AISM (2006) 58: 687–706
DOI 10.1007/s10463-006-0047-7

Zhiwei Zhang · Howard E. Rockette

Semiparametric maximum likelihood for missing covariates in parametric regression

Received: 19 November 2004 / Revised: 19 July 2005 / Published online: 11 May 2006

Abstract We consider parameter estimation in parametric regression models with covariates missing at random. This problem admits a semiparametric maximum likelihood approach which requires no parametric specification of the selection mechanism or the covariate distribution. The semiparametric maximum likelihood estimator (MLE) has been found to be consistent. We show here, for some specific models, that the semiparametric MLE converges weakly to a zero-mean Gaussian process in a suitable space. The regression parameter estimate, in particular, achieves the semiparametric information bound, which can be consistently estimated by perturbing the profile log-likelihood. Furthermore, the profile likelihood ratio statistic is asymptotically chi-squared. The techniques used here extend to other models.

Keywords Asymptotic normality · Efficiency · Infinite-dimensional M-estimation · Missing at random · Missing covariates · Parametric regression · Profile likelihood · Semiparametric likelihood

1 Introduction

Parametric regression models such as generalized linear models are commonly used to assess the effect of a vector X of covariates on an outcome variable Y. Under such a model, the conditional distribution of Y given X is known up to a Euclidean regression parameter θ. Suppose a random sample is taken from the

Z. Zhang (✉)
Division of Biostatistics, U.S. Food and Drug Administration,
1350 Piccard Drive, Rockville, MD 20850, USA
E-mail: zhiwei.zhang@fda.hhs.gov

H.E. Rockette
Department of Biostatistics, University of Pittsburgh,
130 DeSoto Street, Pittsburgh, PA 15261, USA

distribution of (X, Y). If the sample is fully observed, then θ is usually estimated by maximizing the likelihood. The maximum likelihood estimator (MLE) is consistent, asymptotically normal and efficient, under regularity conditions.

A challenge for the statistical analysis arises when X is unobserved for some subjects, either by design or by happenstance. Assume that X is missing at random (MAR) in the sense of Rubin (1976), that is, the conditional probability given (X, Y) that a subject is selected for full observation depends only on Y and not on X. Available methods to estimate θ from this data include complete case analysis, pseudolikelihood (Carroll and Wand, 1991; Pepe and Fleming, 1991), mean score (Reilly and Pepe, 1995), pseudoscore (Chatterjee et al. 2003), the method of Robins et al. (1995a) and maximum likelihood (Ibrahim et al. 1999).

Recently, Zhang and Rockette (2005a) proposed a semiparametric maximum likelihood method which extends the methods of Wild (1991), Roeder et al. (1996) and Lawless et al. (1999) for related problems. This approach requires no parametric specification of the selection mechanism or the covariate distribution. Sufficient conditions are given in Zhang and Rockette (2005a) for the existence and consistency of the semiparametric MLE. Here we show, for specific models, that the semiparametric MLE is asymptotically Gaussian and efficient. Furthermore, the profile likelihood for θ shares many properties with its parametric analogue. The semiparametric MLE can be implemented with an EM algorithm; see Zhang and Rockette (2005b) for computational details and numerical results.

The main results of this paper are deduced from the Z-theorem (van der Vaart and Wellner, 1996, theorem 3.3.1) and the profile likelihood theory (Murphy and van der Vaart 2000). The key arguments are parallel to those of van der Vaart (1994), van der Vaart and Wellner, (1996, example 3.3.10) and Murphy and van der Vaart (2001). The last three references essentially dealt with regression problems with covariates missing completely at random (MCAR), and the present paper can be viewed as an extension to the MAR situation. Even for the MCAR problem, it appears difficult to formulate a single set of conditions that cover most examples of interest. Likewise, we shall focus on specific models but indicate how different models may be treated by a similar argument. Despite this lack of generality, the results of this paper shed light on the large-sample performance of the semiparametric MLE and suggest practical inferential procedures.

The proposed semiparametric MLE is also considered in a recent, independent work of Chen (2004), who also proposed modeling strategies for more general patterns of missing covariates. Chen suggests using the profile likelihood for variance estimation, but offers no theoretical justification; a relevant result is established in the present paper. Chen's proof of asymptotic normality is remarkably concise, and here we hope to make the theory accessible to a broader audience of statisticians.

The rest of the paper is organized as follows. In Sect. 2, we formulate the problem, define the semiparametric MLE and review the key results of Zhang and Rockette (2005a). In Sect. 3, we calculate the efficient score for estimating θ with the covariate distribution unspecified. Then, in Sect. 4, a system of likelihood equations is constructed, which forms the basis for a linearization argument. In Sect. 5, we explore a quadratic expansion of the profile log-likelihood. Extensions to more general models are discussed in Sect. 6. Some technical details are omitted but can be found in a technical report.

2 Semiparametric MLE

Let X and Y be random vectors taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. The distribution of X is denoted by G and is unspecified. The conditional distribution of Y given $X = x$ is specified through $f(\cdot|x;\theta)$, a regular conditional density with respect to some fixed measure μ on $\mathcal{Y}$. Here f is a known function and θ is an unknown Euclidean regression parameter. Let (X_i, Y_i), $i = 1, \ldots, n$, be independent copies of (X, Y). If the (X_i, Y_i) are completely observed, G can be estimated by the empirical distribution of X and θ by the maximizer of $\prod_{i=1}^n f(Y_i|X_i;\theta)$. This can also be seen as the result of jointly maximizing the semiparametric likelihood

$$\prod_{i=1}^{n} f(Y_i|X_i;\theta)G\{X_i\}, \tag{1}$$

where $G\{x\} := G(\{x\})$. The MLE of (θ, G) is asymptotically efficient.

Now suppose that X is unobserved for some subjects. (More general patterns of missing covariates will be considered later in Sect. 6.2.) Let $R = 1$ if X is observed; 0 otherwise. We require that X be missing at random with a certain amount of observability. To be precise, assume that almost surely,

$$\varpi(Y) := E(R|Y) = E(R|X, Y) > 0, \tag{2}$$

$$\varpi^X(X) := E[\varpi(Y)|X] = E(R|X) \geq \delta > 0. \tag{3}$$

The function ϖ will be referred to as the selection mechanism. As before, assume that (X_i, Y_i, R_i), $i = 1, \ldots, n$, are independent copies of (X, Y, R). However, we only observe $(R_i X_i, Y_i, R_i)$, $i = 1, \ldots, n$.

Analogous to (1), a semiparametric likelihood for this reduced data may be defined as

$$L_n(\theta, G) = \prod_{i=1}^{n} [f(Y_i|X_i;\theta)G\{X_i\}]^{R_i} [f(Y_i; G, \theta)]^{1-R_i}, \tag{4}$$

where $f(y; G, \theta) := \int_{\mathcal{X}} f(y|x;\theta)dG(x)$. Note that θ and G are no longer "separated" from each other as they are in (1). Note also that (4) does not involve the selection mechanism, by the MAR assumption (2). Let $(\theta_0, G_0, \varpi_0)$ denote the true value of (θ, G, ϖ). It is natural to estimate (θ_0, G_0) by maximizing L_n over $\Theta \times \mathcal{G}$, where $\Theta \subset \mathbb{R}^d$ is the parameter set for θ and $\mathcal{G}$ the set of all probability measures on $\mathcal{X}$. This turns out to be asymptotically equivalent to a simpler maximization with the restriction that G be supported by the observed values of X (Zhang and Rockette, 2005a, theorem 10). Computationally, the global maximization is infinite-dimensional, whereas the restricted maximization is finite-dimensional. Therefore in this paper we focus on the restricted MLE:

$$(\hat{\theta}_n, \hat{G}_n) = \underset{(\theta,G):G\{X_i:R_i=1\}=1}{\arg\max} L_n(\theta, G),$$

although the arguments apply equally well to the true MLE.

Under (2), (3) and some regularity conditions on the regression model, Zhang and Rockette (2005a) show that almost surely, $\hat{\theta}_n \to \theta_0$ and $\|\hat{G}_n - G_0\|_{\mathcal{H}} \to 0$ for every $L_1(G_0)$-bounded Glivenko-Cantelli class $\mathcal{H}$. Here and in the sequel,

$$\|G\|_{\mathcal{H}} := \sup_{h\in\mathcal{H}} \left| \int h dG \right| \tag{5}$$

for a signed measure G and a class $\mathcal{H}$ of real-valued functions on $\mathcal{X}$, provided the integral exists for every $h \in \mathcal{H}$. This consistency result follows from a lengthy Wald-type argument, which requires a compact parameter set or a suitable compactification of it. The main difficulty arises from the point mass of G in the likelihood (4), for which we substitute an alternative expression. The proof relies heavily on the empirical process theory, and the regularity conditions (mostly integrability conditions) are driven by the desired Glivenko-Cantelli properties of certain classes of functions. The result of Zhang and Rockette (2005a) applies to several important models.

Example (Logistic regression). Suppose Y is a binary variable taking values in $\{0, 1\}$. Write $\theta = (\beta_0, \beta_1)$ and assume that

$$f(1|x;\theta) = P(Y = 1|X = x;\theta) = [1 + \exp(-\beta_0 - \beta_1^{\mathrm{T}} x)]^{-1}.$$

Under this model, the problem of covariates missing at random can also be viewed as one of two-phase sampling, studied by Wild (1991) and Lawless et al. (1999) among others. More generally, the two problems overlap when Y is finitely discrete. In that case the semiparametric MLE considered here is equivalent to the one for two-phase sampling. The semiparametric MLE for two-phase sampling is unavailable (without specifying the selection mechanism) in the next two examples.

Example (Linear regression). $\mathcal{Y} = \mathbb{R}$ and μ is Lebesgue measure. With $\theta = (\beta_0, \beta_1, \sigma)$, assume that

$$f(y|x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{(y - \beta_0 - \beta_1^{\mathrm{T}} x)^2}{-2\sigma^2}\right],$$

that is, given $X = x$, Y is normally distributed with mean $\beta_0 + \beta_1^{\mathrm{T}} x$ and variance σ^2.

Example (Poisson regression). $\mathcal{Y} = \{0\} \cup \mathbb{N}$ and μ is the counting measure. With $\theta = (\beta_0, \beta_1)$, assume that

$$f(y|x;\theta) = \exp[y(\beta_0 + \beta_1^{\mathrm{T}} x) - \exp(\beta_0 + \beta_1^{\mathrm{T}} x)]/y!,$$

that is, given $X = x$, Y follows a Poisson distribution with mean $\exp(\beta_0 + \beta_1^{\mathrm{T}} x)$.

In the next few sections we develop asymptotic distributional results for $(\hat{\theta}_n, \hat{G}_n)$. Although our strategy is quite general, different models may require different techniques, making it difficult to formulate a general result with simple conditions. To fix ideas, we shall work with specific models but indicate how different models may be handled. The first example above, a special case of the two-phase sampling

problem, is covered by the results of van der Vaart and Wellner (2001) and Breslow et al. (2003). We focus instead on the last two examples under the following assumptions:

$$\mathcal{X} \text{ is a compact interval in } \mathbb{R}, \tag{6}$$
$$G_0 \text{ is nondegenerate}, \tag{7}$$
$$\Theta \text{ is compact}, \tag{8}$$
$$\theta_0 \text{ is interior to } \Theta. \tag{9}$$

That X is one-dimensional is assumed for simplicity; higher dimensions can be treated similarly. In practice, covariates can often be considered as bounded and thus admit a compact support. Assumption (7) is necessary for identifiability. The compactness of Θ is assumed to ensure consistency and may be relaxed with a more sophisticated argument.

A linear/Poisson regression model along with assumptions (2), (3) and (6,7,8,9) is implicitly assumed in Sects. 3–5. Wherever the two models admit a unified treatment, general notations will be used to emphasize the main ideas and to facilitate future generalizations to other models. Where necessary, the two models will be treated separately.

3 Information calculation

In this semiparametric model with G and ϖ unspecified, Robins et al. (1995a) have derived an integral equation representation of the efficient score for θ, and Breslow et al. (2003) have noted a minor correction to their formula. Here we derive an alternative representation of the efficient score for θ which is convenient to use in our arguments.

In the present setting, a general observation can be written as

$$V = V(X, Y, R) = \begin{cases} (Y, R) & \text{if } R = 0, \\ (X, Y, R) & \text{if } R = 1, \end{cases}$$

which takes values in $(\mathcal{Y} \times \{0\}) \cup (\mathcal{X} \times \mathcal{Y} \times \{1\})$ and is distributed under (θ, G, ϖ) as $Q_{\theta,G,\varpi}$, defined by

$$Q_{\theta,G,\varpi}(B_Y \times \{0\}) = \int_{B_Y} (1 - \varpi(y)) f(y; G, \theta) d\mu(y),$$
$$Q_{\theta,G,\varpi}(B_{XY} \times \{1\}) = \iint_{B_{XY}} \varpi(y) f(y|x; \theta) dG(x) d\mu(y),$$

for Borel subsets B_Y, $B_{X,Y}$ of $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$, respectively. Define a measure λ on the same space by

$$\lambda(B_Y \times \{0\}) = \mu(B_Y),$$
$$\lambda(B_{XY} \times \{1\}) = (G_0 \times \mu)(B_{XY}).$$

With (G, ϖ) fixed at (G_0, ϖ_0), the class of probabilities Q_{θ,G_0,ϖ_0} is dominated by λ, with densities

$$\frac{dQ_{\theta,G_0,\varpi_0}}{d\lambda}(x, y, r) = [\varpi_0(y) f(y|x; \theta)]^r [(1-\varpi_0(y)) f(y; G_0, \theta)]^{1-r}. \quad (10)$$

The submodel $\theta \mapsto Q_{\theta,G_0,\varpi_0}$ is regular in the sense of Bickel et al. (1993, chapter 2) with score $\dot{\ell}_{\theta_0,G_0}$ at θ_0, where

$$\begin{aligned}
\dot{\ell}_{\theta,G}(x, y, r) &:= r\dot{\ell}^{XY}_{\theta}(x, y) + (1 - r)\dot{\ell}^{Y}_{\theta,G}(y),\\
\dot{\ell}^{XY}_{\theta}(x, y) &:= \partial \log f(y|x; \theta)/\partial\theta,\\
\dot{\ell}^{Y}_{\theta,G}(y) &:= \partial \log f(y; G, \theta)/\partial\theta = E_{\theta,G}[\dot{\ell}^{XY}_{\theta}(X, Y)|Y = y].
\end{aligned}$$

Let $P_{\theta,G}$ denote the distribution of (X, Y, R) under (θ, G, ϖ_0), $P^Y_{\theta,G}$ that of Y under (θ, G), and similarly for $P^{XY}_{\theta,G}$, etc. Define the operators $\Pi^Y_{\theta,G} : L_2(P_{\theta,G}) \to L_2(P^Y_{\theta,G})$, $\Pi^X_{\theta} : L_2(P_{\theta,G}) \to L_2(G)$ and $A_{\theta,G} : L_2(P_{\theta,G}) \to L_2(P_{\theta,G})$ by

$$\begin{aligned}
\Pi^Y_{\theta,G}h(y) &= E_{\theta,G,\varpi_0}[h(X, Y, R)|Y = y]\\
&= f(y; G, \theta)^{-1} \int [\varpi_0(y)h(x, y, 1)\\
&\qquad +(1 - \varpi_0(y))h(x, y, 0)] f(y|x; \theta)dG(x), \quad (11)\\
\Pi^X_{\theta}h(x) &= E_{\theta}[h(X, Y, R)|X = x]\\
&= \int [\varpi_0(y)h(x, y, 1)\\
&\qquad +(1 - \varpi_0(y))h(x, y, 0)] f(y|x; \theta)d\mu(y), \quad (12)\\
A_{\theta,G}h(x, y, r) &= rh(x, y, r) + (1 - r)\Pi^Y_{\theta,G}h(y).
\end{aligned}$$

It is easy to see that $\Pi^Y_{\theta,G}$, Π^X_{θ} and $A_{\theta,G}$ as Hilbert space operators are linear and continuous. In what follows we use the subscript 0 to denote either θ_0 or (θ_0, G_0), depending on the context. Thus $\dot{\ell}^Y_0 = \dot{\ell}^Y_{\theta_0,G_0} = \Pi^Y_{\theta_0,G_0}\dot{\ell}^{XY}_{\theta_0} = \Pi^Y_0\dot{\ell}^{XY}_0$ and $\dot{\ell}_0 = \dot{\ell}_{\theta_0,G_0} = A_{\theta_0,G_0}\dot{\ell}^{XY}_{\theta_0} = A_0\dot{\ell}^{XY}_0$, where the operators act componentwise. We also use operator notation for integrals, so that $P_0\dot{\ell}^{XY}_0 = P_0\dot{\ell}^Y_0 = P_0\dot{\ell}_0 = 0$.

The efficient score function for θ at θ_0 is given by $\dot{\ell}_0$ minus its projection into the tangent space Γ for (G, ϖ) at (G_0, ϖ_0). Recall that Γ is defined as the closed linear span of the set of scores at (G_0, ϖ_0) in all regular one-dimensional submodels passing through (G_0, ϖ_0) with θ fixed at θ_0. For reasons that will become clear later, it suffices to consider regular one-dimensional submodels passing through G_0 with (θ, ϖ) fixed at (θ_0, ϖ_0). Such a submodel can be constructed as follows. Let $h : \mathcal{X} \to \mathbb{R}$ be measurable, bounded and nonconstant under G_0. For a (small) real number t, define G_t by

$$\frac{dG_t}{dG_0} = 1 + t(h - G_0h).$$

Note that when $t = 0$, G_t as defined above is just G_0, so that no ambiguity arises. Clearly, with $|t|$ sufficiently small, G_t is a probability measure on $\mathcal{X}$. The probabilities $Q_{\theta_0, G_t, \varpi_0}$ are again dominated by λ, with densities

$$\frac{dQ_{\theta_0,G_t,\varpi_0}}{d\lambda}(x, y, r) = \{\varpi_0(y) f(y|x; \theta_0)[1 + t(h(x) - G_0 h)]\}^r \times [(1 - \varpi_0(y)) f(y; G_t, \theta_0)]^{1-r}. \quad (13)$$

The submodel $t \mapsto Q_{\theta_0, G_t, \varpi_0}$ is regular and has score $B_0 h - G_0 h$ at $t = 0$, where $B_0 = B_{\theta_0, G_0}$ and $B_{\theta,G}$ is the restriction of $A_{\theta,G}$ to $L_2(G)$. The collection of such functions h is dense in $L_2(G_0)$, and the map $h \mapsto B_0 h - G_0 h$ is L_2-continuous. In view of the closedness of Γ, we now have

$$\Gamma \supset \{B_0 h - G_0 h : h \in L_2(G_0)\} = B_0 L_2^0(G_0) = \text{rge}(B_0) \cap L_2^0(P_0), \quad (14)$$

where rge$(\cdot)$ denotes the range of an operator, $L_2^0(G_0) := \{h \in L_2(G_0) : G_0 h = 0\}$, and similarly for $L_2^0(P_0)$.

Let A_0^* and B_0^* denote the respective Hilbert adjoint operators of A_0 and B_0. Then $A_0^* : L_2(P_0) \to L_2(P_0)$ is given by

$$\begin{aligned} A_0^* h(x, y, r) &= r h(x, y, r) + E_0[(1 - R)h(X, Y, R)|Y = y] \\ &= r h(x, y, r) + f(y; G_0, \theta_0)^{-1} \\ &\quad \times \int (1 - \varpi_0(y)) h(z, y, 0) f(y|z; \theta_0) dG_0(z), \end{aligned} \quad (15)$$

and $B_0^* = \Pi_0^X A_0^*$.

Remark 3.1 It is readily verified that $B_0^* = \Pi_0^X A_0 = \Pi_0^X A_0 A_0 = B_0^* A_0$ on $L_2(P_0^{XY})$ and, in particular, $B_0^* = B_0^* B_0$ on $L_2(G_0)$.

Lemma 3.1 *(a) A_0 restricted to $L_2(P_0^{XY})$ is one-to-one. (b) $B_0^* B_0$ is coutinuously invertible. (c)* rge(B_0) *is closed in $L_2(P_0)$.*

Proof Let $A_0 h = 0$ with $h \in L_2(P_0^{XY})$. In random variable notation, this means $A_0 h(X, Y, R) = 0$ almost surely. It follows that $0 = R A_0 h(X, Y, R) = R h(X, Y)$ almost surely, whence $0 = E_0[R h(X, Y)|X, Y] = \varpi_0(Y) h(X, Y)$ almost surely. By assumption (2), this further implies that $h(X, Y) = 0$ almost surely, i.e., $h = 0$ in $L_2(P_0^{XY})$. Thus (a) is established; in particular, B_0 is one-to-one. It follows that the self-adjoint operator $B_0^* B_0$ is positive-definite, proving (b). In particular, rge$(B_0^*) = L_2(G_0)$, which is closed. By Theorem 4.14 of Rudin, (1973, page 96), this implies that rge(B_0) is closed in $L_2(P_0)$. □

Lemma 3.1 says that the projection of $\dot{\ell}_0$ into rge(B_0) exists; by a standard result in functional analysis, it is given by $B_0(B_0^* B_0)^{-1} B_0^* \dot{\ell}_0$. This can also be written as $B_0(B_0^* B_0)^{-1} B_0^* A_0 \dot{\ell}_0^{XY} = B_0(B_0^* B_0)^{-1} B_0^* \dot{\ell}_0^{XY}$ by Remark 3.1. It is verified that B_0, B_0^* and hence $B_0^* B_0$ and $(B_0^* B_0)^{-1}$ are all mean-preserving, i.e., $P_0 B_0 h = P_0 h$. Therefore $B_0(B_0^* B_0)^{-1} B_0^* \dot{\ell}_0$ is in $L_2^0(P_0)$ and is also the projection of $\dot{\ell}_0$ into the right side of (14) (an intersection of two closed subspaces). Denote

$$\dot{\ell}_e = \dot{\ell}_0 - B_0(B_0^* B_0)^{-1} B_0^* \dot{\ell}_0.$$

It is not yet clear that $\dot{\ell}_e$ is the efficient score for θ. The right side of (14) is in general smaller than Γ, and $I_e := P_0(\dot{\ell}_e \dot{\ell}_e^{\mathrm{T}})$ is larger than the efficient Fisher information in the sense of nonnegative definiteness. If, however, we can demonstrate the existence of a regular estimator of θ with asymptotic variance I_e^{-1}, then $\dot{\ell}_e$ must be the efficient score and I_e the efficient information (Bickel et al. 1993, page 76–77). This will be done for the semiparametric MLE in the next section.

4 Asymptotic normality and efficiency

In this section we show that $(\hat{\theta}_n, \hat{G}_n)$ is asymptotically normal and that $\hat{\theta}_n$ achieves the semiparametric information bound. These results are deduced from the infinite-dimensional Z-theorem (van der Vaart and Wellner, 1996, theorem 3.3.1). Under this approach, the parameter set $\Theta \times \mathcal{G}$ is identified with a subset of some Banach space to be specified later. In the present context, the Z-theorem may be stated as follows.

Z-Theorem *Let Ψ_n and Ψ be random maps and a fixed map, respectively, from $\Theta \times \mathcal{G}$ into a Banach space such that Ψ is Fréchet-differentiable at (θ_0, G_0) with a continuously invertible derivative $\dot{\Psi}$, that*

$$\begin{aligned} &\sqrt{n}(\Psi_n - \Psi)(\hat{\theta}_n, \hat{G}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0, G_0) \\ &\quad = o_p^*(1 + \sqrt{n}\|(\hat{\theta}_n - \theta_0, \hat{G}_n - G_0)\|), \end{aligned} \tag{16}$$

and that the sequence $\sqrt{n}(\Psi_n - \Psi)(\theta_0, G_0)$ converges weakly to a tight random element W. If $\Psi(\theta_0, G_0) = 0$, $\Psi_n(\hat{\theta}_n, \hat{G}_n) = o_p^*(n^{-1/2})$, and $\hat{\theta}_n$ converges in outer probability to θ_0, then

$$\sqrt{n}\dot{\Psi}(\hat{\theta}_n - \theta_0, \hat{G}_n - G_0) = -\sqrt{n}(\Psi_n - \Psi)(\theta_0, G_0) + o_p^*(1).$$

Consequently, $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{G}_n - G_0)$ converges weakly to $-\dot{\Psi}^{-1}W$.

In the rest of this section, we shall construct maps Ψ_n and Ψ, verify the desired properties and derive the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{G}_n - G_0)$ explicitly. These will draw on the arguments of van der Vaart and Wellner, (1996, example 3.3.10) and van der Vaart (1994).

Since $\hat{\theta}_n$ is strongly consistent for θ_0, which is interior to Θ, $\hat{\theta}_n$ eventually lies in the interior of Θ. This and the definition of $(\hat{\theta}_n, \hat{G}_n)$ as a maximizer together imply that

$$0 = \frac{\partial}{\partial\theta}\frac{1}{n}\log L_n(\theta, \hat{G}_n)\bigg|_{\theta=\hat{\theta}_n} = \mathbb{P}_n \dot{\ell}_{\hat{\theta}_n, \hat{G}_n} \tag{17}$$

for large n, almost surely, where $\mathbb{P}_n$ denotes the empirical distribution of (X, Y, R). A similar "differentiation" with respect to G can be carried out as in Sect. 3. For a bounded measurable function $h : \mathcal{X} \to \mathbb{R}$ and a small real number t, define a probability measure $\hat{G}_{n,t}$ by

$$\frac{d\hat{G}_{n,t}}{d\hat{G}_n} = 1 + t(h - \hat{G}_n h).$$

Then $\hat{G}_{n,0} = \hat{G}_n$ and $\hat{G}_{n,t}$ is concentrated on $\{X_i : R_i = 1\}$. By the definition of $\hat{G}_n$, the map $t \mapsto \log L_n(\hat{\theta}_n, \hat{G}_{n,t})$ is maximized at $t = 0$. Differentiating with respect to t and setting the derivative equal to 0 at $t = 0$, we obtain

$$\hat{G}_n h = \mathbb{P}_n B_{\hat{\theta}_n, \hat{G}_n} h, \tag{18}$$

in the operator notation defined in the last section.

We now define a system of likelihood equations. Let $\mathcal{H}$ be a uniformly bounded Glivenko-Cantelli class of real functions on $\mathcal{X}$, and let $\ell^\infty(\mathcal{H})$ denote the collection of bounded real functions on $\mathcal{H}$. The latter is a Banach space under the uniform norm:

$$\|T\|_{\mathcal{H}} = \sup_{h \in \mathcal{H}} |Th|, \quad T \in \ell^\infty(\mathcal{H}).$$

The product space $\mathbb{R}^d \times \ell^\infty(\mathcal{H})$ is a Banach space too, under the product norm:

$$\|(a, T)\| = |a| \vee \|T\|_{\mathcal{H}}, \quad a \in \mathbb{R}^d, \quad T \in \ell^\infty(\mathcal{H}),$$

where $\vee$ denotes maximum and $|\cdot|$ the Euclidean norm. Let the random maps $\Psi_n : \Theta \times \mathcal{G} \to \mathbb{R}^d \times \ell^\infty(\mathcal{H})$ be defined by $\Psi_n(\theta, G) = (\Psi_{n1}(\theta, G), \Psi_{n2}(\theta, G))$, where

$$\Psi_{n1}(\theta, G) = \mathbb{P}_n \dot{\ell}_{\theta, G} \quad \text{and} \quad \Psi_{n2}(\theta, G)h = \mathbb{P}_n B_{\theta, G} h - Gh.$$

Note that $B_{\theta,G}$ preserves boundedness, so that $\Psi_{n2}(\theta, G)$ is indeed in $\ell^\infty(\mathcal{H})$. It follows from (17) and (18) that $\Psi_n(\hat{\theta}_n, \hat{G}_n) = 0$ for large n, almost surely.

For the two examples we consider, it seems convenient to take $\mathcal{H}$ to be the collection of real functions on $\mathcal{X}$ that are uniformly bounded by 1 and are Lipschitz with Lipschitz norm at most 1. This is a Donsker class and hence a Glivenko-Cantelli class (van der Vaart and Wellner, 1996, corollary 2.7.2), so that the consistency result described in Sect. 2 applies. With this choice of $\mathcal{H}$, $\|\cdot\|_{\mathcal{H}}$ defined in (5) generates the weak topology on $\mathcal{G}$ (van der Vaart and Wellner, 1996, theorem 1.12.4). For $h : \mathcal{X} \to \mathbb{R}$, define

$$\|h\|_1 = \sup_{x \in \mathcal{X}} |h(x)| \vee \sup_{x_1, x_2} \frac{|h(x_1) - h(x_2)|}{|x_1 - x_2|}.$$

Let $\mathcal{C}^1(\mathcal{X})$ denote the set of functions h with $\|h\|_1 < \infty$. Then $\mathcal{H}$ is the unit ball of the Banach space $\mathcal{C}^1(\mathcal{X})$ under the 1-norm.

Each $G \in \mathcal{G}$ defines an element of $\ell^\infty(\mathcal{H})$ by $h \mapsto Gh$. Furthermore, for the chosen $\mathcal{H}$ we have that $G_1 = G_2$ whenever $\|G_1 - G_2\|_{\mathcal{H}} = 0$ (van der Vaart and Wellner, 1996, lemma 1.3.12). Thus $\mathcal{G}$ is identified with a subset of $\ell^\infty(\mathcal{H})$, and the Ψ_n can be regarded as maps from $\mathbb{R}^d \times \ell^\infty(\mathcal{H})$ into itself whose domain is the product of Θ with the set of probability measures in $\ell^\infty(\mathcal{H})$ under the given identification.

The population version of Ψ_n is given by $\Psi = (\Psi_1, \Psi_2)$, where

$$\Psi_1(\theta, G) = P_0 \dot{\ell}_{\theta, G} \quad \text{and} \quad \Psi_2(\theta, G)h = P_0 B_{\theta, G} h - Gh. \tag{19}$$

Simple algebra shows that $\Psi(\theta_0, G_0) = 0$. Under the Z-theorem, Ψ is required to be Fréchet-differentiable at (θ_0, G_0), with derivative $\dot{\Psi}$ defined on the linear

span of $\Theta \times \mathcal{G} - (\theta_0, G_0)$. Heuristically, this can be seen as follows. First, for $(\theta, G) \approx (\theta_0, G_0)$,

$$\begin{aligned}\Psi_1(\theta, G) - \Psi_1(\theta_0, G_0) &= P_0(\dot{\ell}_{\theta,G} - \dot{\ell}_0) = P_0(\dot{\ell}_{\theta,G} - \dot{\ell}_{\theta_0,G}) + P_0(\dot{\ell}_{\theta_0,G} - \dot{\ell}_0) \\ &\approx P_0\ddot{\ell}_0(\theta - \theta_0) + \iint [1 - \varpi_0(y)][\dot{\ell}_0^{XY}(x, y) - \dot{\ell}_0^{Y}(y)] \\ &\quad \times f(y|x; \theta_0) d\mu(y) d(G - G_0)(x), \end{aligned} \tag{20}$$

where

$$\begin{aligned}\ddot{\ell}_{\theta,G}(x, y, r) &:= \frac{\partial}{\partial \theta^{\mathrm{T}}} \dot{\ell}_{\theta,G}(x, y, r) = r\ddot{\ell}_{\theta}^{XY}(x, y) + (1 - r)\ddot{\ell}_{\theta,G}^{Y}(y) \\ \ddot{\ell}_{\theta}^{XY}(x, y) &:= \frac{\partial}{\partial \theta^{\mathrm{T}}} \dot{\ell}_{\theta}^{XY}(x, y) = \frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} \log f(y|x; \theta), \\ \ddot{\ell}_{\theta,G}^{Y}(y) &:= \frac{\partial}{\partial \theta^{\mathrm{T}}} \dot{\ell}_{\theta,G}^{Y}(y) = \frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} \log f(y; G, \theta). \end{aligned}$$

In our (and many other) examples, the first term on the right side of (20) is equal to $-I_0(\theta - \theta_0)$, where $I_0 := P_0(\dot{\ell}_0 \dot{\ell}_0^{\mathrm{T}})$ is the Fisher information for θ when G is known to be G_0. In operator notation, the second term can be rewritten as $-(G - G_0)B_0^* \dot{\ell}_0^{XY}$. The derivative of the second component of Ψ can be obtained in a similar fashion. Uniformly over $h \in \mathcal{H}$,

$$\begin{aligned}\Psi_2(\theta, G)h - \Psi_2(\theta_0, G_0)h &= (P_0 - P_{\theta,G})B_{\theta,G}h = (P_0 - P_{\theta,G_0})B_{\theta,G}h \\ &\quad + (P_{\theta,G_0} - P_{\theta,G})B_{\theta,G}h \\ &\approx -(P_0 B_0 h \dot{\ell}_0^{\mathrm{T}})(\theta - \theta_0) - (G - G_0)B_0^* B_0 h. \end{aligned}$$

The foregoing discussion suggests that $\dot{\Psi}$ is given by the map

$$\begin{pmatrix} \theta - \theta_0 \\ G - G_0 \end{pmatrix} \mapsto \begin{pmatrix} \dot{\Psi}_{11} & \dot{\Psi}_{12} \\ \dot{\Psi}_{21} & \dot{\Psi}_{22} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ G - G_0 \end{pmatrix}, \tag{21}$$

where

$$\dot{\Psi}_{11}(\theta - \theta_0) = -I_0(\theta - \theta_0), \tag{22}$$

$$\dot{\Psi}_{12}(G - G_0) = -(G - G_0)B_0^* \dot{\ell}_0, \tag{23}$$

$$\dot{\Psi}_{21}(\theta - \theta_0)h = -(P_0 B_0 h \dot{\ell}_0^{\mathrm{T}})(\theta - \theta_0), \tag{24}$$

$$\dot{\Psi}_{22}(G - G_0)h = -(G - G_0)B_0^* B_0 h. \tag{25}$$

For this derivation to be valid, an intermediate set of sufficient conditions is given in Sect. 6, which can be verified for our specific examples.

It is apparent from the block form (21) of $\dot{\Psi}$ that the continuous invertibility of $\dot{\Psi}$ would follow from the same property of both $\dot{\Psi}_{11}$ and $\dot{\Phi} := \dot{\Psi}_{22} - \dot{\Psi}_{21}\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}$. Recall that $I_0 = P_0[A_0 \dot{\ell}_0^{XY}(A_0 \dot{\ell}_0^{XY})^{\mathrm{T}}]$. Lemma 3.1 and assumption (7) together then imply that I_0 is positive definite, hence invertible. The second operator has the form

$$\dot{\Phi}(G - G_0)h = (G - G_0)[(P_0 B_0 h \dot{\ell}_0^{\mathrm{T}}) I_0^{-1} B_0^* \dot{\ell}_0 - B_0^* B_0 h], \quad h \in \mathcal{H}, G \in \mathcal{G}.$$

$\dot{\Phi}$ is continuously invertible if and only if there exists $c > 0$ such that

$$\|\dot{\Phi}(G_1 - G_2)\|_{\mathcal{H}} \geq c\|G_1 - G_2\|_{\mathcal{H}}, \quad G_1, G_2 \in \mathcal{G}.$$

The latter certainly would follow from the existence of $c > 0$ such that

$$\left\{(P_0 B_0 h \dot{\ell}_0^{\mathrm{T}}) I_0^{-1} B_0^* \dot{\ell}_0 - B_0^* B_0 h : h \in \mathcal{H}\right\} \supset c\mathcal{H}. \tag{26}$$

Remark 4.1 Actually, objects like $B_0^* \dot{\ell}_0$ and $B_0^* B_0 h$ are originally defined as vectors in $L_2(G_0)$ and therefore represent equivalence classes of functions. Rather than redefine these operators, we shall simply take the "natural" versions given by (11), (12) and (15). So it is understood that

$$\begin{aligned} B_0^* \dot{\ell}_0(x) &= B_0^* \dot{\ell}_0^{XY}(x) \\ &= \int_{\mathcal{Y}} \left[\varpi_0(y) \dot{\ell}_0^{XY}(x, y) + (1 - \varpi_0(y)) \frac{\int_{\mathcal{X}} \dot{\ell}_0^{XY}(z, y) f(y|z; \theta_0) dG_0(z)}{f(y; G_0, \theta_0)} \right] \\ &\quad \times f(y|x; \theta_0) d\mu(y), \\ B_0^* B_0 h(x) &= B_0^* h(x) \\ &= \int_{\mathcal{Y}} \left[\varpi_0(y) h(x) + (1 - \varpi_0(y)) \frac{\int_{\mathcal{X}} h(z) f(y|z; \theta_0) dG_0(z)}{f(y; G_0, \theta_0)} \right] \\ &\quad \times f(y|x; \theta_0) d\mu(y). \end{aligned}$$

Clearly, (26) will hold if the operator C defined by

$$Ch = (P_0 B_0 h \dot{\ell}_0^{\mathrm{T}}) I_0^{-1} B_0^* \dot{\ell}_0 - B_0^* B_0 h \tag{27}$$

maps $\mathcal{C}^1(\mathcal{X})$ into itself and is continuously invertible. To this end we write $C = C_1 - C_2 - C_3$, with

$$\begin{aligned} C_1 h &= (P_0 B_0 h \dot{\ell}_0^{\mathrm{T}}) I_0^{-1} B_0^* \dot{\ell}_0, \\ C_2 h &= \varpi_0^X h, \\ C_3 h &= \Pi_0^X [(1 - \varpi_0) \Pi_0^Y h], \end{aligned}$$

and examine each component separately. (Recall that $\varpi_0^X(x) = \int \varpi_0(y) f(y|x; \theta_0) d\mu(y)$.)

Lemma 4.1 *(a) C_1 is compact. (b) C_2 is continuously invertible. (c) C_3 is compact.*

Proof (a) It is elementary to check that all components of $B_0^* \dot{\ell}_0$ are in $\mathcal{C}^1(\mathcal{X})$, so that C_1 indeed maps into $\mathcal{C}^1(\mathcal{X})$. It is linear and continuous, the latter due to the fact that $\mathcal{C}^1(\mathcal{X})$ has a norm stronger than the uniform norm. Furthermore, it has finite rank and therefore is compact.

(b) It can be shown that ϖ_0^X is Lipschitz and bounded. For every $h \in \mathcal{C}^1(\mathcal{X})$, we have

$$\left|\varpi_0^X(x_1) h(x_1) - \varpi_0^X(x_2) h(x_2)\right| \leq 2\left\|\varpi_0^X\right\|_1 \left\|h\right\|_1 |x_1 - x_2|, \quad x_1, x_2 \in \mathcal{X},$$

so that $\varpi_0^X h \in \mathcal{C}^1(\mathcal{X})$. Hence $C_2 : \mathcal{C}^1(\mathcal{X}) \to \mathcal{C}^1(\mathcal{X})$ is linear and continuous. By (3), C_2 is one-to-one and onto. By the inverse mapping theorem, it is continuously invertible.

(c) The map $h \in \mathcal{C}^1(\mathcal{X}) \mapsto (1-\varpi_0)\Pi_0^Y h \in \ell^\infty(\mathcal{Y})$ is linear and continuous because $\|\cdot\|_\infty \leq \|\cdot\|_1$. By Lemma 5.1 of van der Vaart (1994), Π_0^X maps $\ell^\infty(\mathcal{Y})$ into $\mathcal{C}^1(\mathcal{X})$ and is compact. As a composition, $C_3 : \mathcal{C}^1(\mathcal{X}) \to \mathcal{C}^1(\mathcal{X})$ is compact. □

The difference of two compact operators is again compact, so C is the difference of a compact operator and a continuously invertible one. By the theory of Fredholm operators, it is onto and has a continuous inverse if and only if it is one-to-one (Rudin, 1973, pages 99–103). The latter in fact follows from the positivity of the efficient information I_e, which in turn follows from (7).

Lemma 4.2 *(a) I_e is positive definite. (b) $C : \mathcal{C}^1(\mathcal{X}) \to \mathcal{C}^1(\mathcal{X})$ is one-to-one.*

Proof (a) Let $a^{\mathrm{T}} I_e a = 0$. Then, almost surely, $0 = a^{\mathrm{T}}\dot{\ell}_e = A_0[a^{\mathrm{T}}\dot{\ell}_0^{XY} - (B_0^* B_0)^{-1} B_0^*(a^{\mathrm{T}}\dot{\ell}_0^{XY})]$. By Lemma 3.1, A_0 is one-to-one on $L_2(P_0^{XY})$, so that $a^{\mathrm{T}}\dot{\ell}_0^{XY} = (B_0^* B_0)^{-1} B_0^*(a^{\mathrm{T}}\dot{\ell}_0^{XY}) \in L_2(G_0)$. By (7), this implies $a = 0$.

(b) Let $h \in \mathcal{C}^1(\mathcal{X})$ be such that $Ch = 0$ in $\mathcal{C}^1(\mathcal{X})$ (i.e., pointwise). Simple algebraic manipulation then yields

$$(P_0 B_0 h\dot{\ell}_0^{\mathrm{T}})\{I_0^{-1} P_0[B_0(B_0^* B_0)^{-1} B_0^*\dot{\ell}_0\dot{\ell}_0^{\mathrm{T}}] - I\} = 0,$$

where I is the identity matrix. This can be written in a simpler form: $I_e I_0^{-1} P_0(B_0 h\dot{\ell}_0) = 0$. It follows from the positivity (invertibility) of I_e that $P_0(B_0 h\dot{\ell}_0) = 0$. Substituting this into the definition of C gives that $B_0^* B_0 h = 0$ pointwise. By Lemma 3.1, $h = 0$ G_0-almost everywhere, whence $\Pi_0^Y h = 0$ P_0^Y-almost everywhere. Since f is positive everywhere, $\mu \ll P_0^Y$ and we have that $\Pi_0^Y h = 0$ μ-almost everywhere. This shows that $C_3 h = 0$ pointwise, whence $C_2 h = 0$ pointwise. By the strict positivity of ϖ_0^X, $h = 0$ pointwise (i.e., in $\mathcal{C}^1(\mathcal{X})$). □

Remark 4.2 A slight rearrangement of the foregoing discussion yields that $B_0^* B_0 : \mathcal{C}^1(\mathcal{X}) \to \mathcal{C}^1(\mathcal{X})$ is continuously invertible.

It follows from this discussion that C, $\dot{\Phi}$ and hence $\dot{\Psi}$ are continuously invertible, with

$$\dot{\Phi}^{-1} T h = T C^{-1} h, \quad h \in \mathcal{H}, \quad T \in \mathrm{rge}(\dot{\Phi}),$$
$$\dot{\Psi}^{-1} = \begin{pmatrix} \dot{\Psi}_{11}^{-1}(\dot{\Psi}_{11} + \dot{\Psi}_{12}\dot{\Phi}^{-1}\dot{\Psi}_{21})\dot{\Psi}_{11}^{-1} & -\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}\dot{\Phi}^{-1} \\ -\dot{\Phi}^{-1}\dot{\Psi}_{21}\dot{\Psi}_{11}^{-1} & \dot{\Phi}^{-1} \end{pmatrix}. \tag{28}$$

In order to apply the Z-theorem, it remains to verify the stochastic conditions (16) and the weak convergence of $\sqrt{n}(\Psi_n - \Psi)(\theta_0, G_0)$. In view of Lemma 3.3.5 of van der Vaart and Wellner (1996), it suffices to show that

$$P_0|\dot{\ell}_{\theta,G} - \dot{\ell}_0|^2 + \sup_{h\in\mathcal{H}} P_0[(B_{\theta,G} - B_0)h]^2 \to 0, \quad \|(\theta - \theta_0, G - G_0)\| \to 0 \tag{29}$$

and that the class of functions

$$\{\dot{\ell}_{\theta,G},\, B_{\theta,G}h - Gh : \|(\theta - \theta_0, G - G_0)\| < \tau, h \in \mathcal{H}\} \tag{30}$$

is Donsker for some $\tau > 0$. The uniform L_2-convergence (29) essentially follows from the dominated convergence theorem whereas the Donsker property of (30) depends on special structures of the model. Both are verified for the two models we consider.

With all conditions of the Z-theorem satisfied, we have the following asymptotic representation. Write $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$.

Theorem 4.3 *We have*

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_n - \theta_0 \\ \hat{G}_n - G_0\end{pmatrix} = \mathbb{G}_n\begin{pmatrix} I_e^{-1}\dot{\ell}_e \\ B_0C^{-1}h - (P_0B_0C^{-1}h\dot{\ell}_0^{\mathrm{T}})I_0^{-1}\dot{\ell}_0 : h \in \mathcal{H}\end{pmatrix} + o_p^*(1). \tag{31}$$

In particular, $\hat{\theta}_n$ is asymptotically efficient.

Proof It follows from the Z-theorem and the continuous mapping theorem that

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_n - \theta_0 \\ \hat{G}_n - G_0\end{pmatrix} = -\dot{\Psi}^{-1}\mathbb{G}_n\begin{pmatrix}\dot{\ell}_0 \\ B_0h - G_0h : h \in \mathcal{H}\end{pmatrix} + o_p^*(1).$$

A term-by-term examination shows that $\dot{\Psi}^{-1}$ and $\mathbb{G}_n$ can be interchanged in the above display. A direct application of (28) then gives (31), a key observation being that

$$C^{-1}(B_0^*\dot{\ell}_0) = I_0I_e^{-1}(B_0^*B_0)^{-1}B_0^*\dot{\ell}_0,$$

which is easily verified. □

5 Profile likelihood

We now explore a different approach to the analysis of $\hat{\theta}_n$ (not $\hat{G}_n$). This is based on a quadratic expansion of the profile log-likelihood for θ near θ_0, established by Murphy and van der Vaart (2000) for a general semiparametric model. Aside from the asymptotic normality of $\hat{\theta}_n$, the results of this section yield a consistent estimate of I_e and a profile likelihood ratio test.

At the core of this approach is a well-behaved least favorable submodel. In what follows we propose a candidate submodel and verify that it satisfies the conditions imposed by Murphy and van der Vaart (2000). For $G \in \mathcal{G}$ and θ, t in a neighborhood of θ_0, define $G_t(\theta, G)$ by

$$\frac{dG_t(\theta, G)}{dG} = 1 - (t - \theta)^{\mathrm{T}}h_G,$$

where $h_G := h_0 - Gh_0$, and $h_0 := (B_0^*B_0)^{-1}B_0^*\dot{\ell}_0$ is the least favorable direction for estimation of θ at (θ_0, G_0). From the last section, all components of h_0 are in $\mathcal{C}^1(\mathcal{X})$; in particular, they are bounded, so that $G_t(\theta, G) \in \mathcal{G}$ for $|t - \theta|$ sufficiently small. A parametric submodel can then be defined by $t \mapsto (t, G_t(\theta, G))$ (for θ, t in a small neighborhood of θ_0). This submodel clearly passes through (θ, G) at $t = \theta$:

$$G_\theta(\theta, G) = G, \quad \text{every } (\theta, G).$$

Thus condition (8) of Murphy and van der Vaart (2000) is met. Under this submodel, the log-density of V with respect to some dominating measure is, up to a constant,

$$\begin{aligned}
&\ell_{t,\theta,G}(x, y, r) : \\
&\quad = \log \left\{ \left[f(y|x; t) \frac{dG_t(\theta, G)}{dG}(x) \right]^r \left[\int f(y|u; t) dG_t(\theta, G)(u) \right]^{1-r} \right\} \\
&\quad = r \log f(y|x; t) + r \log[1 - (t - \theta)^{\mathrm{T}} h_G(x)] \\
&\qquad + (1 - r) \log \int f(y|u; t)[1 - (t - \theta)^{\mathrm{T}} h_G(u)] dG(u). \qquad (32)
\end{aligned}$$

Remark 5.1 This does not correspond exactly to the semiparametric likelihood (expression (4)) we use, as no point mass appears in the above display. Adding the term $r \log G\{x\}$ to the right side would make an exact correspondence with (4). However, the resulting function would be difficult, if not impossible, to work with, precisely because of the point mass. Inspection of the proof of Murphy and van der Vaart (2000)'s Theorem 1 reveals that, in connection with the likelihood, one may take (in their notation) $l(t, \theta, \eta) = \log 1(t, \eta_t(\theta, \eta)) + j(\theta, \eta)$ for any function j indexed by (θ, η) only. In particular, $\ell_{t,\theta,G}$ defined above is a legitimate choice, provided it satisfies the regularity conditions given in their theorem.

Differentiating (32) with respect to t gives

$$\dot{\ell}_{t,\theta,G} = \dot{\ell}_{t,G_t(\theta,G)} - B_{t,G_t(\theta,G)} h_{t,\theta,G}, \qquad (33)$$

where $h_{t,\theta,G} := h_G/[1-(t-\theta)^{\mathrm{T}} h_G]$. The formula for the second derivative $\ddot{\ell}_{t,\theta,G}$ is more complicated. Clearly, $\dot{\ell}_{\theta_0,\theta_0,G_0} = \dot{\ell}_e$, so that the submodel $t \mapsto (t, G_t(\theta_0, G_0))$ is least favorable for estimating θ at (θ_0, G_0) and condition (9) of Murphy and van der Vaart (2000) is satisfied. It is elementary to check the continuity, Glivenko-Cantelli and Donsker properties of the functions $\dot{\ell}_{t,\theta,G}$ and $\ddot{\ell}_{t,\theta,G}$ at and around $(\theta_0, \theta_0, G_0)$.

The profile likelihood for θ is given by

$$PL_n(\theta) := \max \left\{ L_n(\theta, G) : G \in \mathcal{G}, G\{X_i : R_i = 1\} = 1 \right\}.$$

Denote by $\hat{G}_n(\theta)$ any maximizer in the above display, so that $PL_n(\theta) = L_n(\theta, \hat{G}_n(\theta))$. Then

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} PL_n(\theta) \quad \text{and} \quad \hat{G}_n = \hat{G}_n(\hat{\theta}_n).$$

Condition (10) of Murphy and van der Vaart (2000) requires that $\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}} \xrightarrow{p} 0$ whenever $\bar{\theta}_n \xrightarrow{p} \theta_0$. This follows from an argument similar to the consistency proof for $(\hat{\theta}_n, \hat{G}_n)$. It only remains to verify condition (11), the no-bias condition. Let $\bar{\theta}_n \xrightarrow{p} \theta_0$. We need to show that

$$P_0 \dot{\ell}_{\theta_0,\bar{\theta}_n,\hat{G}_n(\bar{\theta}_n)} = o_p(|\bar{\theta}_n - \theta_0| + n^{-1/2}).$$

In light of the discussion of Murphy and van der Vaart (2000, page 457), this is equivalent to

$$P_0 \dot{\ell}_{\theta_0,\theta_0,\hat{G}_n(\bar{\theta}_n)} = o_p(|\bar{\theta}_n - \theta_0| + n^{-1/2}). \tag{34}$$

Write

$$\begin{aligned} P_0 \dot{\ell}_{\theta_0,\theta_0,\hat{G}_n(\bar{\theta}_n)} &= P_0(\dot{\ell}_{\theta_0,\hat{G}_n(\bar{\theta}_n)} - B_{\theta_0,\hat{G}_n(\bar{\theta}_n)} h_{\hat{G}_n(\bar{\theta}_n)}) \\ &= P_0(\dot{\ell}_{\theta_0,\hat{G}_n(\bar{\theta}_n)} - B_{\theta_0,\hat{G}_n(\bar{\theta}_n)} h_0 + \hat{G}_n(\bar{\theta}_n) h_0) \\ &= \Psi_1(\theta_0, \hat{G}_n(\bar{\theta}_n)) - \Psi_2(\theta_0, \hat{G}_n(\bar{\theta}_n)) h_0 \\ &= \xi \Psi(\theta_0, \hat{G}_n(\bar{\theta}_n)), \end{aligned} \tag{35}$$

where the first step follows from (33), the second from the definitions of h_G and $B_{\theta,G}$, the third from (19), and $\xi : \mathbb{R}^d \times \ell^\infty(\mathcal{H}) \to \mathbb{R}^d$ is defined by $\xi(a, T) = a - T h_0$. Since Ψ is Fréchet-differentiable at (θ_0, G_0),

$$\begin{aligned} \Psi(\theta_0, \hat{G}_n(\bar{\theta}_n)) &= \Psi(\theta_0, \hat{G}_n(\bar{\theta}_n)) - \Psi(\theta_0, G_0) \\ &= \dot{\Psi}(0, \hat{G}_n(\bar{\theta}_n) - G_0) + o_p(\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}}). \end{aligned}$$

With ξ linear and continuous, (35) now becomes

$$\begin{aligned} P_0 \dot{\ell}_{\theta_0,\theta_0,\hat{G}_n(\bar{\theta}_n)} &= \xi \dot{\Psi}(0, \hat{G}_n(\bar{\theta}_n) - G_0) + o_p(\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}}) \\ &= o_p(\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}}), \end{aligned} \tag{36}$$

where the second step is due to the fact that $\xi \dot{\Psi}(0, G - G_0) = 0$ for all G. In view of (36), (34) will follow as soon as

$$\hat{G}_n(\bar{\theta}_n) - G_0 = O_p(|\bar{\theta}_n - \theta_0| + n^{-1/2}). \tag{37}$$

To this end, note that (18) continues to hold with $(\hat{\theta}_n, \hat{G}_n)$ replaced by $(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n))$. In particular, for $h \in \mathcal{H}$,

$$\begin{aligned} \sqrt{n}(\hat{G}_n(\bar{\theta}_n) - G_0)h &= \sqrt{n}\mathbb{P}_n B_{\bar{\theta}_n,\hat{G}_n(\bar{\theta}_n)} h - \sqrt{n} P_0 B_0 h \\ &= \mathbb{G}_n B_{\bar{\theta}_n,\hat{G}_n(\bar{\theta}_n)} h + \sqrt{n} P_0 (B_{\bar{\theta}_n,\hat{G}_n(\bar{\theta}_n)} - B_0) h \\ &= \mathbb{G}_n B_{\bar{\theta}_n,\hat{G}_n(\bar{\theta}_n)} h + \sqrt{n}[\Psi_2(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)) \\ &\quad - \Psi_2(\theta_0, G_0)]h + \sqrt{n}(\hat{G}_n(\bar{\theta}_n) - G_0)h. \end{aligned} \tag{38}$$

Applying once again the differentiability of Ψ at (θ_0, G_0), we obtain

$$\begin{aligned} \sqrt{n}[\Psi_2(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)) - \Psi_2(\theta_0, G_0)]h = &-(P_0 B_0 h \dot{\ell}_0^{\mathrm{T}}) \sqrt{n}(\bar{\theta}_n - \theta_0) \\ &- \sqrt{n}(\hat{G}_n(\bar{\theta}_n) - G_0) B_0^* B_0 h \\ &+ o_p(\sqrt{n}|\bar{\theta}_n - \theta_0|) \\ &+ o_p(\sqrt{n}\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}}) \end{aligned} \tag{39}$$

uniformly in h. It follows from Remark 4.2 that

$$\{B_0^* B_0 h : h \in \mathcal{H}\} \supset c\mathcal{H} \tag{40}$$

for some $c > 0$. Combine (38, 39, 40), take suprema over h, and conclude that

$$\sqrt{n}\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}} \le c^{-1}\|\mathbb{G}_n B_{\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)}\|_{\mathcal{H}} + O_p(\sqrt{n}|\bar{\theta}_n - \theta_0|) + o_p(\sqrt{n}\|\hat{G}_n(\bar{\theta}_n) - G_0\|_{\mathcal{H}}).$$

From this (37) follows, provided the first term on the right is $O_p(1)$. The latter can be ascertained using the next lemma, which can be argued along the lines of van der Vaart (1998, lemma 19.24).

Lemma 5.1 *Let $\mathcal{H}$ be a set, $\mathcal{F} \subset L_2(P)$ a Donsker class, $B : \mathcal{H} \to \mathcal{F}$, and (B_m) a sequence of random maps such that* $\sup_{h\in\mathcal{H}} \|(B_m - B)h\|_{P,2} \xrightarrow{p} 0$. *Then* $\sup_{h\in\mathcal{H}} |\mathbb{G}_m(B_m - B)h| \xrightarrow{p} 0$.

The uniform L_2-continuity (29) and the weak consistency of $(\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n))$ together imply that

$$\sup_{h\in\mathcal{H}} \|(B_{\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)} - B_0)h\|_{P_0,2} \xrightarrow{p} 0.$$

In view of this and the Donsker property of (30), Lemma 5.1 yields

$$\mathbb{G}_n B_{\bar{\theta}_n, \hat{G}_n(\bar{\theta}_n)} = \mathbb{G}_n B_0 + o_p(1) = O_p(1)$$

in $\ell^\infty(\mathcal{H})$. This completes the verification of (37) and hence the no-bias condition.

Thus all conditions of Murphy and van der Vaart (2000)'s Theorem 1 and its corollaries have been established. In return, we have the following result. Write $pl_n(\theta) = \log PL_n(\theta)$.

Theorem 5.2 *For every sequence $\bar{\theta}_n \xrightarrow{p} \theta_0$, we have*

$$pl_n(\bar{\theta}_n) = pl_n(\theta_0) + n(\bar{\theta}_n - \theta_0)^{\mathrm{T}}\mathbb{P}_n\dot{\ell}_e - n(\bar{\theta}_n - \theta_0)^{\mathrm{T}} I_e(\bar{\theta}_n - \theta_0)/2 + o_p(\sqrt{n}\|\bar{\theta}_n - \theta_0\| + 1)^2.$$

In particular,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathbb{G}_n I_e^{-1}\dot{\ell}_e + o_p(1), \tag{41}$$

$$-2[pl_n(\hat{\theta}_n + u_n v_n) - pl_n(\hat{\theta}_n)]/(nu_n^2) = v^{\mathrm{T}} I_e v + o_p(1), \tag{42}$$

$$2[pl_n(\hat{\theta}_n) - pl_n(\theta_0)] = n(\hat{\theta}_n - \theta_0)^{\mathrm{T}} I_e(\hat{\theta}_n - \theta_0) + o_p(1), \tag{43}$$

for all sequences $v_n \xrightarrow{p} v \in \mathbb{R}^d$ and $u_n \xrightarrow{p} 0$ with $(\sqrt{n}u_n)^{-1} = O_p(1)$.

Expression (41) says that $\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, I_e^{-1})$, which is not new to us. What is new here is that, by (42), I_e can be consistently estimated by perturbing the profile log-likelihood for θ around $\hat{\theta}_n$. This makes possible Wald tests and related confidence statements about θ. Furthermore, (43) implies that the profile likelihood ratio statistic, like its parametric analogue, is asymptotically chi-squared with d degrees of freedom. This justifies testing hypotheses about θ using a profile likelihood ratio test and constructing confidence sets by inverting this test, just like in a parametric model.

6 Generalizations

6.1 Different regression models and higher-dimensional X

The main ideas of this paper apply to more general problems than the two examples studied here. However, it appears difficult to formulate a general theorem that covers most problems of interest, because different models may require different techniques. Some results are obtained more easily by appealing to special structures of the model. We now consider how the preceding discussion may be adapted to a different regression model or a higher-dimensional X.

Observe first that the information calculation in Sect. 3 generalizes easily. For a general regression model, one only needs to check that the submodels considered therein are regular and that the corresponding scores can be obtained by differentiating the log-density. Lemma 3.1 relies on assumption (2) but no special structures of the model.

The validity of the likelihood equations requires little more than differentiability in θ and holds quite generally. At this point, $\mathcal{H}$ can be any uniformly bounded Glivenko-Cantelli class such that Ψ_n and Ψ are well defined as maps on a subset of $\mathbb{R}^d \times \ell^\infty(\mathcal{H})$. The last condition can often be ascertained by using Lemma 1.3.12 of van der Vaart and Wellner (1996) if $\mathcal{H}$ consists of continuous functions.

Another consideration in choosing $\mathcal{H}$ is the differentiability of Ψ at (θ_0, G_0). The formulas we have derived appear reasonable, but have to be verified rigorously for a candidate $\mathcal{H}$. The intermediate set of sufficient conditions given below may be helpful.

Lemma 6.1 *Assume that* $P_0\ddot{\ell}_0^{XY} + \mathrm{var}(\dot{\ell}_0^{XY}) = 0$ *and that*

$$P_0\left|\dot{\ell}_\theta^{XY} - \dot{\ell}_0^{XY} - \ddot{\ell}_0^{XY}(\theta-\theta_0)\right| = o(|\theta-\theta_0|), \tag{44}$$

$$P_0\left|\dot{\ell}_{\theta,G}^{Y} - \dot{\ell}_{\theta_0,G}^{Y} - \ddot{\ell}_0^{Y}(\theta-\theta_0)\right| = o(|\theta-\theta_0|), \tag{45}$$

$$\iint [1-\varpi_0(y)](\dot{\ell}_{\theta_0,G}^{Y} - \dot{\ell}_0^{Y})(y) f(y|x;\theta_0) d\mu(y) d(G-G_0)(x) = o(\|G-G_0\|_{\mathcal{H}}), \tag{46}$$

$$\iint \left|f(y|x;\theta) - f(y|x;\theta_0) - \dot{f}(y|x;\theta_0)(\theta-\theta_0)\right| d\mu(y) dG_0(x) = o(|\theta-\theta_0|), \tag{47}$$

$$\sup_{h\in\mathcal{H}} P_0(B_{\theta,G_0}h - B_0h)^2 = o(1), \tag{48}$$

$$\sup_{h\in\mathcal{H}} \left|(G-G_0)(B_{\theta,G}^* B_{\theta,G} - B_0^* B_0)h\right| = o(\|G-G_0\|_{\mathcal{H}}), \tag{49}$$

as $\|(\theta-\theta_0, G-G_0)\| \to 0$. *Then* Ψ *is differentiable at* (θ_0, G_0), *with derivative* $\dot{\Psi}$ *given by* (21–25).

The proof is elementary and is omitted.

Our verification of the continuous invertibility of $\dot{\Psi}$ depends crucially on the fact that $\mathcal{H}$ is the unit ball of the Banach space $\mathcal{C}_1(\mathcal{X})$. In general, one may take $\mathcal{H}$ to be the unit ball of a Banach space $(\mathbb{B}, \|\cdot\|)$ contained in $\ell^\infty(\mathcal{X})$ with $\|\cdot\| \geq \|\cdot\|_\infty$. Then the continuous invertibility of $\dot{\Psi}$ will follow if C defined by (27) maps $\mathbb{B}$ into

itself and is continuously invertible. Examples of $\mathbb{B}$ include Hölder classes (van der Vaart and Wellner, 1996, section 2.7.1) and the space of bounded functions of bounded variation. The theory of Fredholm operators may again be useful here. Suppose that

$$\text{all components of } B_0^* \dot{\ell}_0 \text{ are in } \mathbb{B}, \tag{50}$$

$$B_0^* B_0 : \mathbb{B} \to \mathbb{B} \text{ is continuously invertible.} \tag{51}$$

Then $C : \mathbb{B} \to \mathbb{B}$ is linear and continuous. Furthermore, the first component of C has finite rank and hence is compact. Therefore C is continuously invertible if and only if it is one-to-one. The latter boils down to the positivity of the efficient information I_e, which is essentially equivalent to the condition that

$$a = 0 \text{ if } a^{\mathrm{T}} \dot{\ell}_0^{XY}(x, y) \text{ depends only on } x, \tag{52}$$

i.e., that the components of $\dot{\ell}_0^{XY}$ as vectors in the quotient space $L_2(P_0^{XY}) \big/ L_2(G_0)$ are linearly independent. This discussion is summarized as follows.

Lemma 6.2 *(a) If I_e is positive definite, then* (52) *holds. (b) Conversely,* (2) *and* (52) *together imply the positivity of I_e. (c) Under* (2), (51) *and* (52), *C is one-to-one on $\mathbb{B}$. (d) Under* (2) *and* (50, 51, 52), *$C : \mathbb{B} \to \mathbb{B}$ is continuously invertible.*

The proof of this consists of repetitions of previous arguments and is omitted.

In verifying the stochastic conditions, it will be convenient if $\mathcal{H}$ is a Donsker class. Thus if a Hölder class $\mathcal{C}^\alpha(\mathcal{X})$ is used as $\mathbb{B}$, one may wish to take α at least half of the dimension of X (van der Vaart and Wellner, 1996, corollary 2.7.2). This explains the choice of $\alpha = 1$ for our examples with one-dimensional covariates. The Donsker property of (30) is likely to depend on special structures of the model.

If these steps are all successful, one can then deduce the asymptotic normality result as in Theorem 4.3. Moreover, the arguments of Sect. 5, which are based on the same elementary facts, require minimal modifications.

6.2 General patterns of missing covariates

In practice, X need not be missing as a whole. Suppose only a portion of X can be missing. Write $X = (W, Z)$, where W is always observed and Z is possibly missing. Then the MAR assumption should be interpreted as $E(R|X, Y) = E(R|W, Y)$ almost surely. It is straightforward to extend the semiparametric MLE if W is finitely discrete, taking values in $\{w_j : j = 1, \ldots, k\}$, say. Redefine $G(\cdot|w_j)$ as the conditional distribution of Z given $W = w_j$. Then a semiparametric likelihood analogous to (4) can be written as

$$L(\theta, G) = \prod_{i=1}^{n} \Big[f(Y_i|W_i, Z_i; \theta) G(\{Z_i\}|W_i) \Big]^{R_i} \Big[\int f(Y_i|z, W_i; \theta) G(\mathrm{d}z|W_i) \Big]^{1-R_i},$$

and a semiparametric MLE can be obtained by maximizing this likelihood. We expect that an asymptotic theory for this will follow from the same arguments as in previous sections.

If W is not discrete or, more generally, if multiple patterns of missing covariates can occur, then it seems difficult to treat the covariate distribution completely nonparametrically within the maximum likelihood framework. Partial robustness may be achieved as in Chen (2004), where odds ratio models, rather than conditional distribution models, are specified among the components of X. It should be noted, however, that the interpretation of the MAR assumption can be subtle in the presence of arbitrary patterns of missing covariates. It may be of interest to consider alternative approaches such as estimating equations (e.g., Robins et al. 1994, 1995a,b).

Disclaimer The views expressed in this article are those of the authors and not necessarily of the United States Food and Drug Administration.

References

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., Wellner, J.A. (1993). *Efficient and adaptive estimation for semiparametric models*. Baltimore: Johns Hopkins University Press.

Breslow, N.E., McNeney, B., Wellner, J.A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Annals of Statistics 31*, 1110–1139.

Carroll, R.J., Wand, M.P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B 53*, 573–585.

Chatterjee, N., Chen, Y.H., Breslow, N.E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association 98*, 158–168.

Chen, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association 99*, 1176–1189.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics 55*, 591–596.

Lawless, J.F., Kalbfleisch, J.D., Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B 61*, 413–438.

Murphy, S.A., van der Vaart, A.W. (2000). On profile likelihood (with discussion). *Journal of the American Statistical Association 95*, 449–465.

Murphy, S.A., van der Vaart, A.W. (2001). Semiparametric mixtures in case-control studies. *Journal of Multivariate Analysis 79*, 1–32.

Pepe, M.S., Fleming, T.R. (1991). A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association 86*, 108–113.

Reilly, M., Pepe, M.S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika 82*, 299–314.

Robins, J.M., Rotnitzky, A., Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*, 846–866.

Robins, J.M., Hsieh, F., Newey, W. (1995a). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society, Series B 57*, 409–424.

Robins, J.M., Rotnitzky, A., Zhao, L.P. (1995b). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*, 106–121.

Roeder, K., Carroll, R.J., Lindsay, B.G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, *91*, 722–732.

Rubin, D.B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Rudin, W. (1973). *Functional analysis*. New York: McGraw-Hill.

van der Vaart, A.W. (1994). Maximum likelihood estimation with partially censored data. *Annals of Statistics 22*, 1896–1916.

van der Vaart, A.W. (1998). *Asymptotic statistics*. New York: Cambridge University Press.
van der Vaart, A.W., Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Berlin Heidelberg New York: Springer-Verlag.
van der Vaart, A.W., Wellner, J.A. (2001). Consistency of semiparametric maximum likelihood estimators for two-phase sampling. *Canadian Journal of Statistics 29*, 269–288.
Wild, C.J. (1991). Fitting prospective regression models to case-control data. *Biometrika 78*, 705–717.
Zhang, Z., Rockette, H.E. (2005a). On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference 134*, 206–223.
Zhang, Z., Rockette, H.E. (2005b). An EM algorithm for regression analysis with incomplete covariate information. *Journal of Statistical Computation and Simulation* (in press).