$ and $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\phi}}_n, \widetilde{\mathbf{F}}_n)$ under the DRM and SDRM, respectively. Calculate Δ_n^{obs} based on the observed sample of size n, and the NPMLEs thereof.

Step 2. For $i = 1, \ldots, n$, calculate the conditional distribution of the response based on $(Y_i, \mathbf{Z}_i)$, where $\mathbf{Z}_i = (\mathbf{X}_i^{\mathsf{T}}, \mathbf{A}_i^{\mathsf{T}})^{\mathsf{T}}$, under the DRM, as follows:

$$\widehat{P}_{ij} = \mathrm{P}(\cdot|\mathbf{X}_i, \mathbf{A}_i) = \frac{\widehat{q}_j \exp\{Y_{(j)}(\widehat{\boldsymbol{\alpha}}_n^{\mathsf{T}}\mathbf{A}_i + \widehat{\boldsymbol{\beta}}_n^{\mathsf{T}}\mathbf{X}_i)\}}{\sum_{l=1}^{m} \widehat{q}_l \exp\{Y_{(l)}(\widehat{\boldsymbol{\alpha}}_n^{\mathsf{T}}\mathbf{A}_i + \widehat{\boldsymbol{\beta}}_n^{\mathsf{T}}\mathbf{X}_i)\}} \quad (j = 1, \ldots, m).$$

Then, $\widehat{\mathbf{P}}(\cdot|\mathbf{Z}_i) \equiv \widehat{\mathbf{P}}_i = (\widehat{P}_{i1}, \ldots, \widehat{P}_{im})$ is the estimated probability distribution conditional on $\mathbf{Z}_i$ under the DRM.

Step 3. Generate Y_i^* according to the multinomial distribution $\widehat{\mathbf{P}}(\cdot|\mathbf{Z}_i)(i = 1, \ldots, n)$. Denote the generated random sample of size n as $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*)$.

Step 4. Obtain the NPMLEs $(\widehat{\boldsymbol{\alpha}}_n^*, \widehat{\boldsymbol{\beta}}_n^*, \widehat{\mathbf{F}}_n^*)$ and $(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\boldsymbol{\phi}}_n^*, \widetilde{\mathbf{F}}_n^*)$ based on $\{(Y_i^*, \mathbf{Z}_i), i = 1, \ldots, n\}$ under the DRM and SDRM, respectively. Calculate the test statistic Δ_n^* based on the generated sample and the corresponding NPMLEs.

Step 5. Repeat *Step 3* and *Step 4* B times, obtain $(\Delta_n^{*1}, \ldots, \Delta_n^{*B})$. The p-value is then approximated by

$$\widehat{p}_\Delta = \frac{1}{B} \sum_{b=1}^{B} I\{\Delta_n^{*b} \geqslant \Delta_n^{\mathrm{obs}}\}.$$

The null hypothesis, H_0 : the density-ratio assumption holds, is rejected at a prespecified significance level α if $\widehat{p}_\Delta < \alpha$.

4. Simulation Studies

In this section, we conduct simulation studies to assess the finite-sample performance of the proposed SDRM and the goodness-of-fit test procedure. Over the course of the simulation studies, we provide some insight into how the SDRM is comparable to the DRM when the density-ratio assumption is satisfied/violated.

We first consider the scenario under which the density-ratio assumption holds. We generate data from the model

$$Y_i|(\mathbf{X}_i, A_i) \sim \mathcal{N}\left(\alpha A_i + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i,\ 1\right),\ \ i = 1, \ldots, n,$$

where $\alpha = -0.2$, $\boldsymbol{\beta} = (0.5, -0.5, 0.5)^\mathsf{T}$, A is a binary variable with success probability 0.5, and $\mathbf{X}$ is a covariate vector with three components, X_1, X_2, and X_3, which are standard normal, uniform$(-1, 1)$, and Bernoulli(0.5) variables, respectively. We consider sample sizes 100, 200, and 400, and all simulation results are based on 1,000 replicates. The confidence intervals are constructed based on the normal approximation, where the corresponding standard errors are estimated by inverting the observed Fisher information matrix jointly in all parameters.

The simulation results under the above settings are summarized in Table 1. The NPMLEs under the DRM and SDRM are comparable. Both have small biases; the standard error estimates agree well with the sampling standard deviations; and the 95% confidence intervals have correct coverage probabilities (CPs). As the sample size increases, the biases and standard deviations of the NPMLEs for both models decrease. As expected, the biases and standard error estimates under the DRM are smaller than those under the SDRM, because the DRM is the true model. The relative efficiency (RE), defined as the ratio of the mean squared error (MSE) of the estimator under the SDRM to that under the DRM, is only slightly greater than one, especially for the regression coefficients. This shows that the SDRM results in a limited loss of efficiency when the density-ratio assumption holds.

We next consider the scenario when the density-ratio assumption is violated. Specifically, we generate data from the model

$$Y_i|(\mathbf{X}_i, A_i) \sim \mathcal{N}\left(\alpha A_i + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i,\ \sigma^2(A)\right),\ \ i = 1, \ldots, n.$$

Here, $\sigma^2(A = 1) = 0.7^2$ and $\sigma^2(A = 0) = 1$. The remaining settings are the same as those in the first set of simulations. Because we set $V(\cdot)$ to be the exponential function $\exp(\cdot)$, the dispersion parameter ϕ is $\log(1.0/0.49) = 0.713$. In this case, the true model is the SDRM, whereas the density-ratio assumption in the DRM

Table 1. The scenario in which the density-ratio assumption holds.

	SDRM							DRM							
n	Par.	True	Bias	SE	SEE	95%CP	MSE	Par.	True	Bias	SE	SEE	95%CP	MSE	RE
100	α	-0.200	-	-	-	-	-	α	-0.200	-0.007	0.223	0.216	0.956	0.049	-
	β_1	0.500	0.046	0.140	0.134	0.947	0.022	β_1	0.500	0.043	0.139	0.133	0.954	0.021	1.037
	β_2	-0.500	-0.038	0.224	0.202	0.940	0.052	β_2	-0.500	-0.034	0.219	0.200	0.945	0.049	1.043
	β_3	0.500	0.043	0.232	0.228	0.947	0.056	β_3	0.500	0.040	0.231	0.227	0.949	0.055	1.017
	$F_1(-0.5)$	0.382	0.005	0.092	0.087	0.932	0.008	-	-	-	-	-	-	-	-
	$F_1(0.0)$	0.579	0.003	0.089	0.087	0.947	0.008	-	-	-	-	-	-	-	-
	$F_1(0.5)$	0.758	0.001	0.075	0.070	0.926	0.006	-	-	-	-	-	-	-	-
	$F_2(-0.5)$	0.309	0.008	0.086	0.083	0.930	0.007	$F(-0.5)$	0.309	0.006	0.074	0.072	0.934	0.005	1.364
	$F_2(0.0)$	0.500	0.008	0.092	0.089	0.932	0.008	$F(0.0)$	0.500	0.006	0.079	0.080	0.930	0.006	1.348
	$F_2(0.5)$	0.691	0.008	0.077	0.078	0.923	0.006	$F(0.5)$	0.691	0.005	0.069	0.071	0.942	0.005	1.252
200	α	-0.200	-	-	-	-	-	α	-0.200	-0.005	0.149	0.148	0.960	0.022	-
	β_1	0.500	0.025	0.092	0.091	0.947	0.009	β_1	0.500	0.024	0.091	0.090	0.948	0.009	1.028
	β_2	-0.500	-0.017	0.137	0.137	0.954	0.019	β_2	-0.500	-0.016	0.136	0.137	0.952	0.019	1.011
	β_3	0.500	0.024	0.154	0.156	0.955	0.024	β_3	0.500	0.022	0.153	0.155	0.955	0.024	1.012
	$F_1(-0.5)$	0.382	-0.001	0.059	0.061	0.951	0.004	-	-	-	-	-	-	-	-
	$F_1(0.0)$	0.579	0.001	0.060	0.061	0.956	0.004	-	-	-	-	-	-	-	-
	$F_1(0.5)$	0.758	0.002	0.049	0.049	0.940	0.002	-	-	-	-	-	-	-	-
	$F_2(-0.5)$	0.309	0.001	0.058	0.058	0.941	0.003	$F(-0.5)$	0.309	0.000	0.052	0.051	0.936	0.003	1.282
	$F_2(0.0)$	0.500	0.002	0.063	0.063	0.940	0.004	$F(0.0)$	0.500	0.002	0.056	0.056	0.944	0.003	1.230
	$F_2(0.5)$	0.691	0.004	0.056	0.055	0.940	0.003	$F(0.5)$	0.691	0.003	0.050	0.050	0.946	0.003	1.255

Par., parameter; Bias, difference between the average of parameter estimates and the true parameter value; SE, empirical standard deviation of the parameter estimates; SEE, average of the standard error estimates using the inverse of the Fisher information matrix; 95%CP, empirical coverage probability of the 95% confidence interval based on the normal approximation; MSE, mean squared error estimate, $=\text{Bias}^2 + \text{SE}^2$; RE, relative efficiency, =MSE(SDRM)/MSE(DRM).

Table 2. The scenario in which the density-ratio assumption is violated.

	SDRM							DRM							
n	Par.	True	Bias	SE	SEE	95%CP	MSE	Par.	True	Bias	SE	SEE	95%CP	MSE	RE
100	α	-0.200	-	-	-	-	-	α	-0.200	-0.089	0.255	0.251	0.955	0.073	-
	β_1	0.500	0.053	0.201	0.178	0.947	0.043	β_1	0.500	0.225	0.184	0.162	0.756	0.084	0.512
	β_2	-0.500	-0.050	0.238	0.220	0.946	0.059	β_2	-0.500	-0.221	0.255	0.239	0.879	0.114	0.518
	β_3	0.500	0.034	0.261	0.236	0.928	0.069	β_3	0.500	0.213	0.287	0.268	0.895	0.128	0.542
	ϕ	0.713	0.050	0.460	0.432	0.942	0.214	-	-	-	-	-	-	-	-
	$F_1(-0.5)$	0.334	-0.005	0.093	0.092	0.931	0.009	-	-	-	-	-	-	-	-
	$F_1(-0.2)$	0.500	-0.006	0.101	0.099	0.930	0.010	-	-	-	-	-	-	-	-
	$F_1(0.0)$	0.612	-0.001	0.097	0.095	0.931	0.009	-	-	-	-	-	-	-	-
	$F_1(0.3)$	0.762	0.002	0.082	0.078	0.917	0.007	-	-	-	-	-	-	-	-
	$F_2(-0.5)$	0.309	-0.004	0.083	0.080	0.924	0.007	$F(-0.5)$	0.309	-0.031	0.073	0.069	0.872	0.006	1.090
	$F_2(-0.2)$	0.421	-0.003	0.088	0.087	0.935	0.008	$F(-0.2)$	0.421	-0.014	0.083	0.078	0.923	0.007	1.095
	$F_2(0.0)$	0.500	-0.002	0.089	0.088	0.935	0.008	$F(0.0)$	0.500	0.002	0.084	0.080	0.926	0.007	1.136
	$F_2(0.3)$	0.618	0.001	0.085	0.084	0.931	0.007	$F(0.3)$	0.618	0.024	0.080	0.075	0.900	0.007	1.047
200	α	-0.200	-	-	-	-	-	α	-0.200	-0.080	0.178	0.171	0.919	0.038	-
	β_1	0.500	0.018	0.121	0.117	0.950	0.015	β_1	0.500	0.193	0.117	0.110	0.624	0.051	0.293
	β_2	-0.500	-0.021	0.157	0.145	0.946	0.025	β_2	-0.500	-0.198	0.177	0.163	0.790	0.071	0.357
	β_3	0.500	0.015	0.164	0.156	0.933	0.027	β_3	0.500	0.192	0.190	0.183	0.837	0.073	0.374
	ϕ	0.713	0.039	0.299	0.290	0.948	0.091	-	-	-	-	-	-	-	-
	$F_1(-0.5)$	0.334	-0.003	0.066	0.065	0.946	0.004	-	-	-	-	-	-	-	-
	$F_1(-0.2)$	0.500	-0.001	0.070	0.069	0.946	0.005	-	-	-	-	-	-	-	-
	$F_1(0.0)$	0.612	0.001	0.067	0.067	0.935	0.004	-	-	-	-	-	-	-	-
	$F_1(0.3)$	0.762	0.004	0.055	0.055	0.930	0.003	-	-	-	-	-	-	-	-
	$F_2(-0.5)$	0.309	0.000	0.058	0.057	0.936	0.003	$F(-0.5)$	0.309	-0.029	0.052	0.049	0.867	0.004	0.950
	$F_2(-0.2)$	0.421	-0.001	0.062	0.061	0.935	0.004	$F(-0.2)$	0.421	-0.012	0.058	0.055	0.922	0.003	1.119
	$F_2(0.0)$	0.500	0.000	0.062	0.062	0.941	0.004	$F(0.0)$	0.500	0.004	0.060	0.057	0.927	0.004	1.078
	$F_2(0.3)$	0.618	0.000	0.060	0.059	0.941	0.004	$F(0.3)$	0.618	0.025	0.057	0.053	0.887	0.004	0.950

Par., parameter; Bias, difference between the average of parameter estimates and the true parameter value; SE, empirical standard deviation of the parameter estimates; SEE, average of the standard error estimates using the inverse of the Fisher information matrix; 95%CP, empirical coverage probability of the 95% confidence interval based on the normal approximation; MSE, mean squared error estimate, $=\text{Bias}^2 + \text{SE}^2$; RE, relative efficiency, =MSE(SDRM)/MSE(DRM).

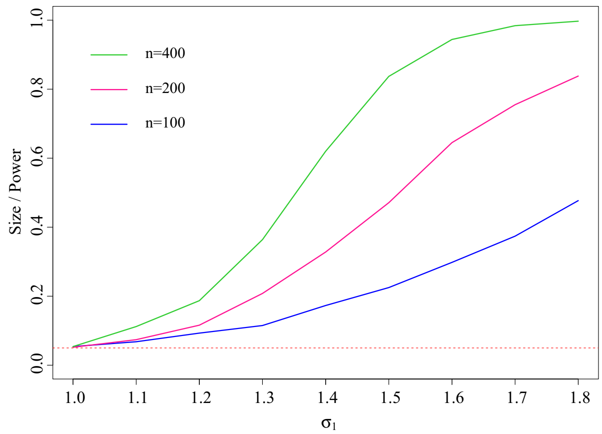

Figure 2. Size/power curve for testing the density-ratio assumption under the normal regression with covariate-dependent errors. The dashed horizontal line corresponds to the nominal 5% significance level.

is violated.

The simulation results under the above scenario are summarized in Table 2. The NPMLEs under the SDRM continue to perform well, with small biases and correct coverage probabilities of the 95% confidence intervals. On the other hand, the NPMLEs of the regression parameters under the DRM are very biased, and the REs decrease quickly as the sample size increases, such that the MSE is dominated by the bias. The biases of the estimators of the baseline distributions under the DRM appear to be much larger than those under the SDRM.

Finally, we evaluate the finite-sample performance of the proposed goodness-of-fit test for testing the density-ratio assumption in the standard DRM. The settings of the mean function of the normal distribution are as before. The variance parameter $\sigma_1 \equiv \sigma(A = 1)$ controls the effect size, and $\sigma(A = 0) \equiv 1$. The type-I error rates and statistical power are calculated based on 2,000 simulations, each with $B = 500$ bootstrap replicates. The size/power curve for the normal regression with covariate-dependent errors is plotted in Figure 2. The proposed goodness-of-fit test can control the type-I error rate accurately, and the power increases when the sample size or the effect size increases.

In addition to the normal regression with covariate-dependent errors, we conduct a power analysis when the baseline distribution is heterogeneous negative-binomial. See Example 2 in Section 1 for the parametrization of the baseline distribution. The linear predictor in the DRM form is $(\alpha A + \sum_{j=1}^{3} \beta_j X_j)$, where A is a binary variable with success probability 0.5; X_1, X_2, and X_3 are standard normal, uniform$(-1, 1)$, and Bernoulli(0.5) variables, respectively; and

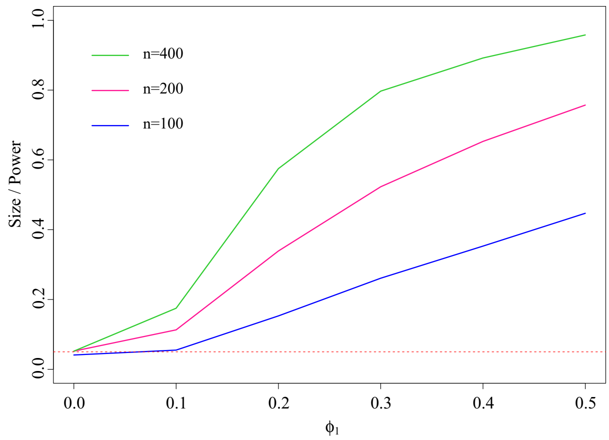

Figure 3. Size/power curve for testing the density-ratio assumption under the heterogeneous negative-binomial regression. The dashed horizontal line corresponds to the nominal 5% significance level.

$(\alpha, \beta_1, \beta_2, \beta_3) = (-1.0, -0.5, 0.5, -0.5)$. The baseline distribution of the response variable is a Poisson distribution with mean $\mu = 3$ conditional on $A = 0$. The dispersion parameter ϕ_1 controls the effect size. The type-I error rates and statistical power are calculated based on 2,000 simulations, each with $B = 500$ bootstrap replicates. The size/power curves are displayed in Figure 3. The proposed goodness-of-fit test can still control the type-I error rate accurately. The test, however, is much less powerful compared to that of the normal regression with covariate-dependent errors. One possible reason for such a dramatic decrease in power is that the response variable follows a discrete probability distribution, of which the true distribution function is a step function.

5. Applications

We first apply the proposed methodology to a blood lead concentration data set. A sample of $n = 100$ observations is available in Fitzmaurice, Laird and Ware (2012) and will be used in our data analysis. In randomized clinical trials, it is common practice to consider the pre-treatment score as a covariate to be adjusted for the post-treatment score, where the difference between the group-wise intercepts may quantify the treatment effect (Crager (1987)). In practice, people typically assume normality. Two issues have long been recognized. First, both pre- and post-treatment scores are random rather than fixed values, which may have different covariance structures. Second, parallel slopes between treatment groups is a critical assumption in the classical ANCOVA model, albeit this has

Table 3. Estimated coefficients for the blood lead concentration data.

SDRM					DRM				
Var.	Coef.	Std. Err.	t	$P > \|t\|$	Var.	Coef.	Std. Err.	t	$P > \|t\|$
pre.trt	4.311	0.993	4.337	< 0.001	pre.trt	1.099	0.225	4.882	< 0.001
ϕ	-2.134	0.453	-4.710	< 0.001	succimer	-3.310	0.593	-5.579	< 0.001

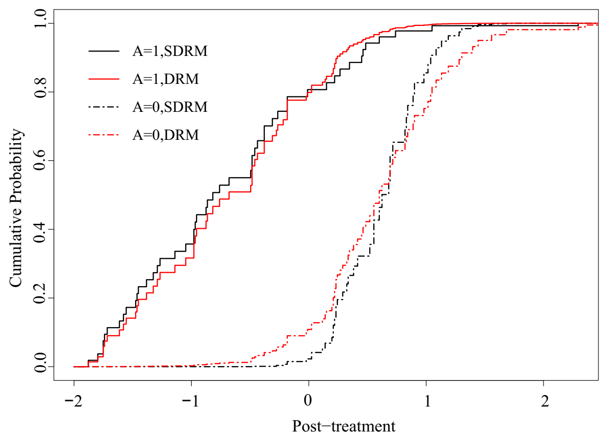

Figure 4. Estimated baseline CDFs for the blood lead concentration data.

been frequently questioned. Motivated by this data set, Funatogawa, Funatogawa and Shyr (2011) studied the type-I error rate of the ANCOVA model under equal-slope, but different covariances, without assuming normality. We now show how the proposed SDRM can be applied in this situation.

Let A be the succimer group indicator, and let *pre.trt* and *post.trt* be the pre- and post-treatment scores, respectively. We standardize these scores for comparison purposes. The fitted linear predictor in the SDRM is $4.311 pre.trt \times e^{-2.134}$, and that in the DRM is $-3.310A+1.099 pre.trt$. The estimated coefficients are all significant. Detailed estimation results can be found in Table 3, and the estimated baseline distribution functions (CDFs) based on the SDRM and DRM are plotted in Figure 4. The proposed goodness-of-fit test is significant, with a p-value of 0.023 based on 2,000 bootstrap samples, and the homogeneity test based on the SDRM is also significant (p-value < 0.001). In other words, the density-ratio assumption is rejected, and the two estimated baseline CDFs based on the SDRM are significantly different.

Although the estimated baseline CDFs based on the DRM and SDRM do not differ substantially in Figure 4, the proposed tests can still detect a significant difference (all give consistent results). This finding demonstrates that when the normality assumption indeed holds, the proposed semiparametric procedure is fairly powerful in distinguishing heterogeneity among the baselines. In addition, both the parametric and semiparametric procedures validate the parallel-slope assumption (with a p-value of 0.194 for testing the interaction effect in the ANCOVA model, and a p-value < 0.001 for testing $\phi = 0$ in the SDRM).

We end this case analysis with a final remark. The quantity $T \equiv \mathrm{E}_{F_1}(post.trt| pre.trt = 0) - \mathrm{E}_{F_0}(post.trt|pre.trt = 0)$ represents the treatment effect incorporating the random subject/group effects (see (2.2) for the conditional mean function). Replacing the unknown parameters in T with $(\widetilde{\beta}, \widetilde{\phi}, \widetilde{F}_1, \widetilde{F}_0)$ based on the SDRM gives the NPMLE of the treatment effect $\widetilde{T}$. We have $\widetilde{T} = -1.317$, which is very close to the value of -1.308 (p-value < 0.001) obtained from the ANCOVA model. Because we have the conditional distribution estimates, in addition to the mean difference, other summary statistics involving β and/or F_0 and F_1, such as the median treatment effect, can also be conveniently calculated using plug-ins.

The second real-data analysis is provided in the Supplementary Material.

6. Discussion

We have proposed a KS-type test statistic (3.2) and a goodness-of-fit procedure. Alternatively, we may consider the so-called Cramér–von-Mises-type (CvM) statistic based on the (weighted) integrated quadratic distance, that is,

$$\Delta_n^{\mathrm{CvM}} = \sum_{k=1}^{K} \int_{\mathcal{Y}_k} w_k(t)\{\widehat{F}_{n,k}(t) - \widetilde{F}_{n,k}(t)\}^2 dt.$$

One may improve the efficiency of the test by choosing a proper weight function $w_k(t)$ $(k = 1, \ldots, K)$, taking into account both the within- and between-strata information, as well as the characteristics of the baseline distribution functions. We take $w_k(t) = n_k/n$, the benchmark sample proportion, as a simple adjustment for the between-strata sample sizes in the KS-type test described in Section 3. The optimality conditions of the weight function are beyond the scope of this study, but deserve to be investigated in future research.

The proposed stratified model and the goodness-of-fit test procedure have several limitations. First, the proposed KS-type test and its potential alternative, the CvM-type test, are not direct and formal diagnostic tests for the density-ratio assumption. The entire testing procedure relies on the stratification of a categorical covariate. On the one hand, searching for such a categorical covariate can be practically burdensome; on the other hand, it may be more reasonable to explain the (condition) variance based on a set of covariates, which is similar to the GLMs with varying dispersion (Smyth (1989)). Most importantly, if the density-ratio assumption is violated for a continuous covariate, then how to properly categorize such a continuous variable becomes more troublesome. Undoubtedly, a suitable stratification will entail a nice result with good interpretability. The second drawback we would like to address is the transformation applied to the variables used to fit the model. Typically, we use the log-transformation to a right-tailed variable naturally bounded from below by zero. This may be acceptable in practice; however, is not yet "formally" justified. Therefore, a formal validation procedure for the density-ratio assumption and the functional forms of the covariates certainly warrants future research. Lastly, we want to emphasize that the regression parameter $\boldsymbol{\beta}$ in the DRM/SDRM, in general, cannot be interpreted as the mean, in contrast to those in the GLMs. This is a major limitation of this model, owing to the "canonical form" $y\boldsymbol{\beta}^{\top}\mathbf{X}$ in the linear predictor. The SPGLMs (Rathouz and Gao, 2009; Huang and Rathouz, 2012; Huang, 2014) have a clear advantage in this regard.

The standard DRM typically assumes that observations are homogeneous within the overall population. This is, in general, untrue if there is a natural stochastic ordering in the response variable across different levels of a potential confounding categorical covariate. El Barmi and McKeague (2013) tested a stochastic ordering based on an integrated localized empirical likelihood ratio statistic. Combining the advantages of the empirical likelihood methodology and the benchmark likelihood ratio test, their approach is elegant and successful in k-sample data, the data structure of which is similar to that of the DRM with a single k-level categorical covariate. Their procedure, however, is not applicable to data with arbitrary covariates, in general. In a blood alcohol concentration (BAC) data set discussed by Ramírez and Vidakovic (2010), Chang (2014) stratified the ages of the drunk drivers into two levels, where below 30 were considered as young, and 30 and above were considered as old. The post-stratified distribution of the BAC of the young group was stochastically larger than that of the old group. The original BAC data set contains many other covariates, though Chang (2014) only considered the covariate *age*, which was stratified. Inspired by El Barmi and McKeague (2013) and Chang (2014), we think that it will be an interesting future research topic to consider the testing of stochastic orderings among the baseline CDFs in the SDRM formulated in this study.

Supplementary Material

The online Supplementary Material provides proofs for the theorems in Section 2 and Section 3, as well as the second real-data example (German health registry data) in Section 5.

Acknowledgments

The authors are grateful to the associate editor and two anonymous referees for their valuable comments and suggestions.

References

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Chan, K. C. G. (2012). Nuisance parameter elimination for proportional likelihood ratio models with nonignorable missingness and random truncation. *Biometrika* **100**, 269–276.

Chang, H. (2014). *Empirical Likelihood Tests for Stochastic Ordering Based on Censored and Biased Data*. Ph.D Dissertation. Columbia University.

Chen, K. (2001). Parametric models for response-biased sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**, 775–789.

Cox, D. R. (1972). Regression model and life-tables (with Discussion). *Journal of the Royal Statistical Society, Series B (Methodological)* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Crager, M. R. (1987). Analysis of covariance in parallel-group clinical trials with pretreatment baselines. *Biometrics* **43**, 895–901.

De Carvalho, M. and Davison, A. C. (2014). Spectral density ratio models for multivariate extremes. *Journal of the American Statistical Association* **109**, 764–776.

Diao, G., Ning, J. and Qin, J. (2012). Maximum likelihood estimation for semiparametric density ratio model. *The International Journal of Biostatistics* **8**.

El Barmi, H. and McKeague, I. W. (2013). Empirical likelihood-based tests for stochastic ordering. *Bernoulli* **19**, 295–307.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2012). *Applied Longitudinal Analysis*. 2nd Edition. Wiley, New Jersey.

Fokianos, K., Kedem, B., Qin, J. and Short, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics* **43**, 56–65.

Funatogawa, T., Funatogawa, I. and Shyr, Y. (2011). Analysis of covariance with pre-treatment measurements in randomized trials under the cases that covariances and post-treatment variances differ between groups. *Biometrical Journal* **53**, 512–524.

Fung, T. and Huang, A. (2016). Semiparametric generalized linear models for time-series data. *arXiv:1603.02802*.

Gilbert, P., Lele, S. and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to aids vaccine trials. *Biometrika* **86**, 27–43.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics* **16**, 1069–1112.

Hilbe, J. M. (2011). *Negative Binomial Regression*. 2nd Edition. Cambridge University Press.

Huang, A. (2014). Joint estimation of the mean and error distribution in generalized linear models. *Journal of the American Statistical Association* **109**, 186–196.

Huang, A. and Rathouz, P. J. (2012). Proportional likelihood ratio models for mean regression. *Biometrika* **99**, 223–229.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

Jørgensen, B. (1997). *The Theory of Dispersion Models.* CRC Press, London.

Kedem, B., Lu, G., Wei, R. and Williams, P. D. (2008). Forecasting mortality rates via density ratio modeling. *The Canadian Journal of Statistics* **36**, 193–206.

Liang, K. Y. and Qin, J. (2000). Regression analysis under non-standard situations: A pairwise pseudo-likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 773–786.

Luo, X. and Tsai, W. Y. (2012). A proportional likelihood ratio model. *Biometrika* **99**, 211–222.

Luo, X. and Tsai, W. Y. (2014). Moment-type estimators for the proportional likelihood ratio model with longitudinal data. *Biometrika* **102**, 121–134.

Marchese, S. and Diao, G. (2017). Density ratio model for multivariate outcomes. *Journal of Multivariate Analysis* **154**, 249–261.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd Edition. CRC Press, London.

Nelder, J. A. and Wedderburn, R. M. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135**, 370–384.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing.* 2nd Edition. Cambridge University Press, Cambridge.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **85**, 619–630.

Qin, J. and Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika* **84**, 609–618.

Ramírez, P. and Vidakovic, B. (2010). Wavelet density estimation for stratified size-biased sample. *Journal of Statistical Planning and Inference* **140**, 419–432.

Rathouz, P. and Gao, L. (2009). Generalized linear models with unspecified reference distribution. *Biostatistics* **10**, 205–218.

Schifano, E. D., Li, L., Christiani, D. C. and Lin, X. (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics* **92**, 744–759.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society: Series B (Methodological)* **51**, 47–60.

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 Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Vardi, Y. (1985). Empirical distribution in selection bias models. *The Annals of Statistics* **13**, 178–203.

Zhang, B. (2000). A goodness of fit test for multiplicative-intercept risk models based on case-control data. *Statistica Sinica* **10**, 839–865.

Zhang, B. (2002). Assessing goodness-of-fit of generalized logit models based on case-control data. *Journal of Multivariate Analysis* **82**, 17–38.

Zhu, H. (2014). Likelihood approaches for proportional likelihood ratio model with right-censored data. *Statistics in Medicine* **33**, 2467–2479.

Zhu, H., Ning, J., Shen, Y. and Qin, J. (2017). Semiparametric density ratio modeling of survival data from a prevalent cohort. *Biostatistics* **18**, 62–75.

Moming Li

Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, CA 94143, USA.

E-mail: moming.li@ucsf.edu

Guoqing Diao

Department of Biostatistics and Bioinformatics, George Washington University, Washington, DC 20052, USA.

E-mail: gdiao@gwu.edu

(Received February 2019; accepted June 2020)