Efficient Estimation of Population-Level Summaries in General Semiparametric Regression Models

Arnab MAITY, Yanyuan MA, and Raymond J. CARROLL

This article considers a wide class of semiparametric regression models in which interest focuses on population-level quantities that combine both the parametric and the nonparametric parts of the model. Special cases in this approach include generalized partially linear models, generalized partially linear single-index models, structural measurement error models, and many others. For estimating the parametric part of the model efficiently, profile likelihood kernel estimation methods are well established in the literature. Here our focus is on estimating general population-level quantities that combine the parametric and nonparametric parts of the model (e.g., population mean, probabilities, etc.). We place this problem in a general context, provide a general kernel-based methodology, and derive the asymptotic distributions of estimates of these population-level quantities, showing that in many cases the estimates are semiparametric efficient. For estimating the population mean with no missing data, we show that the sample mean is semiparametric efficient for canonical exponential families, but not in general. We apply the methods to a problem in nutritional epidemiology, where estimating the distribution of usual intake is of primary interest and semiparametric methods are not available. Extensions to the case of missing response data are also discussed.

KEY WORDS: Generalized estimating equations; Kernel methods; Measurement error; Missing data; Nonparametric regression; Nutrition; Partially linear model; Profile method; Semiparametric efficient score; Semiparametric information bound; Single-index models.

1. INTRODUCTION

This article is about semiparametric regression models when one is interested in estimating a population quantity such as the mean, variance, and probabilities. The unique feature of the problem is that the quantities of interest are functions of both the parametric and the nonparametric parts of the model. We will also allow for partially missing responses, but handling such a modification is relatively easy. The main aim of the article is to estimate population quantities that involve both the parametric and the nonparametric parts of the model and to do so efficiently and in considerable generality.

We will construct estimators of these population-level quantities that exploit the semiparametric structure of the problem, derive their limiting distributions, and show in many cases that the methods are semiparametric efficient. The work is motivated by and illustrated with an important problem in nutritional epidemiology, namely, estimating the distribution of usual intake for episodically consumed foods such as red meat.

A special simple case of our results is already established in the literature (Wang, Linton, and Härdle 2004 and the references therein), namely, the partially linear model

$$Y_i = X_i^{\mathrm{T}}\beta_0 + \theta_0(Z_i) + \xi_i, \tag{1}$$

where $\theta_0(\cdot)$ is an unknown function and $\xi_i = \mathrm{Normal}(0, \sigma_0^2)$. We allow the responses to be partially missing, important in cases where the response is difficult to measure but the predictors are not. Suppose that Y is partially missing and let $\delta = 1$ indicate that Y is observed, so that the observed data are $(\delta_i Y_i, X_i, Z_i, \delta_i)$. Suppose further that Y is missing at random, so that $\mathrm{pr}(\delta = 1|Y, X, Z) = \mathrm{pr}(\delta = 1|X, Z)$.

Usually, of course, the main interest is in estimating β_0 efficiently. This is not the problem we discuss, because in our example the parameters β_0 are themselves of relatively minor interest. In their work, Wang et al. (2004) estimated the marginal mean $\kappa_0 = \mathrm{E}(Y) = \mathrm{E}\{X^{\mathrm{T}}\beta_0 + \theta_0(Z)\}$. Note how this combines both the parametric and the nonparametric parts of the model. One of the results of Wang et al. is that if one uses only the complete data that Y is observed, then fits the standard profile likelihood estimator to obtain $\widehat{\beta}$ and $\widehat{\theta}(\cdot, \widehat{\beta})$, it transpires that a semiparametric efficient estimator of the population mean κ_0 is $n^{-1}\sum_{i=1}^{n}\{X_i^{\mathrm{T}}\widehat{\beta} + \widehat{\theta}(Z_i, \widehat{\beta})\}$. If there are no missing data, the sample mean is also semiparametric efficient.

Actually, quite a bit more is true even in this relatively simple Gaussian case. Let $\mathcal{B} = (\beta^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$ and let $\widehat{\mathcal{B}}$ and $\widehat{\theta}(\cdot, \widehat{\mathcal{B}})$ be the profile likelihood estimates in the complete data; see, for example, Severini and Wong (1992) for local constant estimation and Claeskens and Carroll (2007) for local linear estimation. Consider estimating any functional $\kappa_0 = \mathrm{E}[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$ for some function $\mathcal{F}(\cdot)$ that is thrice continuously differentiable: This, of course, includes such quantities as population mean, and probabilities. Then one very special case of our results is that the semiparametric efficient estimate of κ_0 is just $\widehat{\kappa} = n^{-1}\sum_{i=1}^{n}\mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}$.

In contrast to Wang et al. (2004), we deal with general semiparametric models and general population-level quantities. Thus, consider a semiparametric problem in which the log-likelihood function given (X, Z) is $\mathcal{L}\{Y, X, \theta(Z), \mathcal{B}\}$. If we define $\mathcal{L}_{\mathcal{B}}(\cdot)$ and $\mathcal{L}_{\theta}(\cdot)$ to be derivatives of the log-likelihood with respect to $\mathcal{B}$ and $\theta(Z)$, we have the properties that $\mathrm{E}[\mathcal{L}_{\mathcal{B}}\{Y, X, \theta_0(Z), \mathcal{B}_0\}|X, Z] = 0$ and similarly for $\mathcal{L}_{\theta}(\cdot)$. We use profile likelihood methods computed at the observed data. With missing data, this local linear kernel version of the profile likelihood method of Severini and Wong (1992) works

Arnab Maity is a Graduate Student (E-mail: *amaity@stat.tamu.edu*), Yanyuan Ma is Assistant Professor (E-mail: *ma@stat.tamu.edu*), and Raymond J. Carroll is Distinguished Professor (E-mail: *carroll@stat.tamu.edu*), Department of Statistics, Texas A&M University, College Station, TX 77843. This work was supported by grants from the National Cancer Institute (CA57030 for AM and RJC; CA74552 for YM) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). The authors are grateful to Janet Tooze, Amy Subar, Victor Kipnis, and Douglas Midthune for introducing us to the problem of episodically consumed foods and for allowing us to use their data. The authors thank Naisyin Wang for reading the final manuscript and helping us with replies to a referee. Part of the original work of the last two authors originally occurred during a visit to the Centre of Excellence for Mathematics and Statistics of Complex Systems at the Australian National University, whose support they gratefully acknowledge. The authors especially wish to thank three referees, the associate editor, and the joint editor for helping turn the original submission into a publishable article. Their patience and many helpful suggestions are very greatly appreciated.

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Theory and Methods
DOI 10.1198/016214506000001103

as follows. Let $K(\cdot)$ be a smooth symmetric density function with bounded support, let h be a bandwidth, and let $K_h(z) = h^{-1}K(z/h)$. For any fixed $\mathcal{B}$, let $(\widehat{\alpha}_0, \widehat{\alpha}_1)$ be the local likelihood estimator obtained by maximizing, in (α_0, α_1),

$$\sum_{i=1}^{n} \delta_i K_h(Z_i - z)\mathcal{L}\{Y_i, X_i, \alpha_0 + \alpha_1(Z_i - z), \mathcal{B}\}, \tag{2}$$

and then setting $\widehat{\theta}(z, \mathcal{B}) = \widehat{\alpha}_0$. The profile likelihood estimator of $\mathcal{B}_0$ modified for missing responses is obtained by maximizing in $\mathcal{B}$

$$\sum_{i=1}^{n} \delta_i \mathcal{L}\{Y_i, X_i, \widehat{\theta}(Z_i, \mathcal{B}), \mathcal{B}\}. \tag{3}$$

Our estimator of $\kappa_0 = \mathrm{E}[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$ is then

$$\widehat{\kappa} = n^{-1} \sum_{i=1}^{n} \mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}. \tag{4}$$

We emphasize that the possibility of missing response data and finding a semiparametric efficient estimate of $\mathcal{B}_0$ is not the focus of the article. Instead, the focus is on estimating quantities $\kappa_0 = \mathrm{E}[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$ that depend on both the parametric and the nonparametric parts of the model: This is a very different problem than simply estimating $\mathcal{B}_0$. Previous work in the area considered only the partially linear model and only estimation of the population mean: Our work deals with general semiparametric models and general population-level quantities.

An outline of this article is as follows. In Section 2 we discuss the general semiparametric problem with log-likelihood $\mathcal{L}\{Y, X, \theta(Z), \mathcal{B}\}$ and a general goal of estimating $\kappa_0 = \mathrm{E}[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$. We derive the limiting distribution of (4) and show that it is semiparametric efficient. We also discuss the general problem where the population quantity κ_0 of interest is the expectation of a function of Y alone and describe doubly robust estimators in this context.

In Section 3 we consider the class of generalized partially linear single-index models (Carroll, Fan, Gijbels, and Wand 1997). Single-index modeling, see Härdle and Stoker (1989) and Härdle, Hall, and Ichimura (1993), is an important means of dimension reduction, one that is finding increased use in this age of high-dimensional data. We develop methods for estimating population quantities in the generalized partially linear single-index modeling framework and show that the methods are semiparametric efficient.

Section 4 describes an example from nutritional epidemiology that motivated this work, namely, estimating the distribution of usual intake of episodically consumed foods such as red meat. The model used in this area is far more complex than the simple partially linear Gaussian model (1), and while the population mean is of some interest, of considerably more interest is the probability that usual intake exceeds thresholds. We will illustrate why in this context one cannot simply adopt the percentages of the observed responses that exceed a certain threshold.

Section 5 describes three issues of importance: (1) bandwidth selection (Sec. 5.1), (2) the efficiency and robustness of the sample mean when the population mean is of interest (Sec. 5.2), and numerical and theoretical insights into the partially linear model and the nature of our assumptions (Sec. 5.3). An interesting special case is, of course, the partially linear model when κ_0 is the population mean. For this problem, we show in Section 5.2 that, with no missing data, the sample mean is semiparametric efficient for canonical exponential families but not of course in general, thus extending and clarifying the results of Wang et al. (2004) that were specific to the Gaussian case.

Section 6 gives concluding remarks and results. All technical results are given in the Appendix.

2. SEMIPARAMETRIC MODELS WITH A SINGLE COMPONENT

2.1 Main Results

We benefit from the fact that the limiting expansions for $\widehat{\mathcal{B}}$ and $\widehat{\theta}(\cdot)$ are essentially already well known, with the minor modification of incorporating the missing response indicators. Let $f(z)$ be the density function of Z, which is assumed to have bounded support and to be positive on that support. Let $\Omega(z) = f(z)\mathrm{E}\{\delta\mathcal{L}_{\theta\theta}(\cdot)|Z = z\}$. Let $\mathcal{L}_{i\theta}(\cdot) = \mathcal{L}_{\theta}\{Y_i, X_i, \theta_0(Z_i), \mathcal{B}_0\}$ and so on. Then it follows from standard results (see the App. for more discussion) that as a minor modification of the work of Severini and Wong (1992),

$$\begin{aligned}\widehat{\theta}(z, \widehat{\mathcal{B}}) - \theta_0(z) &= (h^2/2)\theta_0^{(2)}(z) - n^{-1}\sum_{i=1}^{n} \delta_i K_h(Z_i - z)\mathcal{L}_{i\theta}(\cdot)/\Omega(z) \\ &\quad + \theta_{\mathcal{B}}(z, \mathcal{B}_0)(\widehat{\mathcal{B}} - \mathcal{B}_0) + o_p(n^{-1/2}),\end{aligned} \tag{5}$$

$$\widehat{\mathcal{B}} - \mathcal{B}_0 = \mathcal{M}_1^{-1} n^{-1} \sum_{i=1}^{n} \delta_i \epsilon_i + o_p(n^{-1/2}), \tag{6}$$

where

$$\theta_{\mathcal{B}}(z, \mathcal{B}_0) = -\mathrm{E}\{\delta\mathcal{L}_{\mathcal{B}\theta}(\cdot)|Z = z\}/\mathrm{E}\{\delta\mathcal{L}_{\theta\theta}(\cdot)|Z = z\}, \tag{7}$$

$$\epsilon_i = \{\mathcal{L}_{i\mathcal{B}}(\cdot) + \mathcal{L}_{i\theta}(\cdot)\theta_{\mathcal{B}}(Z_i, \mathcal{B}_0)\}, \tag{8}$$

$$\mathcal{M}_1 = \mathrm{E}(\delta\epsilon\epsilon^{\mathrm{T}}) = -\mathrm{E}\big[\delta\{\mathcal{L}_{\mathcal{B}\mathcal{B}}(\cdot) + \mathcal{L}_{\mathcal{B}\theta}(\cdot)\theta_{\mathcal{B}}^{\mathrm{T}}(Z, \mathcal{B}_0)\}\big],$$

and where, under regularity conditions, (5) is uniform in z. Conditions guaranteeing (6) are well known; see the Appendix.

Define

$$D_i(\cdot) = -\mathcal{L}_{i\theta}(\cdot)\frac{\mathrm{E}\{\mathcal{F}_{\theta}(\cdot)|Z_i\}}{\mathrm{E}\{\delta\mathcal{L}_{\theta\theta}(\cdot)|Z_i\}},$$

$$\mathcal{M}_2 = \mathrm{E}\{\mathcal{F}_{\mathcal{B}}(\cdot) + \mathcal{F}_{\theta}(\cdot)\theta_{\mathcal{B}}(Z, \mathcal{B}_0)\}.$$

In the Appendix we show the following result.

Theorem 1. Suppose that $nh^4 \to 0$ and that (5) and (6) hold, the former uniformly in z. Suppose also that Z has compact support, that its density is bounded away from 0 on that support, and that the kernel function also has a finite support. Then the estimator $\widehat{\kappa}$ of $\kappa_0 = \mathrm{E}[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$ is semiparametric efficient in the sense of Newey (1990). In addition, as $n \to \infty$,

$$\begin{aligned}n^{1/2}(\widehat{\kappa} - \kappa_0) &= n^{-1/2}\sum_{i=1}^{n}\big\{\mathcal{F}_i(\cdot) - \kappa_0 + \mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\delta_i\epsilon_i \\ &\quad + \delta_i D_i(\cdot)\big\} + o_p(1)\end{aligned} \tag{9}$$

$$\Rightarrow \mathrm{Normal}\big[0, \mathrm{E}\{\mathcal{F}(\cdot) - \kappa_0\}^2 + \mathcal{M}_2^{\mathrm{T}} \mathcal{M}_1^{-1} \mathcal{M}_2 + \mathrm{E}\{\delta D^2(\cdot)\}\big]. \tag{10}$$

Remark 1. To obtain asymptotically correct inference about κ_0, there are two possible routes. The first is to use the bootstrap: Whereas Chen, Linton, and Van Keilegom (2003) only justified the bootstrap for estimating $\mathcal{B}_0$, we conjecture that the bootstrap works for κ_0 as well. More formally, one requires only a consistent estimate of the limiting variance in (10). This is a straightforward exercise, although programming intense: One merely replaces all the expectations by sums in that expression and all the regression functions by kernel estimates.

Remark 2. Our analysis of semiparametric efficiency in the sense of Newey (1990) has this outline. We first assume pathwise differentiability of κ; see Section A.2.2 for a definition. Working with this assumption, we derive the semiparametric efficient score. With this score in hand, we then prove pathwise differentiability. Details are given in the Appendix.

Remark 3. With a slight modification using a device introduced to semiparametric methods by Bickel (1982), Theorem 1 also holds for estimated bandwidths. We confine our discussion to bandwidths of order $n^{-1/3}$; see Section 5.1.2 for a reason. Write such bandwidths as $h_n = cn^{-1/3}$, where, following Bickel, the values for c are allowed to take values in the set $\mathcal{U} = a\{0, \pm 1, \pm 2, \ldots\}$, where a is an arbitrary small number. We discretize bandwidths so that they take on values $cn^{-1/3}$ with $c \in \mathcal{U}$. Denote estimators as $\widehat{\kappa}(h_n)$ and note that for an arbitrary c_* and an arbitrary fixed, deterministic sequence $c_n \to c_0$ for finite c_0, Theorem 1 shows that $n^{1/2}\{\widehat{\kappa}(c_n n^{-1/3}) - \widehat{\kappa}(c_0 n^{-1/3})\} = o_p(1)$ and that $n^{1/2}\{\widehat{\kappa}(c_0 n^{-1/3}) - \widehat{\kappa}(c_* n^{-1/3})\} = o_p(1)$. Hence, it follows from Bickel (1982, p. 653, just after eq. 3.7) that if $\widehat{h}_n = \widehat{c}_n n^{-1/3}$, with $\widehat{c} \in \mathcal{U}$, is an estimated bandwidth with the property that $\widehat{h}_n = O_p(n^{-1/3})$, then $n^{1/2}\{\widehat{\kappa}(\widehat{c}_n n^{-1/3}) - \widehat{\kappa}(c_* n^{-1/3})\} = o_p(1)$. Hence, Theorem 1 holds for these estimated bandwidths.

2.2 General Functions of the Response and Double Robustness

It is important to consider estimation in problems where κ_0 can be constructed outside the model. Suppose that $\kappa_0 = \mathrm{E}\{\mathcal{G}(Y)\}$ and define $\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\} = \mathrm{E}\{\mathcal{G}(Y)|X, Z\}$. We will discuss two estimators with the properties that (1) if there are no missing response data, the semiparametric model is not used and the estimator is consistent; and (2) under certain circumstances, the estimator is consistent if either the semiparametric model is correct or if a model for the missing-data process is correct.

Our motivating example discussed in Section 4 dose not fall into the category discussed in this section.

The two estimators are based on different constructions for estimating the missing-data process. The first is based on a nonparametric formulation for estimating $\mathrm{pr}(\delta = 1|Z) = \pi_{\mathrm{marg}}$, where the subscript indicates a marginal estimation of the probability that Y is observed. The second is based on a parametric formulation for estimating $\mathrm{pr}(\delta = 1|Y, X, Z) = \pi(X, Z, \zeta)$, where ζ is an unknown parameter estimated by standard logistic regression of δ on (X, Z).

The first estimator, similar to one defined by Wang et al. (2004) and efficient in the Gaussian partially linear model, can be constructed as follows. Estimate π_{marg} by local linear logistic regression of δ on Z, leading to the usual asymptotic expansion

$$\widehat{\pi}_{\mathrm{marg}}(z) - \pi_{\mathrm{marg}}(z) = n^{-1} \sum_{j=1}^{n} \{\delta_j - \pi_{\mathrm{marg}}(Z_j)\} K_h(z - Z_j)/f_Z(z) + o_p\big(n^{-1/2}\big), \tag{11}$$

assuming that $nh^4 \to 0$. Then construct the estimator

$$\widehat{\kappa}_{\mathrm{marg}} = n^{-1} \sum_{i=1}^{n} \left[\frac{\delta_i}{\widehat{\pi}_{\mathrm{marg}}(Z_i)} \mathcal{G}(Y_i) + \left\{1 - \frac{\delta_i}{\widehat{\pi}_{\mathrm{marg}}(Z_i)}\right\} \times \mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\} \right].$$

The estimator has two useful properties: (1) if there are no missing data, it does not depend on the model and is, hence, consistent for κ_0; and (2) if observation of the response Y depends only on Z, it is consistent even if the semiparametric model is not correct.

In a similar vein, the second estimate, also similar to another estimate of Wang et al. (2004), is given as

$$\widehat{\kappa} = n^{-1} \sum_{i=1}^{n} \left[\frac{\delta_i}{\pi(X_i, Z_i, \widehat{\zeta})} \mathcal{G}(Y_i) + \left\{1 - \frac{\delta_i}{\pi(X_i, Z_i, \widehat{\zeta})}\right\} \times \mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\} \right].$$

This estimator has the double-robustness property that if either the parametric model $\pi(X, Z, \zeta)$ or the underlying semiparametric model for $\{\mathcal{B}, \theta(\cdot)\}$ is correct, then $\widehat{\kappa}$ is consistent and asymptotically normally distributed. Generally, the second terms in both $\widehat{\kappa}_{\mathrm{marg}}$ and $\widehat{\kappa}$ improve efficiency: They are also important for the double-robustness property of $\widehat{\kappa}$.

If both models are correct, then the following results are obtained as a consequence of (5) and (6); see the Appendix for a sketch.

Lemma 1. Define

$$\mathcal{M}_{2,\mathrm{marg}} = \mathrm{E}\left[\left\{1 - \frac{\delta}{\pi_{\mathrm{marg}}(Z)}\right\} \{\mathcal{F}_{\mathcal{B}}(\cdot) + \mathcal{F}_\theta(\cdot)\theta_{\mathcal{B}}(Z, \mathcal{B}_0)\}^{\mathrm{T}}\right],$$

$$D_{i,\mathrm{marg}}(\cdot) = -\mathcal{L}_{i\theta}(\cdot) \mathrm{E}\left[\left\{1 - \frac{\delta_i}{\pi_{\mathrm{marg}}(Z_i)}\right\} \mathcal{F}_{i\theta}(\cdot) \middle| Z_i\right] \Big/ \mathrm{E}\{\delta \mathcal{L}_{\theta\theta}(\cdot)|Z_i\}.$$

Then, to terms of order $o_p(1)$,

$$n^{1/2}(\widehat{\kappa}_{\mathrm{marg}} - \kappa_0) \approx n^{-1/2} \sum_{i=1}^{n} \left[\frac{\delta_i}{\pi_{\mathrm{marg}}(Z_i)} \mathcal{G}(Y_i) + \left\{1 - \frac{\delta_i}{\pi_{\mathrm{marg}}(Z_i)}\right\} \mathcal{F}_i(\cdot) - \kappa_0 \right]$$

$$+ \mathcal{M}_{2,\rm marg}\mathcal{M}_1^{-1} n^{-1/2}\sum_{i=1}^n \delta_i\epsilon_i$$
$$+ n^{-1/2}\sum_{i=1}^n \delta_i D_{i,\rm marg}(\cdot). \tag{12}$$

Lemma 2. Define $\pi_\zeta(X,Z,\zeta) = \partial\pi(X,Z,\zeta)/\partial\zeta$. Assume that $n^{1/2}(\widehat\zeta - \zeta) = n^{-1/2}\sum_{i=1}^n \psi_{i\zeta}(\cdot) + o_p(1)$ with $\mathrm{E}\{\psi_\zeta(\cdot)| X,Z\} = 0$. Then, to terms of order $o_p(1)$,

$$n^{1/2}(\widehat\kappa - \kappa_0)$$
$$\approx n^{-1/2}\sum_{i=1}^n\Bigg[\frac{\delta_i}{\pi(X_i,Z_i,\zeta)}\{\mathcal{G}(Y_i) - \kappa_0\}$$
$$+ \bigg\{1 - \frac{\delta_i}{\pi(X_i,Z_i,\zeta)}\bigg\}\{\mathcal{F}_i(\cdot) - \kappa_0\}\Bigg]. \tag{13}$$

Remark 4. The expansions (12) and (13) show that $\widehat\kappa_{\rm marg}$ and $\widehat\kappa$ are asymptotically normally distributed. One can show that the asymptotic variances are given as

$$\mathcal{V}_{\kappa,\rm marg} = \mathrm{var}\Bigg[\frac{\delta}{\pi_{\rm marg}(Z)}\mathcal{G}(Y) + \bigg\{1 - \frac{\delta}{\pi_{\rm marg}(Z)}\bigg\}\mathcal{F}(\cdot)$$
$$+ \mathcal{M}_{2,\rm marg}\mathcal{M}_1^{-1}\delta\epsilon + \delta D_{\rm marg}(\cdot)\Bigg],$$
$$\mathcal{V}_\kappa = \mathrm{var}\Bigg[\frac{\delta_i}{\pi(X_i,Z_i,\zeta)}\mathcal{G}(Y_i)$$
$$+ \bigg\{1 - \frac{\delta_i}{\pi(X_i,Z_i,\zeta)}\bigg\}\mathcal{F}_i(\cdot)\Bigg],$$

respectively, from which estimates are readily derived.

Finally, we note that Claeskens and Carroll (2007) showed that in general likelihood problems, if there is an omitted covariate, then under contiguous alternatives the effect on estimators is to add an asymptotic bias, without changing the asymptotic variance.

3. SINGLE–INDEX MODELS

One means of dimension reduction is single-index modeling. Single-index models can be viewed as a generalized version of projection pursuit, in that only the most influential direction is retained to keep the model tractable and to reduce dimension. Since its introduction in Härdle and Stoker (1989), single-index modeling has been widely studied and used. A comprehensive summary of the model is given in Härdle, Müller, Sperlich, and Werwatz (2004). Let $Z = (R, S^{\rm T})^{\rm T}$ where R is a scalar. We consider here the generalized partially linear single-index model (GPLSIM) of Carroll et al. (1997), namely, the exponential family (20) with $\eta(X,Z) = X^{\rm T}\beta_0 + \theta_0(Z^{\rm T}\alpha_0)$, where $\theta_0(\cdot)$ is an unknown function and for identifiability purposes $\|\alpha_0\| = 1$. Because identifiability requires that one of the components of Z be a nontrivial predictor of Y, for convenience we will make the very small modification that one component of Z, what we call R, is a known nontrivial predictor of Y. The reason for making this modification can be seen in theorem 4 of Carroll et al. (1997) where the final limit distribution of the estimate of α_0 has a singular covariance matrix. In addition, their main asymptotic expansion, given in their equation (A.12), is about the nonsingular transformation $(I - \alpha_0\alpha_0^{\rm T})(\widehat\alpha - \alpha_0)$.

With this modification, we write the model as

$$\mathrm{E}(Y|X,Z) = \mathcal{C}^{(1)}\big[c\{\eta(X,Z)\}\big] = \mu\{X^{\rm T}\beta_0 + \theta_0(R + S^{\rm T}\gamma_0)\}, \tag{14}$$

where γ_0 is unrestricted.

Carroll et al. (1997) used profile likelihood to estimate $\mathcal{B}_0 = (\gamma_0, \beta_0)$ and $\theta_0(\cdot)$, although they presented no results concerning the estimate of ϕ_0, their interest largely being in logistic regression where $\phi_0 = 1$ is known. Rewrite the likelihood function (20) as $L\{Y, X, \beta, \theta(R + S^{\rm T}\gamma), \phi\}$. Then, given $\mathcal{B} = (\gamma^{\rm T}, \beta^{\rm T})^{\rm T}$, they formed $U(\gamma) = R + S^{\rm T}\gamma$ and computed the estimate $\widehat\theta\{u(\gamma), \mathcal{B}\}$ by local likelihood of Y on $\{X, U(\gamma)\}$ as in Severini and Staniswalis (1994), using the data with $\delta = 1$. Then they maximized $\sum_{i=1}^n \delta_i \log[L\{Y_i, X_i, \beta, \widehat\theta(R_i + S_i^{\rm T}\gamma, \mathcal{B}), \phi\}]$ in $\mathcal{B}$ and ϕ.

Our goal is to estimate $\kappa_{\rm SI} = \mathrm{E}[\mathcal{F}\{X, \theta_0(R + S^{\rm T}\gamma_0), \beta_0, \phi_0\}]$. Our proposed estimate is $\widehat\kappa_{\rm SI} = n^{-1}\sum_{i=1}^n \mathcal{F}\{X_i, \widehat\theta(R_i + S_i^{\rm T}\widehat\gamma, \widehat{\mathcal{B}}), \widehat\beta, \widehat\phi\}$.

Our main result is as follows. First, define $U = R + S^{\rm T}\gamma_0$ and

$$\mathcal{G} = \mathcal{D}_\phi(Y, \phi_0) - \big[Yc\{X^{\rm T}\beta_0 + \theta_0(U)\} - \mathcal{C}\{c(\cdot)\}\big]/\phi_0^2.$$

Also define $\Lambda = \{S^{\rm T}\theta_0^{(1)}(U), X^{\rm T}\}^{\rm T}$, $\rho_\ell(\cdot) = \{\mu^{(1)}(\cdot)\}^\ell/V(\cdot)$, and $\epsilon = [Y - \mu\{X^{\rm T}\beta_0 + \theta_0(U)\}]\rho_1\{X^{\rm T}\beta_0 + \theta_0(U)\}$. Define $\mathcal{N}_i = \Lambda_i - [\mathrm{E}\{\delta\rho_2(\cdot)|U_i\}]^{-1}\mathrm{E}\{\delta_i\Lambda_i\rho_2(\cdot)|U_i\}$ and $\mathcal{Q} = \mathrm{E}\{\delta\mathcal{N}\mathcal{N}^{\rm T} \times \rho_2(\cdot)\}$. Make the further definitions $\mathcal{F}_\beta(\cdot) = \partial\mathcal{F}\{X, \theta_0(U), \beta_0, \phi_0\}/\partial\beta_0$, $\mathcal{F}_\phi(\cdot) = \partial\mathcal{F}\{X, \theta_0(U), \beta_0, \phi_0\}/\partial\phi_0$, and $\mathcal{F}_\theta(\cdot) = \partial\mathcal{F}\{X, \theta_0(U), \beta_0, \phi_0\}/\partial\theta_0(U)$. Also define

$$J(U) = [\mathrm{E}\{\delta\rho_2(\cdot)|U\}]^{-1}\mathrm{E}\{\mathcal{F}_\theta(\cdot)|U\},$$
$$D = \begin{bmatrix} \mathrm{E}\{\mathcal{F}_\theta(\cdot)\theta^{(1)}(U)S\} - \mathrm{E}(\mathcal{F}_\theta(\cdot)[\mathrm{E}\{\delta\rho_2(\cdot)|U\}]^{-1}\theta^{(1)}(U)\mathrm{E}\{\delta S\rho_2(\cdot)|U\}) \\ \mathrm{E}\{\mathcal{F}_\beta(\cdot)\} - \mathrm{E}(\mathcal{F}_\theta(\cdot)[\mathrm{E}\{\delta\rho_2(\cdot)|U\}]^{-1}\mathrm{E}\{\delta X\rho_2(\cdot)|U\}) \end{bmatrix}.$$

Then we have the following result regarding the asymptotic distribution of $\widehat\kappa_{\rm SI}$.

Theorem 2. Assume that $(Y_i, \delta_i, X_i, Z_i), i = 1, 2, \ldots, n$, are iid and that the conditions in Carroll et al. (1997) hold, in particular, that $nh^4 \to 0$. Then

$$n^{1/2}(\widehat\kappa_{\rm SI} - \kappa_{\rm SI})$$
$$= n^{-1/2}\sum_{i=1}^n\big[\mathcal{F}\{X_i, \theta_0(U_i), \beta_0, \phi_0\} - \kappa_{\rm SI}$$
$$+ D^{\rm T}\mathcal{Q}^{-1}\delta_i\mathcal{N}_i\epsilon_i + \delta_i J(U_i)\epsilon_i$$
$$+ \delta_i\mathcal{G}_i\mathrm{E}\{\mathcal{F}_\phi(\cdot)\}/\mathrm{E}(\delta\mathcal{G}^2)\big] + o_p(1)$$
$$\Rightarrow \mathrm{Normal}(0, \mathcal{V}), \tag{15}$$

where $\mathcal{V} = \mathrm{E}[\mathcal{F}\{X, \theta_0(U), \beta_0, \phi_0\} - \kappa_{\rm SI}]^2 + D^{\rm T}\mathcal{Q}^{-1}D + \mathrm{var}\{\delta \times J(U)\epsilon\} + \mathrm{E}(\delta\mathcal{G}^2)[\mathrm{E}\{\mathcal{F}_\phi(\cdot)\}]^2/\{\mathrm{E}(\delta\mathcal{G}^2)\}^2$. Further, $\widehat\kappa_{\rm SI}$ is semiparametric efficient.

4. MOTIVATING EXAMPLE

4.1 Introduction

There is considerable interest in understanding the distribution of dietary intake in various populations. For example, as obesity rates continue to rise in the United States (Flegal, Carroll, Ogden, and Johnson 2002), the demand for information about diet and nutrition is increasing. Information on dietary intake has implications for establishing population norms, conducting research, and making public policy decisions (Woteki 2003).

We wish to emphasize that there are no missing response data in this example. We also emphasize that the problem is vastly different from simply estimating the population mean using a Gaussian partially linear model. The strength of our approach is that once we have proposed a semiparametric model, then our methodology, asymptotics, and semiparametric efficiency results are readily employed.

This article was motivated by the analysis of the Eating at America's Table Study (EATS) (Subar et al. 2001), where estimating the distribution of the consumption of episodically consumed foods is of interest. The data consist of four 24-hour recalls over the course of a year as well as the National Cancer Institute's (NCI) dietary history questionnaire (DHQ), a particular version of a food frequency questionnaire (FFQ; see Willett et al. 1985 and Block et al. 1986). The goal is to estimate the distribution of usual intake, defined as the average daily intake of a dietary component by an individual in a fixed time period, a year in the case of EATS. There were $n = 886$ individuals in the dataset.

When the responses are continuous random variables, this is a classical problem of measurement error, with a large literature. However, little of the literature is relevant to episodically consumed foods, as we now describe. Consider, for example, consumption of red meat, dark-green vegetables, and deep-yellow vegetables, all of interest in nutritional surveillance. In the EATS data, 45% of the 24-hour recalls reported no red-meat consumption. In addition, 5.5% of the individuals reported no red-meat consumption on any of the four separate 24-hour recalls: For deep-yellow vegetables these numbers are 63% and 20%, respectively, while for dark-green vegetables the numbers are 78% and 46%, respectively. Clearly, methods aimed at understanding usual intakes for continuous data are inappropriate for episodically consumed foods with so many zero-reported intakes.

4.2 Model

To handle episodically consumed foods, two-part models have been developed (Tooze, Grunwald, and Jones 2002). These are basically zero-inflated repeated-measures examples. Our methods are applicable to such problems when the covariate Z is evaluated only once for each subject, as it is in our example.

We describe here a simplification of this approach, used to illustrate our methodology. On each individual, we measure age and gender, the collection being what we call R. We also observe energy (calories) as measured by the DHQ, the logarithm of which we call Z. The reader should note that Z is evaluated only once per individual, and, hence, while there are repeated measures on the responses, there are no repeated measures on Z: $\theta_0(Z)$ occurs only once in the likelihood function, and our methodology applies.

Let $X = (R, Z)$. The response data for an individual i consist of four 24-hour recalls of red-meat consumption. Let $\Delta_{ij} = 1$ if red meat is reported consumed on the jth 24-hour recall for $j = 1, \ldots, 4$. Let $\mathcal{Y}_{ij}$ be the product of Δ_{ij} and the logarithm of reported red-meat consumption, with the convention that $0\log(0) = 0$. Then the response data are $Y_i = (\Delta_{ij}, \mathcal{Y}_{ij})_{j=1}^4$.

4.2.1 Modeling the Probability of Zero Response. The first part of the model is whether the subject reports red-meat consumption. We model this as a repeated-measures logistic regression, so that

$$\mathrm{pr}(\Delta_{ij} = 1 | R_i, Z_i, U_{i1}) = H(\beta_0 + X_i^{\mathrm{T}}\beta_1 + U_{i1}), \tag{16}$$

where $H(\cdot)$ is the logistic distribution function and $U_{i1} = \mathrm{Normal}(0, \sigma_{u1}^2)$ is a person-specific random effect. Note that, for simplicity, we have modeled the effect of energy consumption as linear, because in the data there is little hint of nonlinearity.

4.2.2 Modeling Positive Responses. The second part of the model consists of a distribution of the logarithm of red-meat consumption on days when consumption is reported, namely,

$$[\mathcal{Y}_{ij} | \Delta_{ij} = 1, R_i, Z_i, U_{i2}] = \mathrm{Normal}\{R_i^{\mathrm{T}}\beta_2 + \theta(Z_i) + U_{i2}, \sigma^2\}, \tag{17}$$

where $U_{i2} = \mathrm{Normal}(0, \sigma_{u2}^2)$ is a person-specific random effect, which we take to be independent of U_{i1}. Note that (17) means that the nonzero $\mathcal{Y}$ data within an individual marginally have the same mean $R_i^{\mathrm{T}}\beta_2 + \theta(Z_i)$, variance $\sigma^2 + \sigma_{u2}^2$, and common covariance σ_{u2}^2.

4.2.3 Likelihood Function. The collection of parameters is $\mathcal{B}$, consisting of β_0, β_1, β_2, σ_{u1}^2, σ_{u2}^2, and σ^2. The log-likelihood function $\mathcal{L}(\cdot)$ is readily computed with numerical integration as follows:

$$\begin{aligned}\exp\{\mathcal{L}(\cdot)\} = \frac{1}{\sigma_{u1}} \int \phi\left(\frac{u_1}{\sigma_{u1}}\right) \prod_{j=1}^{4} &\{H(\beta_0 + X^{\mathrm{T}}\beta_1 + u_1)\}^{\Delta_{ij}} \\ &\times \{1 - H(\beta_0 + X^{\mathrm{T}}\beta_1 + u_1)\}^{1-\Delta_{ij}}\, du_1 \\ &\times \sigma_{u2}^{-1}\sigma^{-\sum_j \Delta_{ij}} \int \phi\left(\frac{u_2}{\sigma_{u2}}\right) \\ &\times \prod_{j=1}^{4}\left(\phi\left[\frac{\mathcal{Y}_{ij} - \{R_i^{\mathrm{T}}\beta_2 + \theta(Z_i) + u_2\}}{\sigma}\right]\right)^{\Delta_{ij}} du_2.\end{aligned}$$

Of course, the second numerical integral is not necessary, because the integration can be done analytically.

4.2.4 Defining Usual Intake at the Individual Level. Noting from (17) that reported intake on days of consumption follows a log-normal distribution, the usual intake for an individual is defined as

$$\begin{aligned}&G\{X, U_1, U_2, \mathcal{B}, \theta(Z)\} \\ &\quad = H(\beta_0 + X_i^{\mathrm{T}}\beta_1 + U_1) \\ &\qquad \times \exp\{R^{\mathrm{T}}\beta_2 + \theta(Z) + U_2 + \sigma^2/2\}.\end{aligned} \tag{18}$$

The goal is to understand the distribution of $G\{X, U_1, U_2, \mathcal{B}, \theta(Z)\}$ across a population. In particular, for arbitrary c we wish to estimate $\mathrm{pr}[G\{X, U_1, U_2, \mathcal{B}, \theta(Z)\} > c]$. Define $\mathcal{F}\{X, \mathcal{B}, \theta(Z)\} = \mathrm{pr}[G\{X, U_1, U_2, \mathcal{B}, \theta(Z)\} > c | X, Z]$, a quantity that can be computed by numerical integration. Then $\kappa_0 = \mathrm{E}[\mathcal{F}\{X, \mathcal{B}, \theta(Z)\}]$ is the percentage of the population whose long-term reported daily average consumption of red meat exceeds c.

4.3 Bias in Naive Estimates, and a Simulation Study

We emphasize that the distribution of mean intake cannot be estimated consistently by the simple device of computing the sample percentage of the observed 24-hour recalls that exceed c, and, as a consequence, going through the model-fitting process is actually necessary. To see this, suppose only one 24-hour recall per person was computed and the percentage of these 24-hour recalls exceeding c was computed. In large samples, this percentage converges to

$$\kappa_{24\text{hr}} = \mathrm{E}\big(H(\beta_0 + X^{\mathrm{T}}\beta_1 + U_1) \times \Phi\big[\{R^{\mathrm{T}}\beta_2 + \theta(Z) - \log(c)\}/(\sigma^2 + \sigma_2^2)^{1/2}\big]\big).$$

In contrast, for $\sigma_2 > 0$,

$$\kappa_0 = \mathrm{E}\big\{\Phi\big(\big[R^{\mathrm{T}}\beta_2 + \theta(Z) + \sigma^2/2 - \log\{c/H(\beta_0 + X^{\mathrm{T}}\beta_1 + U_1)\}\big]/\sigma_2\big)\big\}.$$

As the number of replicates m of the 24-hour recall approaches ∞, the percentage $\kappa_{m,24\text{hr}}$ of the means of the 24-hour recalls that exceed $c \to \kappa_0$, so we would expect that the fewer the replicates, the less our estimate agrees with the sample version of $\kappa_{m,24\text{hr}}$, a phenomenon observed in our data; see below.

To see this numerically, we ran the following simulation study. Gender, age, and the DHQ were kept the same as in the EATS. The parameters $(\beta_0, \beta_1, \beta_2, \sigma^2, \sigma_1^2, \sigma_2^2)$ were the same as our estimated values; see below. The function $\theta(\cdot)$ was roughly in accord with our estimated function, for simplicity, being quadratic in the logarithm of the DHQ, standardized to have minimum .0 and maximum 1.0, with intercept, slope, and quadratic parameters being .50, 1.50, and −.75, respectively. The true survival function, that is, 1 − the cdf, was computed analytically, while the survival functions for the mean of two 24-hour recalls and the mean of four 24-hour recalls were computed by 1,000 simulated datasets.

The results are given in Figure 1, where the bias from not using a model is evident.

We used our methods with a nonparametrically estimated function, a bandwidth $h = .30$, and the Epanechnikov kernel function. We generated 300 datasets, with results displayed in Figure 2. The mean over the simulation was almost exactly the correct function, not surprising given that the sample size is large ($n = 886$). In Figure 2 we also display a 90% confidence range from the simulated datasets, indicating that in the EATS data at least, the results of our approach are relatively accurate.

4.4 Data Analysis

We standardized age to have mean 0 and variance 1. In the logistic part of the model, the intercept was estimated as −8.15, with the coefficients for (gender, age, DHQ) = (.13, .14, 1.09). The random-effect variance was estimated as $\widehat{\sigma}_1^2 = .66$. In the continuous part of the model, we used bandwidths ranging from .05 to .40, with little change in any of the estimates, as described in more detail in Section 5.1. With a bandwidth $h = .30$, our estimates were $\widehat{\sigma}^2 = .76$, $\widehat{\sigma}_2^2 = .043$, and the coefficients for gender and age were −.25 and .02, respectively. The coefficient for the person-specific random effect σ_2^2 appears intrinsic to the data: We used other methods such as mixed models with polynomial fits and obtained roughly the same answers.

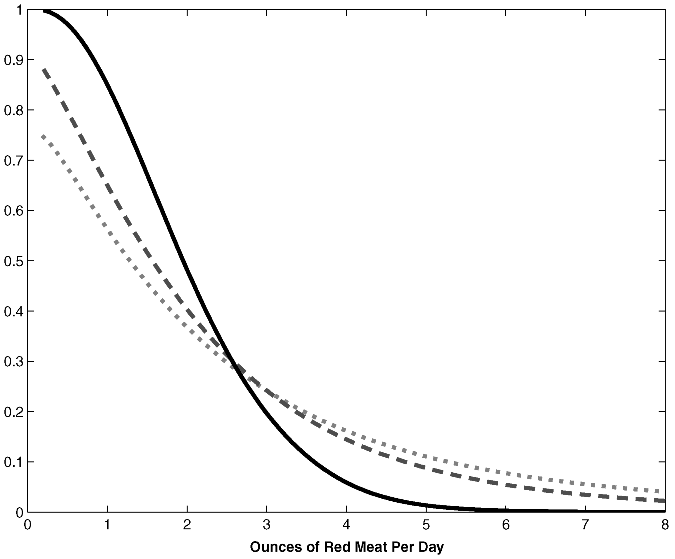

Figure 1. Results of the Simulation Study Meant to Mimic the EATS Study. All results are averages over 1,000 simulated datasets. The mean of the semiparametric estimator (——) of the survival curse, which is almost identical to the true survival curve. The empirical survival function of the mean of two 24-hour recalls (•••••) from 1,000 simulated datasets. The empirical survival function of the mean of four 24-hour recalls (▬ ▬ ▬) from 1,000 simulated datasets.

We display the computed survival function in Figure 3. Displayed there are our method, along with the empirical survival functions for the mean of the first two 24-hour recalls and the mean of all four 24-hour recalls. While these are biased, it is interesting to note that using the mean of only two 24-hour recalls is more different from our method than using the mean of four 24-hour recalls, which is expected as described previously. The similarity of Figures 1 and 3 is striking, mainly indicating that naive approaches, such as using the mean of two 24-hour recalls, can result in badly biased estimates of κ_0.

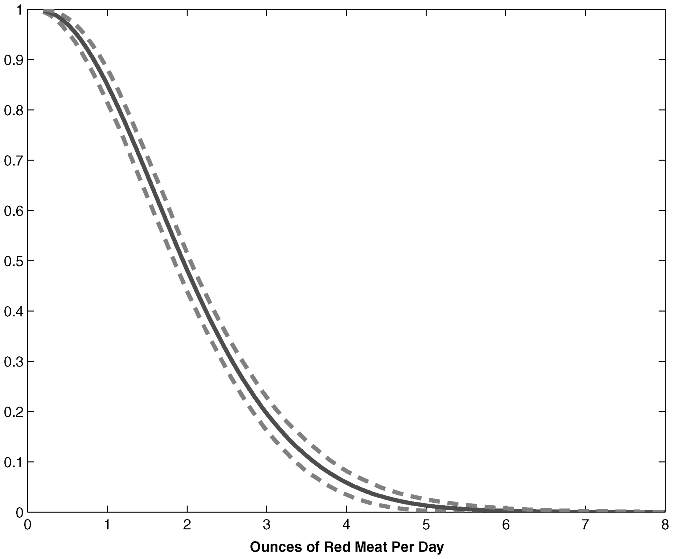

Figure 2. Results of the Simulation Study Meant to Mimic the EATS Study. Plotted is the mean survival function for 300 simulated datasets, along with the 90% pointwise confidence intervals. The mean fitted function is almost exact.

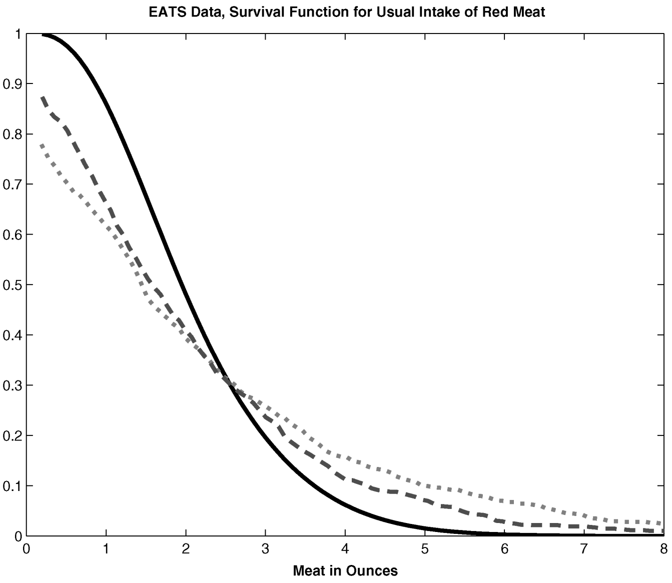

Figure 3. Results From the EATS Example. Plotted are estimates of the survival function (1 − the cdf) of usual intake of red meat. The solid line is the semiparametric method described in Section 4. The dotted line is the empirical survival function of the mean of the first two 24-hour recalls per person, while the dashed line is survival function of the mean of all the 24-hour recalls per person.

5. BANDWIDTH SELECTION, THE PARTIALLY LINEAR MODEL, AND THE SAMPLE MEAN

5.1 Bandwidth Selection

5.1.1 Background. We have used a standard first-order kernel density function, that is, one with mean 0 and positive variance. With this choice, in Theorem 1 we have assumed that the bandwidth satisfies $nh^4 \to 0$: for estimation of the population mean in the partially linear model. In contrast, if one were interested only in $\mathcal{B}_0$, then it is well known that by using profile likelihood the usual bandwidth order $h \sim n^{-1/5}$ is acceptable, and off-the-shelf bandwidth selection techniques yield an asymptotically normal limit distribution.

The reason for the technical need for undersmoothing is the inclusion of $\theta_0(\cdot)$ in κ_0. For example, suppose that $\kappa_0 = \mathrm{E}\{\theta_0(Z)\}$. Then it follows from (5) that $\widehat{\kappa} - \kappa_0 = O_p(h^2 + n^{-1/2})$. Thus, in order for $n^{1/2}(\widehat{\kappa} - \kappa_0) = O_p(1)$, we require that $nh^4 = O_p(1)$. The additional restriction that $nh^4 \to 0$ merely removes the bias term entirely.

Note that κ_0 is not a parameter in the model, being a mixture of the parametric part $\mathcal{B}_0$, the nonparametric part $\theta_0(\cdot)$, and the joint distribution of (X, Z). Thus, it does not appear that κ_0 can be estimated by profiling ideas.

5.1.2 Optimal Estimation. As seen in Theorem 1, the asymptotic distribution of $n^{1/2}(\widehat{\kappa} - \kappa_0)$ is unaffected by the bandwidth, at least to first order. In Section 5.1.3 we give intuitive and numerical evidence of the lack of sensitivity to the bandwidth choice; see also Section 5.3 for further numerical evidence. In Section 5.1.4 we describe three different, simple practical methods for bandwidth selection in this problem, all of which work quite well in our simulations and example.

Because first-order calculations do not get squarely at the choice of bandwidth, other than to suggest that it is not particularly crucial, an alternative theoretical device is to do second-order calculations. Define $\eta(n, h) = n^{1/2}h^2 + (n^{1/2}h)^{-1}$. In a problem similar to ours, Sepanski, Knickerbocker, and Carroll (1994) showed that the variance of linear combinations of the estimate of $\mathcal{B}_0$ has a second-order expansion as follows. Suppose we want to estimate $\xi^{\mathrm{T}}\mathcal{B}_0$. Then, for constants (a_1, a_2),

$$n^{1/2}(\xi^{\mathrm{T}}\widehat{\mathcal{B}} - \xi^{\mathrm{T}}\mathcal{B}_0) = V_n + o_p\{\eta(n, h)\},$$
$$\mathrm{cov}(V_n) = \mathrm{constant} + \{a_1 n^{1/2}h^2 + a_2(hn^{1/2})^{-1}\}^2.$$

This means that the optimal bandwidth is on the order of $h = cn^{-1/3}$ for, a constant c depending on (a_1, a_2), which, in turn, depend on the problem, that is, on the distribution of (Y, X, Z) as well as on $\mathcal{B}_0$ and $\theta_0(\cdot)$. In their practical implementation, translated from the Gaussian kernel function to our Epanechnikov kernel function, Sepanski et al. (1994) suggested the following device, namely, that if the optimal bandwidth for estimating $\theta_0(\cdot)$ is $h_o = cn^{-1/5}$, then one should use the correct-order bandwidth $h = cn^{-1/3}$. They also did sensitivity analysis, for example, $h = (1/2)cn^{-1/3}$, but found little change in their simulations. One of our three methods of practical bandwidth selection is exactly this one.

A problem not within our framework but carrying a similar flavor was considered by Powell and Stoker (1996) and Newey, Hsieh, and Robins (2004), namely, the estimation of the weighted average derivative $\kappa_{\mathrm{AD}} = \mathrm{E}\{Y\theta_0^{(1)}(Z)\}$. As done by Sepanski et al. (1994), Powell and Stoker (1996) showed that the optimal bandwidth constructed from second-order calculations is an undersmoothed bandwidth. Newey et al. (2004) suggested that a simple device of choosing the bandwidth is to choose something optimal when using a standard second-order kernel function but to then undersmooth, in effect, by using a higher order kernel such as the twicing kernel. This is our second bandwidth selection method described in Section 5.1.4. Like the first, it appears to be an effective means of eliminating the bias term.

In our problem, the article by Sepanski et al. (1994) is more relevant. Preliminary calculations based on the basic tools in that article suggest that for our problem, the optimal bandwidth is also of order $n^{-1/3}$. We intend to pursue these very calculations in another article.

5.1.3 Lack of Sensitivity to Bandwidth. We have used the term *technical need for undersmoothing* because that is what it really is. In practice, as Theorem 1 states, the asymptotic distribution of $\widehat{\kappa}$ is unaffected by bandwidth choice for very broad ranges of bandwidths. This is totally different from what happens with estimation of the function $\theta_0(\cdot)$, where bandwidth selection is typically critical in practice, and this is seen in theory through the usual bias–variance tradeoff.

In practice, we expect little effect of the bandwidth selection on estimation of $\mathcal{B}_0$, and even less effect on estimation of κ_0. The reason is that broad ranges of bandwidths lead to no asymptotic effect on the distribution of $\widehat{\mathcal{B}}$. The extra amount of smoothing inherent in the summation in (4) should mean that $\widehat{\kappa}$ will be even less sensitive to the bandwidth, the so-called *double-smoothing* phenomenon.

To see this issue, consider the simulation in Wang et al. (2004). They set X and Z to be independent, with $X = \mathrm{Normal}(1, 1)$ and $Z = \mathrm{Uniform}[0, 1]$. In the partially linear model, they set $\mathcal{B}_0 = 1.5$, $\epsilon = \text{Normal}(0,1)$, and $\theta_0(z) = 3.2z^2 - 1$. They used the kernel function $(15/16)(1-z^2)^2 \times I(|z| \le 1)$, and they fixed the bandwidth to be $h = n^{-2/3}$, which at least asymptotically is very great undersmoothing, because $h \sim n^{-1/3}$ is already acceptable and typically something like $nh^2/\log(n) \to \infty$ is usually required. In their case 3, they used effective sample sizes for complete data of 18, 36, and 60, with corresponding bandwidths .146, .092, and .065, respectively.

We reran the simulation of Wang et al. (2004), with complete response data and $n = 60$. We used bandwidths .02, .06, .10, and .14, ranging from a very small bandwidth, less than 1/3 that used by Wang et al. (2004), to a larger bandwidth, more than double that used. As another perspective, if one sets $h = \sigma_z n^{-c}$, where σ_z is the standard deviation of Z, then the bandwidths used are equivalent to $c = .73$, .46, .34, and .26. In other words, a bandwidth here of $h = .02$ is very great undersmoothing, while even $h = .14$ satisfies the theoretical constraint on the bandwidth.

In Figure 4 we plot the results for a *single* dataset, where, as in Wang et al. (2004), interest lies in estimating $\kappa_0 = \mathrm{E}(Y)$. As is obvious from this figure, the bandwidth choice is very important for estimation of the function, but trivially unimportant for estimation of κ_0, the estimate of which ranged from 1.818 to 1.828.

In Figure 5 we plot the mean estimated functions from 100 simulated datasets. Again, the bandwidth matters a great deal for estimating the function $\theta_0(\cdot)$. Again, too, the bandwidth matters hardly at all for estimating κ_0. Thus, for estimating κ_0, the mean estimates across the bandwidths range from 1.513 to 1.526, and the standard deviations of the estimates range from .249 to .252. There is somewhat more effect of bandwidth on the estimate of $\mathcal{B}_0$: For $h \ge .06$, there is almost no effect, but choosing $h = .02$ results in a 50% increase in standard deviation.

In other words, as expected by theory and intuition, bandwidth selection has little effect on the estimate of $\mathcal{B}_0$, except when the bandwidth is much too small, and very little effect on the estimation of $\kappa_0 = \mathrm{E}(Y)$. Similar results occur when one looks at the variance of the errors as the parameter, and κ_0 is the population variance.

5.1.4 Bandwidth Selection. As described in Section 5.1.3, bandwidth selection is not a vital issue for estimating κ_0: Of course, it is vital for estimating $\theta_0(\cdot)$. Effectively, what this means is that the real need is simply to get bandwidths that satisfy the technical assumption of undersmoothing but are not too ridiculously small: A precise target is often unnecessary. In addition, because the asymptotic distribution of $\widehat{\kappa}$ does not depend on the bandwidth, simple first-order methods of the type that are used in bandwidth selection for function estimation are not possible. Thus, in our example, we used three different methods, all of which gave answers that were as nearly identical as in the simulation of Wang et al. (2004).

All the methods are based on a so-called "typical device" to get an optimal bandwidth for estimating θ_0, of the form $h_{\text{opt}} = c\sigma_z n^{-1/5}$. In practice, this can be accomplished by constructing a finite grid of bandwidths of the form $h_{\text{grid}} = c_{\text{grid}}\sigma_z n^{-1/5}$: We use a grid from .20 to 5.0. After estimating $\mathcal{B}_0$ by $\widehat{\mathcal{B}}(h_{\text{grid}})$, this value is fixed, and then a log-likelihood cross-validation score is obtained. The maximizer of the log-likelihood cross-validation score is selected as h_{opt}.

- If $h_{\text{opt}} = c\sigma_z n^{-1/5}$, an extremely simple device is simply to set $h = h_{\text{opt}} n^{-2/15} = c\sigma_z n^{-1/3}$, which satisfies the technical condition of undersmoothing without becoming ridiculously small. This device may seem terribly ad hoc, but the theory, the simulation of Wang et al. (2004), the discussion in Section 5.1.3, and our own work suggest that this method actually works reasonably well. Note, too, that in Section 5.1.2 we give evidence that this bandwidth rate is most likely optimal.
- A second approach is taken by Newey et al. (2004) and is also an effective practical device. The technical need for undersmoothing comes from the fact that the bias term in a first-order local likelihood kernel regression is of order $O(h^2)$. One can use higher order kernels to get the bias to be of order $O(h^{2s})$ for $s \ge 2$, but this does not really help in that the variance remains of order $O\{(nh)^{-1}\}$, so that the optimal mean squared error kernel estimator has $h = O\{n^{-1/(4s+1)}\}$, and thus undersmoothing to estimate κ_0 is still required. However, as Newey et al. (2004) pointed out, if one uses the optimal bandwidth $h_{\text{opt}} = c\sigma_z n^{-1/5}$, but then does the estimation procedure replacing the first-order kernel by a higher order kernel, then the bias is $O(h_{\text{opt}}^{2s}) = o(n^{-1/2})$ if $s \ge 2$. A convenient higher order kernel is the second-order twicing kernel $K_{\text{tw}}(u) = 2K(u) - \int K(u-v)K(v)\,dv$, where $K(\cdot)$ is a first-order kernel.
- One can also use log-likelihood cross-validation, but with the grid of values being of the form $h_{\text{grid}} = c_{\text{grid}}\sigma_z n^{-1/3}$. Because cross-validation scores often have multiple modes, this is not the same as optimal smoothing.

It may be worth pointing out again that Wang et al. (2004) set $h = n^{-2/3}$, and even then, with too much undersmoothing (asymptotically), the performance of the method is rather good.

5.2 Efficiency and Robustness of the Sample Mean

In general problems with complete data, with no assumptions about the response Y other than that it has a second moment, the sample mean $\overline{Y}$ is semiparametric efficient for estimating the population mean $\kappa_0 = \mathrm{E}(Y)$; see, for example, Newey (1990). Somewhat remarkably, Wang et al. (2004) showed that in the partially linear model with Gaussian errors, with complete data the sample mean is still semiparametric efficient. This fact is crucial, of course, in establishing that with missing response data, their estimators are still semiparametric efficient.

It is clear that with complete data, the sample mean will not be semiparametric efficient for all semiparametric likelihood models. Simple counterexamples abound, for example, the partially linear model for Laplace or t errors. More complex examples can be constructed, for example, the partially linear model in the Gamma family with log-linear mean $\exp\{X^{\mathrm{T}}\mathcal{B}_0 + \theta_0(Z)\}$: Details follow from Lemma 4.

The model robustness of the sample mean for estimating the population mean in complete data is nonetheless a powerful feature. It is, therefore, of considerable interest to know whether there are cases of semiparametric likelihood problems where the sample mean is still semiparametric efficient and, thus, would be used because of its model robustness. It turns out that such cases exist. In particular, the sample mean for complete response data is semiparametric efficient in canonical exponential families with partially linear form.

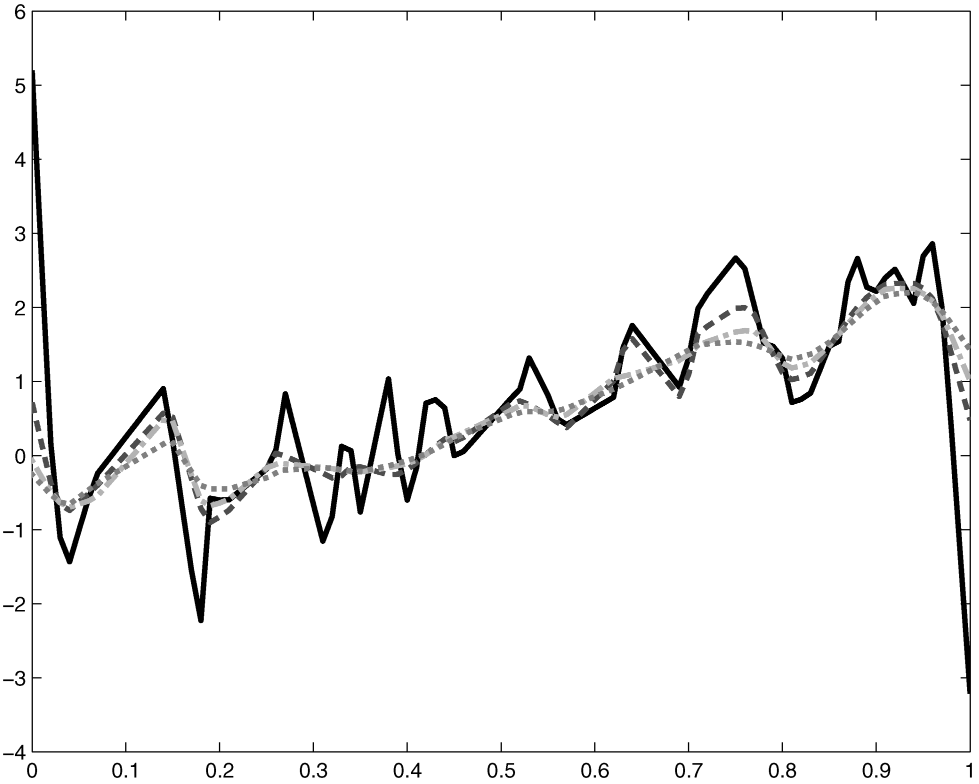

Figure 4. Results for a Single Dataset in a Simulation as in Wang et al. (2004), the Partially Linear Model With n = 60, Complete Response Data, and When $\kappa_0 = E(Y)$. Various bandwidths are used, and the estimates of the function $\theta_0(\cdot)$ are displayed. Note how the bandwidth has a major impact on the function estimate when the bandwidth is too small (h = .02), but very little effect on the estimate of κ_0. (——, h = .02, κ = 1.828; – – –, h = .06, κ = 1.818; –·–, h = .10, κ = 1.826; ·····, h = .14, κ = 1.818.)

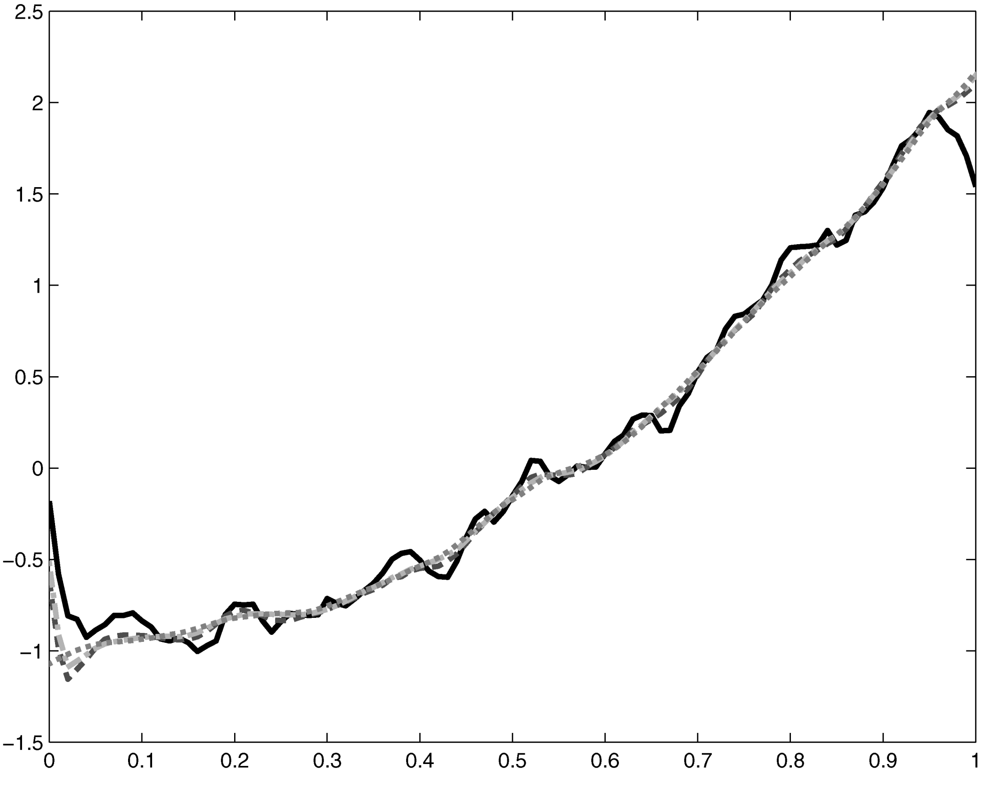

Figure 5. Results for 100 Simulated Datasets in a Simulation as in Wang et al. (2004), the Partially Linear Model With n = 60 and Complete Response Data. Various bandwidths are used, and the mean estimates of the function $\theta_0(\cdot)$ are displayed. Note how even over these simulations, the bandwidth has a clear impact on the function estimate: There is almost no impact on estimates of the population mean and variance. (——, h = .02; – – –, h = .06; –·–, h = .10; ·····, h = .14.)

Lemma 3. Recall that ϵ is defined in (8). If there are no missing data, the sample mean is a semiparametric efficient estimator of the population mean only if

$$Y - \mathrm{E}(Y|X,Z) = \mathrm{E}(Y\epsilon^{\mathrm{T}})\mathcal{M}_1^{-1}\epsilon + \mathcal{L}_\theta(\cdot)\frac{\mathrm{E}\{Y\mathcal{L}_\theta(\cdot)|Z\}}{\mathrm{E}\{\mathcal{L}_\theta^2(\cdot)|Z\}}. \quad (19)$$

It is interesting to consider (19) in the special case of exponential families with likelihood function

$$f(y|x,z) = \exp\left[\frac{yc\{\eta(x,z)\} - \mathcal{C}[c\{\eta(x,z)\}]}{\phi} + \mathcal{D}(y,\phi)\right], \quad (20)$$

where $\eta(x,z) = x^{\mathrm{T}}\beta_0 + \theta_0(z)$, so that $\mathrm{E}(Y|X,Z) = \mathcal{C}^{(1)}[c\{\eta(X,Z)\}] = \mu\{\eta(X,Z)\} = \mu(X,Z)$ and $\mathrm{var}(Y|X,Z) = \phi\mathcal{C}^{(2)}[c\{\eta(X,Z)\}] = \phi V[\mu\{\eta(X,Z)\}]$.

As it turns out, effectively, (19) holds and the sample mean is semiparametric efficient only in the canonical exponential family for which $c(t) = t$. More precisely, we show in Section A.6 the following result.

Lemma 4. If there are no missing data, under the exponential model (20), the sample mean is a semiparametric efficient estimate of the population mean if $\partial c\{X^{\mathrm{T}}\beta + \theta(Z)\}/\partial\theta(Z)$ is a function only of Z for all β, for example, the canonical exponential family. Otherwise, the sample mean is generally not semiparametric efficient: The precise condition is given in (A.29) in the Appendix. In particular, outside the canonical exponential family, the only possibility for the sample mean to be semiparametric efficient is that if for some known (a,b), $c\{x^{\mathrm{T}}\beta + \theta(z)\} = a + b\log\{x^{\mathrm{T}}\beta + \theta(z)\}$.

Remark 5. We consider Lemmas 3 and 4 to be *positive* results, although an earlier version of the paper had a misplaced emphasis. Effectively, we have characterized the cases, with complete data, that the sample mean is both model free and semiparametric efficient. In these cases, one would use the sample mean, or perhaps a robust version of it, rather than fit a potentially complex semiparametric model that can do no better and if that model is incorrect, can incur nontrivial bias.

5.3 Numerical Experience and Theoretical Insights in the Partially Linear Model, and Some Tentative Conclusions

In responding to a referee about the estimation of the population mean in the partially linear model (1), we collect here a few remarks based on our numerical experience. Because the problem of estimating the population mean is the problem focused on by Chen et al. (2003), we focus on the simulation setup in their article, although some of the conclusions we reach may be supportable in general cases. To remind the reader, in their simulation, X and Z are independent, with $X = \mathrm{Normal}(0,1)$, $Z = \mathrm{Uniform}[0,1]$, $\beta = 1.5$, $\theta(z) = 3.2z^2 - 1$, and $\epsilon = \mathrm{Normal}(0,1)$.

5.3.1 Can Semiparametric Methods Improve Upon the Sample Mean? When there are missing response data, the simulations in Wang et al. (2004) show conclusively that substantial gains in efficiency can be made over using the sample mean of the observed responses alone. In addition, if missingness depends on (X,Z), the sample mean of the observed responses will be biased.

This leaves the issue of what happens when there are no missing data. Obviously, if one thought that ϵ were normally distributed, it would be delusional to use anything other than the sample mean, it being efficient.

Theoretically, some insight can be gained by the following considerations. Suppose that X and Z are independent. Suppose also that ϵ has a symmetric density function known up to a scale parameter. Let σ_ϵ^2 be the variance of ϵ and let $\zeta \le \sigma_\epsilon^2$ be the inverse of the Fisher information for estimating the mean in the model $Y = \mu + \epsilon$. Then, it can be shown that $\mathrm{E}\{\mathcal{F}_\mathcal{B}(\cdot)\} = 0$, that $\theta_\mathcal{B}(z,\mathcal{B}) = 0$, and that the asymptotic mean squared error (MSE) efficiency of the semiparametric efficient estimate of the population mean compared to the sample mean is

$$\text{MSE efficiency of sample mean} = \frac{\beta^2\,\mathrm{var}(X) + \mathrm{var}\{\theta(Z)\} + \zeta}{\beta^2\,\mathrm{var}(X) + \mathrm{var}\{\theta(Z)\} + \sigma_\epsilon^2} \le 1.$$

Note that there are cases where ζ/σ_ϵ^2 may be quite small, especially when ϵ is heavy tailed, so that if $\beta = 0$ and $\theta(\cdot)$ is approximately constant, the MSE efficiency of the sample mean would be ζ/σ_ϵ^2, and then substantial gains in efficiency would be gained. However, the usual motivation for fitting semiparametric models is that the regression function is not constant, in which case the MSE efficiency gain will be attenuated toward 1.0, often dramatically.

We conclude then that with no missing data, in the partially linear model, substantial improvements upon the sample mean will be realized mainly when the regression errors are heavy tailed and the regression signal is slight.

We point out that in the example that motivated this work (Sec. 4), there is no simple analog to the sample mean, one that could avoid fitting models to the data.

5.3.2 How Critical Are Our Assumptions on Z? We have made two assumptions on Z: It has a compact support, and its density function is positive on that support. We have indicated in Section A.1.2 that all general articles in the semiparametric kernel-based literature make this assumption and that it appears to be critical for deriving asymptotic results for problems such as our example in Section 4. It is certainly well beyond our capabilities to weaken this assumption as it applies to problems such as our motivating example.

The condition that the density of Z be bounded away from 0 warns users that the method will deteriorate if there are a few sparsely observed outlying Z values; see below for numerical evidence of this phenomenon.

Estimation in subpopulations formed by compact subsets of Z can also be of considerable interest in practice, and these compact subsets can be chosen to avoid density spareness and meet our assumptions. A simple example might be where Z is age, and one might be interested in population summaries for those in the 40- to 60-year age range.

The partially linear model is a special case, however, because all estimates are explicit and what few Taylor expansions are necessary simplify tremendously. That is, the estimates are simple functions of sums of random variables. Cheng (1994) considered a different problem where there is no X and where local constant estimation of the nonparametric function is used, rather than local linear estimation, so that $\widehat{\theta}(z_0) =$

$\sum_{i=1}^{n} K_h(Z_i - z_0)Y_i / \sum_{i=1}^{n} K_h(Z_i - z_0)$. He indicated that the essential condition for this case is that the tails of the density of Z decay exponentially fast.

We tested this numerically in the normal-based simulation of Wang et al. (2004) with the sample size of $n = 500$: Similar results were found with $n = 100$. We use the Epanechnikov kernel and estimated the bandwidth using the following methods. First, we regressed Y and X separately on Z, using the direct plug-in (DPI) bandwidth selection method of Ruppert, Sheather, and Wand (1995) to form different estimated bandwidths on each. We then calculated the residuals from these fits and regressed the residual in Y on the residual in X to get a preliminary estimate $\widehat{\beta}_{\text{start}}$ of β. Following this, we regressed $Y - X^{\mathrm{T}}\widehat{\beta}_{\text{start}}$ on Z to get a common bandwidth, then undersmoothed it by multiplication by $n^{-2/15}$ to get a bandwidth of order $n^{-1/3}$ to eliminate bias, and then reestimated β and $\theta(\cdot)$.

We found that for various Beta distributions on Z, for example, the Beta(2, 1) that violates our assumptions, the sample mean and the semiparametric efficient method were equally efficient. The same occurs for the case that Z is normally distributed. However, when Z has a t distribution with 9 degrees of freedom, the sample mean greatly outperforms the undersmoothed estimator (MSE efficiency ≈ 2.0), which, in turn, outperformed the method that did not employ undersmoothing (MSE efficiency ≈ 2.5). An interesting quote from Ma, Chiou, and Wang (2006) is relevant here: Also operating in a partial linear model, they stated "This condition enables us to simplify asymptotic expression of certain sums of functions of variables... also excludes pathological cases where the number of observations in a window defined by the bandwidth may not increase to infinity when $n \to \infty$."

We conclude that if the design density in Z is at all heavy tailed, then the semiparametric methods will be badly affected. If such a phenomenon happens in the simple case of the partially linear model, it is likely to hold in most other cases. Otherwise, in practice at least, as long as there are no design "stragglers," the assumption is likely to be one required by the technicalities of the problem. How well this generalizes to complex nonlinear problems is unknown.

6. DISCUSSION

In this article we considered the problem of estimating population-level quantities κ_0 such as the mean, variance, and probabilities. Previous literature on the topic applies only to the simple special case of estimating a population mean in the Gaussian partially linear model. The problem was motivated by an important issue in nutritional epidemiology, estimating the distribution of usual intake for episodically consumed food, where we considered a zero-inflated mixture measurement error model: Such a problem is very different from the partially linear model, and the main interest is not in the population mean.

The key feature of the problem that distinguishes it from most work in semiparametric modeling is that the quantities of interest are based on both the parametric and the nonparametric parts of the model. Results were obtained for two general classes of semiparametric ones: (1) general semiparametric regression models depending on a function $\theta_0(Z)$ and (2) generalized linear single-index models. Within these semiparametric frameworks, we suggested a straightforward estimation methodology, derived its limiting distribution, and showed semiparametric efficiency. An interesting part of the approach is that we also allow for partially missing responses.

In the case of standard semiparametric models, we have considered the case where the unknown function $\theta_0(Z)$ is a scalar function of a scalar argument. The results, though, readily extend to the case of a multivariate function of a scalar argument.

We have also assumed that $\kappa_0 = \mathrm{E}[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$ and $\mathcal{F}(\cdot)$ are scalar, which, in principle, excludes the estimation of the population variance and standard deviation. It is, however, readily seen that both $\mathcal{F}(\cdot)$ and κ_0 or κ_{SI} can be multivariate, and, hence, the obvious modification of our estimates is semiparametric efficient.

APPENDIX: SKETCH OF TECHNICAL ARGUMENTS

In what follows, the arguments for $\mathcal{L}$ and its derivatives are in the form $\mathcal{L}(\cdot) = \mathcal{L}\{Y, X, \mathcal{B}_0, \theta_0(Z)\}$. The arguments for $\mathcal{F}$ and its derivatives are $\mathcal{F}(\cdot) = \mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}$.

Also, note that in our arguments about semiparametric efficiency, we use the symbol d exactly as it was used by Newey (1990). It does not stand for differential.

A.1 Assumptions and Remarks

A.1.1 General Considerations. The main results needed for the asymptotic distribution of our estimator are (5) and (6). The single-index model assumptions were given already in Carroll et al. (1997).

Results (5) and (6) hold under smoothness and moment conditions for the likelihood function and under smoothness and boundedness conditions for $\theta(\cdot)$. The strength of these conditions depends on the generality of the problem. For the partially linear Gaussian model of Wang et al. (2004), because the profile likelihood estimator of β is an explicit function of regressions of Y and X on Z, the conditions are simply conditions about uniform expansions for kernel regression estimators, as in, for example, Claeskens and Van Keilegom (2003). For generalized partially linear models, Severini, and Staniswalis (1994) gave a series of moment and smoothness conditions toward this end. For general likelihood problems, Claeskens and Carroll (2007) stated that the conditions needed are as follows.

(C1) The bandwidth sequence $h_n \to 0$ as $n \to \infty$ in such a way that $nh_n/\log(n) \to \infty$ and $h_n \geq \{\log(n)/n\}^{1-2/\lambda}$ for λ as in condition (C4).

(C2) The kernel function K is a symmetric, continuously differentiable pdf on $[-1, 1]$ taking on the value 0 at the boundaries. The design density $f(\cdot)$ is differentiable on an interval $B = [b_1, b_2]$, the derivative is continuous, and $\inf_{z \in B} f(z) > 0$. The function $\theta(\cdot, \mathcal{B})$ has two continuous derivatives on B and is also twice differentiable with respect to $\mathcal{B}$.

(C3) For $\mathcal{B} \neq \mathcal{B}'$, the Kullback–Leibler distance between $\mathcal{L}\{\cdot, \cdot, \mathcal{B}, \theta(\cdot, \mathcal{B})\}$ and $\mathcal{L}\{\cdot, \cdot, \mathcal{B}', \theta(\cdot, \mathcal{B}')\}$ is strictly positive. For every (y, x), third partial derivatives of $\mathcal{L}\{y, x, \mathcal{B}, \theta(z)\}$ with respect to $\mathcal{B}$ exist and are continuous in $\mathcal{B}$. The fourth partial derivative exists for almost all (y, x). Further, mixed partial derivatives $\frac{\partial^{r+s}}{\partial \mathcal{B}^r \, \partial v^s} \mathcal{L}\{y, x, \mathcal{B}, v\}|_{v=\theta(z)}$, with $0 \leq r, s \leq 4, r + s \leq 4$ exist for almost all (y, x) and $\mathrm{E}\{\sup_{\mathcal{B}} \sup_v |\frac{\partial^{r+s}}{\partial \mathcal{B}^r \, \partial v^s} \mathcal{L}\{y, x, \mathcal{B}, v\}|^2\} < \infty$. The Fisher information, $G(z)$, possesses a continuous derivative and $\inf_{z \in B} G(z) > 0$.

(C4) There exists a neighborhood $\mathcal{N}\{\mathcal{B}_0, \theta_0(z)\}$ such that

$$\max_{k=1,2} \sup_{z \in B} \left\| \sup_{(\mathcal{B},\theta) \in \mathcal{N}\{\mathcal{B}_0,\theta_0(z)\}} \left| \frac{\partial^k}{\partial \theta^k} \log\{\mathcal{L}(Y, X, \mathcal{B}, \theta)\} \right| \right\|_{\lambda,z} < \infty$$

for some $\lambda \in (2, \infty]$, where $\| \cdot \|_{\lambda,z}$ is the L^{λ}-norm, conditional on $Z = z$. Further,

$$\sup_{z \in B} \mathrm{E}_z \left[\sup_{(\mathcal{B},\theta) \in \mathcal{N}\{\mathcal{B}_0,\theta_0(z)\}} \left| \frac{\partial^3}{\partial \theta^3} \log\{\mathcal{L}(Y, X, \mathcal{B}, \theta)\} \right| \right] < \infty.$$

The preceding regularity conditions are the same as those used in a local likelihood setting where one wishes to obtain strong uniform consistency of the local likelihood estimators. Condition (C3) requires the fourth partial derivative of the log profile likelihood to have a bounded second moment; it further requires the Fisher information matrix to be invertible and to be differentiable with respect to z. Condition (C4) requires a bound on the first and second derivatives of the log profile likelihood and of the first moment of the third partial derivative, in a neighborhood of the true parameter values.

A.1.2 Compactly Supported Z. Multiple reviewers of earlier drafts of this article commented that the assumption that Z be compactly supported with density positive on this support is too strong.

However, this assumption is completely standard in the kernel-based semiparametric literature for estimation of $\mathcal{B}_0$, because it is needed for uniform expansions for estimation of $\theta_0(\cdot)$. The assumption was made in the founding articles on semiparametric likelihood estimation (Severini and Wong 1992, p. 1875, part e); the first article on generalized linear models (Severini and Staniswalis 1994, p. 511, assumption D), the first article on efficient estimation of partially linear single index models (Carroll et al. 1997, p. 485, condition 2a); and the precursor article to ours that was focused on estimation of the population mean in a partially linear model (Wang et al. 2004, p. 341, condition C.T). The uniform expansions for local likelihood given in Claeskens and Van Keilegom (2003) also make this assumption; see their page 1869, condition R0. Thus, our assumption on the design density of Z is a standard one.

The reason this assumption is made has to do with kernel technology, where proofs generally require a uniform expansion for the kernel regression or at least uniform in all observed values of Z, which is the same thing. The Nadaraya–Watson estimator, for example, has a denominator that is a density estimate, and the condition on Z stops this denominator from getting too close to 0. Ma et al. (2006), who made the same assumption (their condition 6 on p. 83), stated that it is necessary to avoid "pathological cases."

A.2 Proof of Theorem 1

A.2.1 Asymptotic Expansion. We first show (9). First, note that $\mathcal{L}$ is a log-likelihood function conditioned on (X, Z), so that we have

$$\begin{aligned} \mathrm{E}\{\delta\mathcal{L}_{\theta\theta}(\cdot)|X,Z\} &= -\mathrm{E}\{\delta\mathcal{L}_{\theta}(\cdot)\mathcal{L}_{\theta}(\cdot)|X,Z\}, \\ \mathrm{E}\{\delta\mathcal{L}_{\theta\mathcal{B}}(\cdot)|X,Z\} &= -\mathrm{E}\{\delta\mathcal{L}_{\theta}(\cdot)\mathcal{L}_{\mathcal{B}}(\cdot)|X,Z\}. \end{aligned} \tag{A.1}$$

By a Taylor expansion,

$$\begin{aligned} &n^{1/2}(\widehat{\kappa}-\kappa_0) \\ &= n^{-1/2}\sum_{i=1}^{n}\bigl[\mathcal{F}_i(\cdot)-\kappa_0 \\ &\qquad + \{\mathcal{F}_{i\mathcal{B}}(\cdot)+\mathcal{F}_{i\theta}(\cdot)\theta_{\mathcal{B}}(Z_i,\mathcal{B}_0)\}^{\mathrm{T}}(\widehat{\mathcal{B}}-\mathcal{B}_0) \\ &\qquad + \mathcal{F}_{i\theta}(\cdot)\{\widehat{\theta}(Z_i,\mathcal{B}_0)-\theta_0(Z_i)\}\bigr] + o_p(1) \\ &= \mathcal{M}_2^{\mathrm{T}} n^{1/2}(\widehat{\mathcal{B}}-\mathcal{B}_0) \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\bigl[\mathcal{F}_i(\cdot)-\kappa_0+\mathcal{F}_{i\theta}(\cdot)\{\widehat{\theta}(Z_i,\mathcal{B}_0)-\theta_0(Z_i)\}\bigr] \\ &\quad + o_p(1). \end{aligned}$$

Because $nh^4 \to 0$, using (5), we see that

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^{n}\mathcal{F}_{i\theta}(\cdot)\{\widehat{\theta}(Z_i,\mathcal{B}_0)-\theta_0(Z_i)\} \\ &= -n^{-1/2}\sum_{i=1}^{n}\mathcal{F}_{i\theta}(\cdot)n^{-1}\sum_{j=1}^{n}\delta_j K_h(Z_j-Z_i)\mathcal{L}_{j\theta}(\cdot)/\Omega(Z_i)+o_p(1) \\ &= -n^{-1/2}\sum_{i=1}^{n}\delta_i\mathcal{L}_{i\theta}(\cdot)n^{-1}\sum_{j=1}^{n}K_h(Z_j-Z_i)\mathcal{F}_{j\theta}(\cdot)/\Omega(Z_j)+o_p(1) \\ &= n^{-1/2}\sum_{i=1}^{n}\delta_i D_i(\cdot)+o_p(1), \end{aligned}$$

the last step following because the interior sum is a kernel regression converging to D_i; see Carroll et al. (1997) for details. Result (9) now follows from (6). The limiting variance (10) is an easy calculation; noting that (A.1) implies that

$$\begin{aligned} &\mathrm{E}\{\delta\epsilon\mathcal{L}_{\theta}(\cdot)|Z\} \\ &\quad = \mathrm{E}\{\delta\mathcal{L}_{\theta}(\cdot)\mathcal{L}_{\mathcal{B}}(\cdot)+\delta\mathcal{L}_{\theta}(\cdot)\mathcal{L}_{\theta}(\cdot)\theta_{\mathcal{B}}(Z,\mathcal{B}_0)|Z\} \\ &\quad = -\mathrm{E}\{\delta\mathcal{L}_{\mathcal{B}\theta}(\cdot)+\delta\mathcal{L}_{\theta\theta}(\cdot)\theta_{\mathcal{B}}(Z,\mathcal{B}_0)|Z\} \\ &\quad = 0 \end{aligned} \tag{A.2}$$

by the definition of $\theta_{\mathcal{B}}(\cdot)$ given in (7), and, hence, the last two terms in (9) are uncorrelated. We will use (A.2) repeatedly in what follows.

A.2.2 Pathwise Differentiability. We now turn to the semiparametric efficiency, using the results of Newey (1990). The relevant text of his article is in his section 3, especially through his equation (9). A parameter $\kappa = \kappa(\Theta)$ is pathwise differentiable under two conditions. The first is that $\kappa(\Theta)$ is differentiable for all smooth parametric submodels: In our case, the parametric submodels include $\mathcal{B}$, parametric submodels for $\theta(\cdot)$, and parametric submodels for the distribution of (X, Z) and the probability function $\mathrm{pr}(\delta = 1|X, Z)$. This condition is standard in the literature and fairly well required. Our motivating example clearly satisfies this condition.

The second condition is that there exists a random vector d such that $\mathrm{E}(d^{\mathrm{T}}d) < \infty$ and $\partial\kappa(\Theta)/\partial\Theta = \mathrm{E}(dS_{\Theta}^{\mathrm{T}})$, where S_{Θ} is the log-likelihood score for the parametric submodel. Newey noted that pathwise differentiability also holds if the first condition holds and if there is a regular estimator in the semiparametric problem. Generally, as Newey noted, finding a suitable random variable d can be difficult.

Assuming pathwise differentiability, which we show later, the efficient influence function is calculated by projecting d onto the nuisance tangent space. One innovation here is that we can calculate the efficient influence function without having an explicit representation for d.

Our development in Section A.2.3 will consist of two steps. In the first, we will assume pathwise differentiability and derive the efficient score function under that assumption. Using this derivation, we will then exhibit a random variable d that has the requisite property.

A.2.3 Efficiency. Recall that $\mathrm{pr}(\delta = 1|X, Z) = \pi(X, Z)$. Let $f_{X,Z}(x, z)$ be the density function of (X, Z). Let the model under consideration be denoted by M_0. Now consider a smooth parametric submodel M_λ, with $f_{X,Z}(x, z, \alpha_1)$, $\theta(z, \alpha_2)$, and $\pi(X, Z, \alpha_3)$ in place of $f_{X,Z}(x, z)$, $\theta_0(z)$, and $\pi(X, Z)$, respectively. Then, under M_λ, the log-likelihood is given by

$$\begin{aligned} L(\cdot) = \delta\mathcal{L}(\cdot) &+ \delta\log\{\pi(X,Z,\alpha_3)\} \\ &+ (1-\delta)\log\{1-\pi(X,Z,\alpha_3)\} \\ &+ \log\{f_{X,Z}(X,Z,\alpha_1)\}, \end{aligned}$$

where $(\cdot)$ represents the argument $\{Y, X, \theta(Z, \alpha_2), \mathcal{B}_0\}$. Then the score functions in this parametric submodel are given by

$$\begin{aligned} \partial L(\cdot)/\partial\mathcal{B} &= \delta\mathcal{L}_{\mathcal{B}}(\cdot), \\ \partial L(\cdot)/\partial\alpha_1 &= \partial\log\{f_{X,Z}(X,Z,\alpha_1)\}/\partial\alpha_1, \\ \partial L(\cdot)/\partial\alpha_2 &= \delta\mathcal{L}_{\theta}(\cdot)\,\partial\theta(Z,\alpha_2)/\partial\alpha_2, \\ \partial L(\cdot)/\partial\alpha_3 &= \{\partial\pi(X,Z,\alpha_3)/\partial\alpha_3\}\{\delta-\pi(X,Z,\alpha_3)\} \\ &\qquad /\bigl[\pi(X,Z,\alpha_3)\{1-\pi(X,Z,\alpha_3)\}\bigr]. \end{aligned}$$

Thus, the tangent space is spanned by the functions $\delta\mathcal{L}_{\mathcal{B}}(\cdot)^{\mathrm{T}}, s_f(x,z)$, $\delta\mathcal{L}_\theta(\cdot)g(Z)$, and $a(X,Z)\{\delta - \pi(X,Z)\}$, where $s_f(x,z)$ is any function with mean 0, while $g(z)$ and $a(X,Z)$ are any functions. For computational convenience, we rewrite the tangent space as the linear span of four subspaces $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$, and $\mathcal{T}_4$ that are orthogonal to each other (see below) and defined as follows:

$$\begin{aligned}
\mathcal{T}_1 &= \delta\mathcal{L}_{\mathcal{B}}(\cdot)^{\mathrm{T}} + \delta\mathcal{L}_\theta(\cdot)\theta_{\mathcal{B}}^{\mathrm{T}}(Z,\mathcal{B}_0),\\
\mathcal{T}_2 &= s_f(x,z),\\
\mathcal{T}_3 &= \delta\mathcal{L}_\theta(\cdot)g(Z),\\
\mathcal{T}_4 &= a(X,Z)\{\delta - \pi(X,Z)\}.
\end{aligned}$$

To show that these spaces are orthogonal, we first note that, by assumption, the data are missing at random, and, hence, $\mathrm{pr}(\delta = 1|Y,X,Z) = \pi(X,Z)$. This means that $\mathcal{T}_4$ is orthogonal to the other three spaces. Note also that, by assumption, $\mathrm{E}\{\mathcal{L}_{\mathcal{B}}(\cdot)|X,Z\} = \mathrm{E}\{\mathcal{L}_\theta(\cdot)|X,Z\} = 0$. This shows that $\mathcal{T}_2$ is orthogonal to $\mathcal{T}_1$ and $\mathcal{T}_3$. It remains to show that $\mathcal{T}_1$ and $\mathcal{T}_3$ are orthogonal, which we showed in (A.2). Thus, the spaces $\mathcal{T}_1$–$\mathcal{T}_4$ are orthogonal.

Note that, under model M_λ,

$$\kappa_0 = \int \mathcal{F}\{X, \theta(Z,\alpha_2), \mathcal{B}_0\} f_{X,Z}(x,z,\alpha_1)\,dx\,dz.$$

Hence, we have

$$\begin{aligned}
\partial\kappa_0/\partial\mathcal{B} &= \mathrm{E}\{\mathcal{F}_{\mathcal{B}}(\cdot)\},\\
\partial\kappa_0/\partial\alpha_1 &= \mathrm{E}\big[\mathcal{F}(\cdot)\,\partial\log\{f_{X,Z}(X,Z,\alpha_1)\}/\partial\alpha_1\big],\\
\partial\kappa_0/\partial\alpha_2 &= \mathrm{E}\{\mathcal{F}_\theta(\cdot)\,\partial\theta(Z,\alpha_2)/\partial\alpha_2\},\\
\partial\kappa_0/\partial\alpha_3 &= 0.
\end{aligned}$$

Now, by pathwise differentiability and equation (7) of Newey (1990), there exists a random variable d, which we need not compute, such that

$$\mathrm{E}\{\mathcal{F}_{\mathcal{B}}(\cdot)\} = \mathrm{E}\big[d\{\delta\mathcal{L}_{\mathcal{B}}(\cdot)\}\big], \tag{A.3}$$

$$\mathrm{E}\{\mathcal{F}(\cdot)s_f(X,Z)\} = \mathrm{E}\{ds_f(X,Z)\}, \tag{A.4}$$

$$\mathrm{E}\{\mathcal{F}_\theta(\cdot)g(Z)\} = \mathrm{E}\{d\delta\mathcal{L}_\theta(\cdot)g(Z)\}, \tag{A.5}$$

$$0 = \mathrm{E}\big[da(X,Z)\{\delta - \pi(X,Z)\}\big]. \tag{A.6}$$

Next, we compute the projections of d onto $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$, and $\mathcal{T}_4$. First, note that, by (A.4), for any function $s_f(X,Z)$ with expectation 0, we have $\mathrm{E}[\{d - \mathcal{F}(\cdot) + \kappa_0\}s_f(X,Z)] = 0$, which implies that the projection of d onto $\mathcal{T}_2$ is given by

$$\Pi(d|\mathcal{T}_2) = \mathcal{F}(\cdot) - \kappa_0. \tag{A.7}$$

Also, by (A.1) and (A.5), for any function $g(Z)$, we have

$$\begin{aligned}
&\mathrm{E}[\{d - \delta D(\cdot)\}\delta g(Z)\mathcal{L}_\theta(\cdot)]\\
&\quad = \mathrm{E}\{\mathcal{F}_\theta(\cdot)g(Z)\} + \mathrm{E}\big[\delta g(Z)\mathcal{L}_\theta^2(\cdot)\mathrm{E}\{\mathcal{F}_\theta(\cdot)|Z\}/\mathrm{E}\{\delta\mathcal{L}_{\theta\theta}(\cdot)|Z\}\big]\\
&\quad = 0,
\end{aligned}$$

and, hence, the projection of d onto $\mathcal{T}_3$ is given by

$$\Pi(d|\mathcal{T}_3) = \delta D(\cdot). \tag{A.8}$$

In addition, by (A.3) and (A.5),

$$\begin{aligned}
&\mathrm{E}[\{d - \mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\delta\epsilon\}\delta\epsilon^{\mathrm{T}}]\\
&\quad = \mathrm{E}\{\mathcal{F}_{\mathcal{B}}^{\mathrm{T}}(\cdot)\} - \mathrm{E}\{\mathcal{F}_\theta(\cdot)\theta_{\mathcal{B}}^{\mathrm{T}}(Z,\mathcal{B}_0)\} - \mathrm{E}(\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\delta\epsilon\epsilon^{\mathrm{T}})\\
&\quad = 0.
\end{aligned}$$

Hence, the projection of d onto $\mathcal{T}_1$ is given by

$$\Pi(d|\mathcal{T}_1) = \delta\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\epsilon. \tag{A.9}$$

Also, by (A.6), we have $\Pi(d|\mathcal{T}_4) = 0$. Using (A.7), (A.8), and (A.9), we get that the efficient influence function for κ_0 is

$$\begin{aligned}
\psi_{\mathrm{eff}} &= \Pi(d|\mathcal{T}_1) + \Pi(d|\mathcal{T}_2) + \Pi(d|\mathcal{T}_3) + \Pi(d|\mathcal{T}_4)\\
&= \mathcal{F}(\cdot) - \kappa_0 + \delta\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\epsilon + \delta D(\cdot),
\end{aligned}$$

which is the same as (9), hence completing the proof under the assumption of pathwise differentiability. In the calculations that follow, we will write $\mathcal{F}_{\mathcal{B}}$ rather than $\mathcal{F}_{\mathcal{B}}(\cdot)$, a rather than $a(X,Z)$, and so on.

We now show pathwise differentiability and, hence, semiparametric efficiency; that is, we show that (A.3)–(A.6) hold for $d = \mathcal{F} - \kappa_0 + \delta D + \delta\mathcal{M}_2^{T}\mathcal{M}_1^{-1}\epsilon$.

To verify (A.3), we see that

$$\begin{aligned}
\mathrm{E}(d\delta\mathcal{L}_{\mathcal{B}}) &= \mathrm{E}[(\mathcal{F} - \kappa_0 + \delta D + \delta\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\epsilon)\,\delta\mathcal{L}_{\mathcal{B}}]\\
&= \mathrm{E}[\delta D\mathcal{L}_{\mathcal{B}} + \delta\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\epsilon\mathcal{L}_{\mathcal{B}}]\\
&= -\mathrm{E}\left\{\mathcal{L}_\theta\frac{E(\mathcal{F}_\theta|Z)}{E(\delta\mathcal{L}_{\theta\theta}|Z)}\mathcal{L}_{\mathcal{B}}\delta\right\}\\
&\quad + \mathrm{E}\{\delta\mathcal{L}_{\mathcal{B}}(\mathcal{L}_{\mathcal{B}} + \mathcal{L}_\theta\theta_{\mathcal{B}})^{\mathrm{T}}\}\mathcal{M}_1^{-1}\mathcal{M}_2\\
&= \mathrm{E}\left\{\delta\mathcal{L}_{\theta\mathcal{B}}\frac{E(\mathcal{F}_\theta|Z)}{E(\delta\mathcal{L}_{\theta\theta}|Z)}\right\} - \mathrm{E}\{\delta(\mathcal{L}_{\mathcal{B}\mathcal{B}} + \mathcal{L}_{\mathcal{B}\theta}\theta_{\mathcal{B}}^{\mathrm{T}})\}\mathcal{M}_1^{-1}\mathcal{M}_2\\
&= -\mathrm{E}(\mathcal{F}_\theta\theta_{\mathcal{B}}) + \mathcal{M}_2\\
&= \mathrm{E}(\mathcal{F}_{\mathcal{B}}).
\end{aligned}$$

To verify (A.4), we see that

$$\begin{aligned}
\mathrm{E}(ds_f) &= \mathrm{E}\{(\mathcal{F} - \kappa_0 + \delta D + \delta\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\epsilon)s_f\}\\
&= \mathrm{E}(\mathcal{F}s_f) - \kappa_0 E(s_f) + \mathrm{E}\{\mathrm{E}(\delta D + \delta\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\epsilon|X,Z)s_f\}\\
&= \mathrm{E}(\mathcal{F}s_f).
\end{aligned}$$

To verify (A.5), we see that

$$\begin{aligned}
\mathrm{E}(d\delta\mathcal{L}_\theta g) &= \mathrm{E}\{(\mathcal{F} - \kappa_0 + \delta D + \delta\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\epsilon)\,\delta\mathcal{L}_\theta g\}\\
&= \mathrm{E}(D\mathcal{L}_\theta\delta g) + \mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}E(\epsilon\mathcal{L}_\theta\,\delta g)\\
&= -\mathrm{E}\left\{\mathcal{L}_\theta\frac{E(\mathcal{F}_\theta|Z)}{E(\delta\mathcal{L}_{\theta\theta}|Z)}\mathcal{L}_\theta\,\delta g\right\}\\
&\quad + \mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\mathrm{E}\{(\mathcal{L}_{\mathcal{B}} + \mathcal{L}_\theta\theta_{\mathcal{B}})\mathcal{L}_\theta\,\delta g\}\\
&= \mathrm{E}(\mathcal{F}_\theta g) - \mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\mathrm{E}\{(\mathcal{L}_{\mathcal{B}\theta} + \mathcal{L}_{\theta\theta}\theta_{\mathcal{B}})\,\delta g\}\\
&= \mathrm{E}(\mathcal{F}_\theta g) - \mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\mathrm{E}\{E(\delta\mathcal{L}_{\mathcal{B}\theta} + \delta\mathcal{L}_{\theta\theta}\theta_{\mathcal{B}}|Z)g\}\\
&= \mathrm{E}(\mathcal{F}_\theta g),
\end{aligned}$$

where again we have used (A.2). Finally, because the responses are missing at random, (A.6) is immediate. This completes the proof.

A.3 Sketch of Lemma 1

We have

$$\begin{aligned}
\widehat{\kappa}_{\mathrm{marg}} = n^{-1}\sum_{i=1}^{n}\Bigg[&\frac{\delta_i}{\widehat{\pi}_{\mathrm{marg}}(Z_i)}\mathcal{G}(Y_i)\\
&+ \left\{1 - \frac{\delta_i}{\widehat{\pi}_{\mathrm{marg}}(Z_i)}\right\}\mathcal{F}\{X_i, \widehat{\theta}(Z_i,\widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}\Bigg] = A_1 + A_2.
\end{aligned}$$

By calculations that are similar to those given previously, and using (11), we can readily show that

$$A_1 = n^{-1}\sum_{i=1}^n \frac{\delta_i}{\pi_{\mathrm{marg}}(Z_i)}\mathcal{G}(Y_i) - n^{-1}\sum_{i=1}^n \{\delta_i - \pi_{\mathrm{marg}}(Z_i)\}\mathrm{E}\left[\frac{\delta_i \mathcal{G}(Y_i)}{\{\pi_{\mathrm{marg}}(Z_i)\}^2}\Big| Z_i\right] + o_p(n^{-1/2}).$$

We can write

$$A_2 = B_1 + B_2 + o_p(n^{-1/2}),$$

$$B_1 = n^{-1}\sum_{i=1}^n \left\{1 - \frac{\delta_i}{\pi_{\mathrm{marg}}(Z_i)}\right\}\mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\},$$

$$B_2 = n^{-1}\sum_{i=1}^n \frac{\delta_i \mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}}{\{\pi_{\mathrm{marg}}(Z_i)\}^2}\{\widehat{\pi}_{\mathrm{marg}}(Z_i) - \pi_{\mathrm{marg}}(Z_i)\}.$$

Using (5) and (6), we can show that

$$B_1 = n^{-1}\sum_{i=1}^n \left\{1 - \frac{\delta_i}{\pi_{\mathrm{marg}}(Z_i)}\right\}\mathcal{F}_i(\cdot) + \mathcal{M}_{2,\mathrm{marg}}\mathcal{M}_1^{-1} n^{-1}\sum_{i=1}^n \delta_i \epsilon_i + n^{-1}\sum_{i=1}^n \delta_i D_{i,\mathrm{marg}}(\cdot) + o_p(n^{-1/2}).$$

Using (11) once again, we see that

$$B_2 = n^{-1}\sum_{i=1}^n \{\delta_i - \pi_{\mathrm{marg}}(Z_i)\}\mathrm{E}\left[\frac{\delta_i \mathcal{F}_i(\cdot)}{\{\pi_{\mathrm{marg}}(Z_i)\}^2}\Big| Z_i\right] + o_p(n^{-1/2}).$$

Collecting terms and noting that

$$0 = \mathrm{E}\left[\frac{\delta_i\{\mathcal{G}(Y_i) - \mathcal{F}_i(\cdot)\}}{\{\pi_{\mathrm{marg}}(Z_i)\}^2}\Big| Z_i\right]$$

proves (12).

A.4 Sketch of Lemma 2

We have

$$\widehat{\kappa} = n^{-1}\sum_{i=1}^n \left[\frac{\delta_i}{\pi(X_i, Z_i, \widehat{\zeta})}\mathcal{G}(Y_i) + \left\{1 - \frac{\delta_i}{\pi(X_i, Z_i, \widehat{\zeta})}\right\}\mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}\right] = A_1 + A_2,$$

say. By a simple Taylor series expansion,

$$A_1 = n^{-1}\sum_{i=1}^n \frac{\delta_i}{\pi(X_i, Z_i, \zeta)}\mathcal{G}(Y_i) - \mathrm{E}\left\{\frac{1}{\pi(X, Z, \zeta)}\mathcal{G}(Y)\pi_\zeta(X, Z, \zeta)\right\}^{\mathrm{T}} n^{-1}\sum_{i=1}^n \psi_{i\zeta} + o_p(n^{-1/2}).$$

In addition,

$$A_2 = B_1 + B_2 + o_p(n^{-1/2}),$$

$$B_1 = n^{-1}\sum_{i=1}^n \left\{1 - \frac{\delta_i}{\pi(X_i, Z_i, \zeta)}\right\}\mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\},$$

$$B_2 = n^{-1}\sum_{i=1}^n \frac{\delta_i \mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}}{\{\pi(X_i, Z_i, \zeta)\}^2}\pi_\zeta(X_i, Z_i, \zeta)^{\mathrm{T}}(\widehat{\zeta} - \zeta) + o_p(n^{-1/2}).$$

Using the fact that

$$0 = \mathrm{E}\left\{1 - \frac{\delta_i}{\pi(X_i, Z_i, \zeta)}\Big| X, Z\right\},$$

we can easily show that

$$B_1 = n^{-1}\sum_{i=1}^n \left\{1 - \frac{\delta_i}{\pi(X_i, Z_i, \zeta)}\right\}\mathcal{F}_i(\cdot) + o_p(n^{-1/2}).$$

It also follows that

$$B_2 = \mathrm{E}\left\{\frac{1}{\pi(X, Z, \zeta)}\mathcal{F}(\cdot)\pi_\zeta(X, Z, \zeta)\right\}^{\mathrm{T}} n^{-1}\sum_{i=1}^n \psi_{i\zeta}(\cdot) + o_p(n^{-1/2}).$$

Collecting terms and using the fact that $\mathrm{E}\{\mathcal{G}(Y)|X, Z\} = \mathcal{F}(\cdot)$, we obtain the result.

A.5 Proof of Theorem 2

A.5.1 Asymptotic Expansion. We first show the expansion (15). Recall that $\mathcal{B} = (\gamma, \beta)$. The only things that differ with the calculations of Carroll et al. (1997) is that we add terms involving δ_i and we need not worry about any constraint on γ, and, thus, we avoid terms such as their $P\alpha$ on their page 487.

In their equation (A.12), they showed that

$$n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) = n^{-1/2}\mathcal{Q}^{-1}\sum_{i=1}^n \delta_i \mathcal{N}_i \epsilon_i + o_p(1). \tag{A.10}$$

Define $H(u) = [\mathrm{E}\{\rho_2(\cdot)|U = u\}]^{-1}$. In their equation (A.13), Carroll et al. (1997) showed that

$$\begin{aligned}\widehat{\theta}(R + S^{\mathrm{T}}\widehat{\gamma}, \widehat{\mathcal{B}}) - \theta_0(R + S^{\mathrm{T}}\gamma_0) &= \theta_0^{(1)}(R + S^{\mathrm{T}}\gamma_0)S^{\mathrm{T}}(\widehat{\gamma} - \gamma_0) \\ &\quad + \widehat{\theta}(R + S^{\mathrm{T}}\gamma_0, \widehat{\mathcal{B}}) - \theta_0(R + S^{\mathrm{T}}\gamma_0) + o_p(n^{-1/2}).\end{aligned} \tag{A.11}$$

Also, in their equation (A.11), they showed that

$$\begin{aligned}\widehat{\theta}(u, \widehat{\mathcal{B}}) - \theta_0(u) &= n^{-1}\sum_{i=1}^n \delta_i K_h(U_i - u)\epsilon_i H(u)/f(u) \\ &\quad - H(u)\big[\mathrm{E}\{\delta\Lambda\rho_2(\cdot)|U = u\}\big]^{\mathrm{T}}(\widehat{\mathcal{B}} - \mathcal{B}_0) + o_p(n^{-1/2}).\end{aligned} \tag{A.12}$$

Carroll et al. (1997) did not consider an estimate of ϕ. Define

$$\mathcal{G}\{\phi, Y, X, \mathcal{B}, \theta(U)\} = \mathcal{D}_\phi(Y, \phi) - \big[Yc\{X^{\mathrm{T}}\beta + \theta(U)\} - \mathcal{C}\{c(\cdot)\}\big]/\phi^2.$$

Of course, $\mathcal{G}(\cdot)$ is the likelihood score for ϕ. If there are no arguments, $\mathcal{G} = \mathcal{G}\{\phi_0, Y, X, \mathcal{B}_0, \theta_0(R + S^{\mathrm{T}}\gamma_0)\}$. The estimating function for ϕ solves

$$0 = n^{-1/2}\sum_{i=1}^n \delta_i \mathcal{G}\{\widehat{\phi}, Y_i, X_i, \widehat{\mathcal{B}}, \widehat{\theta}(R_i + S_i^{\mathrm{T}}\widehat{\gamma}, \widehat{\mathcal{B}})\}.$$

Because $\mathcal{G}$ is a likelihood score, it follows that

$$\mathrm{E}\big[\mathcal{G}_\phi\{\phi_0, Y, X, \mathcal{B}_0, \theta_0(R + S^{\mathrm{T}}\gamma_0)\}|X, R, S\big] = -\mathrm{E}\{\mathcal{G}^2|X, R, S\}.$$

By a Taylor series,

$$\begin{aligned}\mathrm{E}(\delta\mathcal{G}^2)n^{1/2}(\widehat{\phi} - \phi_0) &= n^{-1/2}\sum_{i=1}^n \delta_i \mathcal{G}\{\phi_0, Y_i, X_i, \widehat{\mathcal{B}}, \widehat{\theta}(R_i + S_i^{\mathrm{T}}\widehat{\gamma}, \widehat{\mathcal{B}})\} + o_p(1) \\ &= n^{-1/2}\sum_{i=1}^n \delta_i \mathcal{G}_i + \mathrm{E}(\delta\mathcal{G}_{\mathcal{B}}^{\mathrm{T}})n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) \\ &\quad + n^{-1/2}\sum_{i=1}^n \delta_i \mathcal{G}_{i\theta}\{\widehat{\theta}(R_i + S_i^{\mathrm{T}}\widehat{\gamma}, \widehat{\mathcal{B}}) - \theta_0(R_i + S_i^{\mathrm{T}}\gamma_0)\} + o_p(1).\end{aligned}$$

However, it is readily verified that $\mathrm{E}(\delta\mathcal{G}_{\mathcal{B}}|X,R,S)=0$ and that $\mathrm{E}(\delta\mathcal{G}_{\theta}|X,R,S)=0$. It, thus, follows via a simple calculation using (A.11) that

$$\begin{aligned}&\mathrm{E}(\delta\mathcal{G}^2)n^{1/2}(\widehat{\phi}-\phi_0)\\&\quad=n^{-1/2}\sum_{i=1}^{n}\delta_i\mathcal{G}_i+n^{-1/2}\sum_{i=1}^{n}\delta_i\mathcal{G}_{i\theta}\{\widehat{\theta}(U_i,\mathcal{B}_0)-\theta_0(U_i)\}+o_p(1)\\&\quad=n^{-1/2}\sum_{i=1}^{n}\delta_i\mathcal{G}_i+o_p(1),\end{aligned}$$

the last step following from an application of (A.12).

With some considerable algebra, (15) now follows from calculations similar to those in Section A.2. The variance calculation follows because it is readily shown that, for any function $h(U)$,

$$0=\mathrm{E}\big[(\mathcal{N}\epsilon)\{\delta h(U)\epsilon\}\big].\tag{A.13}$$

A.5.2 Efficiency. We now turn to semiparametric efficiency. Recall that the GPLSIM follows the form (20) with $X^{\mathrm{T}}\beta_0+\theta_0(R+S^{\mathrm{T}}\gamma_0)$ and that $U=R+S^{\mathrm{T}}\gamma_0$. It is immediate that $V\{\mu(t)\}=\mu^{(1)}(t)/c^{(1)}(t)$, that $c^{(1)}(t)=\rho_1(t)$, and that $\rho_2(t)=\rho_1^2(t)V\{\mu(t)\}=c^{(1)}(t)\mu^{(1)}(t)$. We also have

$$\mathrm{E}(\epsilon|X,Z)=0,\tag{A.14}$$

$$\begin{aligned}&\mathrm{E}(\epsilon^2|X,Z)\\&\quad=\mathrm{E}\big(\big[Y-\mu\{X^{\mathrm{T}}\beta_0+\theta_0(U)\}\big]^2|X,Z\big)\big[\rho_1\{X^{\mathrm{T}}\beta_0+\theta_0(U)\}\big]^2\\&\quad=\mathrm{var}(Y|X,Z)\big[\rho_1\{X^{\mathrm{T}}\beta_0+\theta_0(U)\}\big]^2\\&\quad=\phi\rho_2(\cdot).\end{aligned}\tag{A.15}$$

Let the semiparametric model be denoted by M_0. Consider a parametric submodel M_λ with $f_{X,Z}(X,Z;\nu_1)$, $\theta_0(R+S^{\mathrm{T}}\gamma_0,\nu_2)$, and $\pi(X,Z,\nu_3)$. The joint log-likelihood of Y, X, and Z under M_λ is given by

$$\begin{aligned}L(\cdot)=&(\delta/\phi)\big(Yc\{X^{\mathrm{T}}\beta_0+\theta_0(R+S^{\mathrm{T}}\gamma_0,\nu 2)\}\\&-\mathcal{C}\big[c\{X^{\mathrm{T}}\beta_0+\theta_0(R+S^{\mathrm{T}}\gamma_0,\nu 2)\}\big]\big)\\&+\delta\mathcal{D}(Y,\phi)+\log\{f_{X,Z}(X,Z,\nu_1)\}\\&+\delta\log\{\pi(X,Z,\nu_3)\}+(1-\delta)\log\{1-\pi(X,Z,\nu_3)\}.\end{aligned}$$

As before, recall that $\epsilon=\rho_1(\cdot)\{Y-\mu(\cdot)\}=c^{(1)}(\cdot)\{Y-\mu(\cdot)\}$. Then the score functions evaluated at M_0 are

$$\begin{aligned}\partial L/\partial\beta&=\delta Xc^{(1)}(\cdot)\{Y-\mu(\cdot)\}/\phi=\delta X\epsilon/\phi,\\\partial L/\partial\gamma&=\delta\theta^{(1)}(U)Sc^{(1)}(\cdot)\{Y-\mu(\cdot)\}\phi=\delta\theta^{(1)}(U)S\epsilon/\phi,\\\partial L/\partial\nu_1&=s_f(X,Z),\\\partial L/\partial\nu_2&=\delta h(U)c^{(1)}(\cdot)\{Y-\mu(\cdot)\}/\phi=\delta h(U)\epsilon/\phi,\\\partial L/\partial\nu_3&=a(X,Z)\{\delta-\pi(X,Z)\},\\\partial L/\partial\phi&=\delta\mathcal{D}_\phi(Y,\phi)-\delta\big[Yc(\cdot)-\mathcal{C}\{c(\cdot)\}\big]/\phi^2=\delta\mathcal{G},\end{aligned}$$

where $\mathcal{D}_\phi(Y,\phi)$ is the derivative of $\mathcal{D}(Y,\phi)$ with respect to ϕ, $s_f(X,Z)$ is a mean-zero function and $h(U)$, and $a(X,Z)$ are any functions. This means that the tangent space is spanned by

$$\begin{aligned}\big(&\mathcal{T}_1=\delta\big\{S^{\mathrm{T}}\theta_0^{(1)}(U),X^{\mathrm{T}}\big\}\epsilon/\phi,\ \mathcal{T}_2=s_f(X,Z),\\&\mathcal{T}_3=\delta h(U)\epsilon/\phi,\ \mathcal{T}_4=a(X,Z)\{\delta-\pi(X,Z)\},\\&\mathcal{T}_5=\delta\mathcal{G}\big).\end{aligned}$$

An orthogonal basis of the tangent space is given by $[\mathcal{T}_1=\delta\mathcal{N}^{\mathrm{T}}\epsilon,\mathcal{T}_2=s_f(X,Z),\mathcal{T}_3=\delta h(U)\epsilon,\mathcal{T}_4=a(X,Z)\{\delta-\pi(X,Z)\}]$ and $\mathcal{T}_5=\delta\mathcal{G}$; the orthogonality is a straightforward calculation. Now notice that

$$\kappa_0=\int\mathcal{F}\{x,\theta_0(z;\nu_2),\mathcal{B}_0,\phi_0\}f_{X,Z}(x,z;\gamma)\,dx\,dz$$

and, hence,

$$\begin{aligned}\partial\kappa_0/\partial\beta&=\mathrm{E}\{\mathcal{F}_\beta(\cdot)\},\\\partial\kappa_0/\partial\gamma&=\mathrm{E}\big\{\mathcal{F}_\theta(\cdot)\theta^{(1)}(U)S\big\},\\\partial\kappa_0/\partial\nu_1&=\mathrm{E}\{\mathcal{F}(\cdot)s_f(X,Z)\},\\\partial\kappa_0/\partial\nu_2&=\mathrm{E}[\mathcal{F}_\theta(\cdot)h(Z)],\\\partial\kappa_0/\partial\nu_3&=0,\\\partial\kappa_0/\partial\phi&=\mathrm{E}\{\mathcal{F}_\phi(\cdot)\}.\end{aligned}$$

As before, we first assume pathwise differentiability to construct the efficient score. We then verify this in Section A.5.3.

By equation (7) of Newey (1990), there is a random quantity d such that

$$\mathrm{E}(d\delta X\epsilon/\phi)=\mathrm{E}\{\mathcal{F}_\beta(\cdot)\},\tag{A.16}$$

$$\mathrm{E}\big\{d\delta\theta^{(1)}(U)S\epsilon/\phi\big\}=\mathrm{E}\big\{\mathcal{F}_\theta(\cdot)\theta^{(1)}(U)S\big\},\tag{A.17}$$

$$\mathrm{E}\{ds_f(X,Z)\}=\mathrm{E}\{\mathcal{F}(\cdot)s_f(X,Z)\},\tag{A.18}$$

$$\mathrm{E}\{d\delta h(U)\epsilon/\phi\}=\mathrm{E}\{\mathcal{F}_\theta(\cdot)h(U)\},\tag{A.19}$$

$$\mathrm{E}\big[da(X,Z)\{\delta-\pi(X,Z)\}\big]=0,\tag{A.20}$$

$$\mathrm{E}(d\delta\mathcal{G})=\mathrm{E}\{\mathcal{F}_\phi(\cdot)\}.\tag{A.21}$$

Now we compute the projection of d onto the tangent space. It is immediate that $\Pi(d|\mathcal{T}_2)=\mathcal{F}(\cdot)-\kappa_0$ and that $\Pi(d|\mathcal{T}_4)=0$. Because $\mathrm{E}[\{\delta J(U)\epsilon\}\{\delta h(U)\epsilon/\phi\}]=\mathrm{E}\{h(U)\mathcal{F}_\theta(\cdot)\}$, it is readily shown that $\Pi(d|\mathcal{T}_3)=\delta J(U)\epsilon$. It is a similarly direct calculation to show that $\Pi(d|\mathcal{T}_1)=D^{\mathrm{T}}\mathcal{Q}^{-1}\delta\mathcal{N}\epsilon$. Finally, $\Pi(d|\mathcal{T}_5)=\delta\mathcal{G}\mathrm{E}\{\mathcal{F}_\phi(\cdot)\}/\mathrm{E}(\delta\mathcal{G}^2)$.

These calculations, thus, show that, assuming pathwise differentiability, the efficient influence function for κ_0 is

$$\Psi=D^{\mathrm{T}}\mathcal{Q}^{-1}\,\delta\mathcal{N}\epsilon+\mathcal{F}(\cdot)-\kappa_0+\delta J(U)\epsilon+\delta\mathcal{G}\mathrm{E}\{\mathcal{F}_\phi(\cdot)\}/\mathrm{E}(\delta\mathcal{G}^2).$$

Hence, from (15), we see that $\widehat{\kappa}_{\mathrm{SI}}$ has the semiparametric optimal influence function and is asymptotically efficient.

A.5.3 Pathwise Differentiability. For $d=D^{\mathrm{T}}\mathcal{Q}^{-1}\delta\mathcal{N}\epsilon+\mathcal{F}(\cdot)-\kappa_0+\delta J(U)\epsilon+\delta\mathcal{G}\mathrm{E}\{\mathcal{F}_\phi(\cdot)\}/\mathrm{E}(\delta\mathcal{G}^2)$, we have to show that (A.16)–(A.21) hold. Let

$$\begin{aligned}d_1&=D^{\mathrm{T}}\mathcal{Q}^{-1}\delta\mathcal{N}\epsilon,\\d_2&=\mathcal{F}(\cdot)-\kappa_0,\\d_3&=\delta J(U)\epsilon,\\d_4&=0,\\d_5&=\delta\mathcal{G}\mathrm{E}\{\mathcal{F}_\phi(\cdot)\}/\mathrm{E}(\delta\mathcal{G}^2).\end{aligned}$$

Then $d=d_1+\cdots+d_5$. Because $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, $\mathcal{T}_4$, and $\mathcal{T}_5$ are orthogonal and $d_i\in\mathcal{T}_i$ for $i=1,\dots,5$, we have

$$\mathrm{E}(d_1\mathcal{T}_1)=\mathrm{E}(D^{\mathrm{T}}\mathcal{Q}^{-1}\,\delta\mathcal{N}\mathcal{N}^{\mathrm{T}}\epsilon^2)=\phi\mathrm{E}(D^{\mathrm{T}}),\tag{A.22}$$

$$\mathrm{E}(d_2\mathcal{T}_2)=\mathrm{E}\big[\{\mathcal{F}(\cdot)-\kappa_0\}s_f(X,Z)\big]=\mathrm{E}\{\mathcal{F}(\cdot)s_f(X,Z)\},\tag{A.23}$$

$$\begin{aligned}\mathrm{E}(d_3\mathcal{T}_3)&=\mathrm{E}\{\delta J(U)h(U)\epsilon^2\}\\&=\mathrm{E}\{\pi(X,Z)J(U)h(U)\phi\rho_2(\cdot)\},\end{aligned}\tag{A.24}$$

$$\mathrm{E}(d_4\mathcal{T}_4)=0,\tag{A.25}$$

$$\mathrm{E}(d_5\mathcal{T}_5)=\mathrm{E}\big[\delta\mathcal{G}^2\mathrm{E}\{\mathcal{F}_\phi(\cdot)\}/\mathrm{E}(\delta\mathcal{G}^2)\big]=\mathrm{E}\{\mathcal{F}_\phi(\cdot)\},\tag{A.26}$$

$$\mathrm{E}(d_i\mathcal{T}_j)=0,\qquad i\neq j.\tag{A.27}$$

To verify (A.16) and (A.17), we have to prove

$$\mathrm{E}\begin{bmatrix} d\delta\theta^{(1)}(U)S\epsilon/\phi \\ d\delta X\epsilon/\phi \end{bmatrix}^{\mathrm{T}} = \mathrm{E}\begin{bmatrix} \mathcal{F}_{\theta}(\cdot)\theta^{(1)}(U)S \\ \mathcal{F}_{\beta}(\cdot) \end{bmatrix}^{\mathrm{T}}.$$

Recall that $\Lambda = \{\theta^{(1)}(U)S^{\mathrm{T}}, X^{\mathrm{T}}\}^{\mathrm{T}}$. Therefore,

$$\begin{aligned} &\mathrm{E}\begin{bmatrix} d\delta\theta^{(1)}(U)S\epsilon/\phi \\ d\delta X\epsilon/\phi \end{bmatrix}^{\mathrm{T}} \\ &\quad = \mathrm{E}(d\delta\Lambda^{\mathrm{T}}\epsilon/\phi) \\ &\quad = \mathrm{E}\bigl\{d\delta\bigl(\mathcal{N}^{\mathrm{T}} + [\mathrm{E}\{\delta\rho_2(\cdot)|U_i\}]^{-1}\mathrm{E}\{\delta_i\Lambda_i^{\mathrm{T}}\rho_2(\cdot)|U_i\}\bigr)\epsilon/\phi\bigr\} \\ &\quad = \mathrm{E}\{d\delta\mathcal{N}^{\mathrm{T}}\epsilon/\phi\} + \mathrm{E}\bigl\{d\delta\bigl(\bigl[\mathrm{E}\{\delta\rho_2(\cdot)|U_i\}\bigr]^{-1}\mathrm{E}\{\delta_i\Lambda_i^{\mathrm{T}}\rho_2(\cdot)|U_i\}\bigr)\epsilon/\phi\bigr\} \\ &\quad = \mathrm{E}(d\mathcal{T}_1/\phi) + \mathrm{E}(d\delta h(U)\epsilon/\phi) \\ &\quad = B_1 + B_2, \end{aligned}$$

where $h(U) = [\mathrm{E}\{\delta\rho_2(\cdot)|U\}]^{-1}\mathrm{E}\{\delta\Lambda^{\mathrm{T}}\rho_2(\cdot)|U\}$. Hence, using (A.22), (A.24), and (A.27), we see $B_1 = \mathrm{E}(d_1\mathcal{T}_1/\phi) = E(D^{\mathrm{T}})$ and

$$\begin{aligned} B_2 &= \mathrm{E}(d_3\delta h(U)\epsilon/\phi) \\ &= \mathrm{E}\{\pi(X,Z)J(U)h(U)\rho_2(\cdot)\} \\ &= \mathrm{E}\{\delta J(U)h(U)\rho_2(\cdot)\} \\ &= \mathrm{E}\bigl[J(U)h(U)\mathrm{E}\{\delta\rho_2(\cdot)|U\}\bigr] \\ &= \mathrm{E}\bigl\{\mathcal{F}_{\theta}(\cdot)\bigl[\mathrm{E}\{\delta\rho_2(\cdot)|U\}\bigr]^{-1}\mathrm{E}\{\delta\Lambda^{\mathrm{T}}\rho_2(\cdot)|U\}\bigr\}, \end{aligned}$$

and, hence,

$$\begin{aligned} B_1 + B_2 &= \mathrm{E}(D^{\mathrm{T}}) + \mathrm{E}\bigl\{\mathcal{F}_{\theta}(\cdot)\bigl[\mathrm{E}\{\delta\rho_2(\cdot)|U\}\bigr]^{-1}\mathrm{E}\{\delta\Lambda^{\mathrm{T}}\rho_2(\cdot)|U\}\bigr\} \\ &= \mathrm{E}\begin{bmatrix} \mathcal{F}_{\theta}(\cdot)\theta^{(1)}(U)S \\ \mathcal{F}_{\beta}(\cdot) \end{bmatrix}^{\mathrm{T}}. \end{aligned}$$

To verify (A.19), we use (A.24) and (A.27) to get

$$\begin{aligned} \mathrm{E}\{d\delta h(U)\epsilon/\phi\} &= \mathrm{E}(d_3\mathcal{T}_3/\phi) \\ &= \mathrm{E}\{\pi(X,Z)J(U)h(U)\rho_2(\cdot)\} \\ &= \mathrm{E}\{\delta J(U)h(U)\rho_2(\cdot)\} \\ &= \mathrm{E}\bigl[J(U)h(U)\mathrm{E}\{\delta\rho_2(\cdot)|U\}\bigr] \\ &= \mathrm{E}[\mathcal{F}_{\theta}(\cdot)h(U)]. \end{aligned}$$

Finally, (A.18) follows directly from (A.23) and (A.27), (A.20) follows directly from (A.25) and (A.27), and (A.21) follows directly from (A.26) and (A.27).

A.6 Proof of Lemma 3

Denote the model under consideration by M_0. Now consider any regular parametric submodel M_{λ}, with $f_{X,Z}(x,z,\alpha_1)$ and $\theta(z,\alpha_2)$ in place of $f_{X,Z}(x,z)$ and $\theta_0(z)$, respectively. For the model M_{λ}, we have the joint log-likelihood of Y, X, and Z,

$$L(y,z,x) = \mathcal{L}(\cdot) + \log\{f_{X,Z}(x,z,\alpha_1)\},$$

where $(\cdot)$ represents the argument $\{Y, X, \theta(Z,\alpha_2), \mathcal{B}_0\}$. The score functions are given by

$$\begin{aligned} \partial L/\partial\mathcal{B} &= \mathcal{L}_{\mathcal{B}}(\cdot), \\ \partial L/\partial\alpha_1 &= \partial\log\{f_{X,Z}(x,z,\alpha_1)\}/\partial\alpha_1, \\ \partial L/\partial\alpha_2 &= \mathcal{L}_{\theta}(\cdot)\,\partial\theta(z,\alpha_2)/\partial\alpha_2. \end{aligned}$$

The tangent space is spanned by $S_{\lambda} = \{\mathcal{L}_{\mathcal{B}}(\cdot)^{\mathrm{T}}, s_f(x,z)^{\mathrm{T}}, \mathcal{L}_{\theta}(\cdot)g(z)^{\mathrm{T}}\}$ or, equivalently, by

$$\begin{aligned} \mathcal{T} = \bigl\{\mathcal{T}_1 = \mathcal{L}_{\mathcal{B}}(\cdot)^{\mathrm{T}} + \mathcal{L}_{\theta}(\cdot)\theta_{\mathcal{B}}^{\mathrm{T}}(Z,\mathcal{B}_0) = \epsilon^{\mathrm{T}}, \\ \mathcal{T}_2 = s_f(X,Z)^{\mathrm{T}}, \mathcal{T}_3 = g(Z)^{\mathrm{T}}\mathcal{L}_{\theta}(\cdot)\bigr\}, \end{aligned}$$

where $s_f(x,z)$ is any function with expectation 0 and $g(z)$ is any function of z. Note that, under model M_{λ}, $\kappa_0 = \int Y\exp\{\mathcal{L}(\cdot)\}f_{X,Z}(x,z,\alpha_1)\,dy\,dx\,dz$. Hence, we have

$$\begin{aligned} \partial\kappa_0/\partial\mathcal{B} &= \mathrm{E}\{Y\mathcal{L}_{\mathcal{B}}(\cdot)\} = \mathrm{E}\{Y(\partial L/\partial\mathcal{B})\}, \\ \partial\kappa_0/\partial\alpha_1 &= \mathrm{E}\{Ys_f(X,Z)\} = \mathrm{E}\{Y(\partial L/\partial\alpha_1)\}, \\ \partial\kappa_0/\partial\alpha_2 &= \mathrm{E}\{Y\mathcal{L}_{\theta}(\cdot)g(Z)\} = \mathrm{E}\{Y(\partial L/\partial\alpha_2)\}. \end{aligned}$$

Hence, we see that κ_0 is pathwise differentiable and $d = Y$. The projection of d onto $\mathcal{T}$ is then given by

$$\begin{aligned} \Pi(d|\mathcal{T}_1) &= \mathrm{E}(Y\epsilon^{\mathrm{T}})\mathcal{M}_1^{-1}\epsilon, \\ \Pi(d|\mathcal{T}_2) &= \mathrm{E}(Y|X,Z) - \kappa_0, \\ \Pi(d|\mathcal{T}_3) &= \mathcal{L}_{\theta}(\cdot)\mathrm{E}\{Y\mathcal{L}_{\theta}(\cdot)|Z\}/\mathrm{E}\bigl[\{\mathcal{L}_{\theta}(\cdot)\}^2|Z\bigr], \end{aligned}$$

and, hence, the efficient influence function is

$$\begin{aligned} \Pi(d|\mathcal{T}) = \mathrm{E}(Y\epsilon^{\mathrm{T}})\mathcal{M}_1^{-1}\epsilon + \{\mathrm{E}(Y|X,Z) - \kappa_0\} \\ + \mathcal{L}_{\theta}(\cdot)\mathrm{E}\{Y\mathcal{L}_{\theta}(\cdot)|Z\}/\mathrm{E}\bigl[\{\mathcal{L}_{\theta}(\cdot)\}^2|Z\bigr]. \end{aligned}$$

But we see that the influence function of the sample mean is $Y - \kappa_0$. Hence, the sample mean is semiparametric efficient if and only if (19) holds.

A.7 Proof of Lemma 4

It suffices to consider only the case that $\phi = 1$ is known, because the estimates of β_0 and $\theta_0(z)$ do not depend on the value of φ.

It is convenient to write $c\{\eta(x,z)\}$ as $d(x,z)$ and to denote the derivative of $d(x,z)$ with respect to $\theta_0(z)$ as $d_{\theta}(x,z)$. Note that the derivative with respect to β is $d_{\beta}(x,z) = Xd_{\theta}(x,z)$. Direct calculations show that

$$\begin{aligned} \mathcal{L}_{\theta}(\cdot) &= d_{\theta}(X,Z)\{Y - \mu(X,Z)\}, \\ \mathcal{L}_{\beta}(\cdot) &= Xd_{\theta}(X,Z)\{Y - \mu(X,Z)\}, \\ \theta_{\beta}(Z) &= -\frac{\mathrm{E}[Xd_{\theta}^2(X,Z)V\{\mu(X,Z)\}|Z]}{\mathrm{E}[d_{\theta}^2(X,Z)V\{\mu(X,Z)\}|Z]}, \\ \epsilon &= \{X + \theta_{\beta}(Z)\}d_{\theta}(X,Z)\{Y - \mu(X,Z)\}, \\ \mathrm{E}(\mathrm{Y}\epsilon) &= \mathrm{E}\bigl[\{X + \theta_{\beta}(Z)\}d_{\theta}(X,Z)V\{\mu(X,Z)\}\bigr], \end{aligned}$$

$$\begin{aligned} &\mathcal{L}_{\theta}(\cdot)\frac{\mathrm{E}\{Y\mathcal{L}_{\theta}(\cdot)|Z\}}{\mathrm{E}\{\mathcal{L}_{\theta}^2(\cdot)|Z\}} \\ &\quad = \{Y - \mu(X,Z)\}d_{\theta}(X,Z)\frac{\mathrm{E}[d_{\theta}(X,Z)V\{\mu(X,Z)\}|Z]}{\mathrm{E}[d_{\theta}^2(X,Z)V\{\mu(X,Z)\}|Z]}. \end{aligned}$$

If $d_{\theta}(x,z)$ depends only on z, then $\theta_{\beta}(Z) = -\mathrm{E}[XV\{\mu(X,Z)\}|Z]/\mathrm{E}[V\{\mu(X,Z)\}|Z]$, $\mathrm{E}(Y\epsilon) = 0$, and

$$1 \equiv d_{\theta}(X,Z)\frac{\mathrm{E}[d_{\theta}(X,Z)V\{\mu(X,Z)\}|Z]}{\mathrm{E}[d_{\theta}^2(X,Z)V\{\mu(X,Z)\}|Z]}, \tag{A.28}$$

so that by Lemma 3 the sample mean is semiparametric efficient.

The cases where the sample mean is not semiparametric efficient are the following. Consider problems not of canonical exponential forms. First of all, it cannot be semiparametric efficient if $\mathrm{E}(Y\epsilon) = 0$ and $d_{\theta}(x,z)$ depends on x, for then (A.28) fails. This means then that $d_{\theta}(x,z)$ cannot be a function of x; that is, the data must follow a canonical exponential family.

If $\mathrm{E}(Y\epsilon) \neq 0$, we must have

$$\begin{aligned} 1 \equiv d_{\theta}(X,Z)\biggl(\mathrm{E}(Y\epsilon^{\mathrm{T}})\mathcal{M}_1^{-1}\{X + \theta_{\beta}(Z)\} \\ + \frac{\mathrm{E}[d_{\theta}(X,Z)V\{\mu(X,Z)\}|Z]}{\mathrm{E}[d_{\theta}^2(X,Z)V\{\mu(X,Z)\}|Z]}\biggr). \end{aligned} \tag{A.29}$$

Examples where (A.29) fails to hold are easily constructed. Because the term inside the parentheses in (A.29) is linear in X and a function of Z, (A.29) can only hold, in principle, if $d(x, z) = c\{x^{\mathrm{T}}\beta + \theta(z)\} = a + b \log\{x^{\mathrm{T}}\beta + \theta(z)\}$ for known constants (a, b).

[Received January 2006. Revised August 2006.]

REFERENCES

Bickel, P. J. (1982), "On Adaptive Estimation," *The Annals of Statistics*, 10, 647–671.

Block, G., Hartman, A. M., Dresser, C. M., Carroll, M. D., Gannon, J., and Gardner, L. (1986), "A Data-Based Approach to Diet Questionnaire Design and Testing," *American Journal of Epidemiology*, 124, 453–469.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Chen, X., Linton, O., and Van Keilegom, I. (2003), "Estimation of Semiparametric Models When the Criterion Function Is not Smooth," *Econometrica*, 71, 1591–1608.

Cheng, P. E. (1994), "Nonparametric Estimation of Mean Functionals With Data Missing at Random," *Journal of the American Statistical Association*, 89, 81–87.

Claeskens, G., and Carroll, R. J. (2007), "Post-Model Selection Inference in Semiparametric Models," *Biometrica*, to appear.

Claeskens, G., and Van Keilegom, I. (2003), "Bootstrap Confidence Bands for Regression Curves and Their Derivatives," *The Annals of Statistics*, 31, 1852–1884.

Flegal, K. M., Carroll, M. D., Ogden, C. L., and Johnson, C. L. (2002), "Prevalence and Trends in Obesity Among US Adults, 1999–2000," *Journal of the American Medical Association*, 288, 1723–1727.

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *The Annals of Statistics*, 21, 157–178.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, Berlin: Springer-Verlag.

Härdle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 408, 986–995.

Ma, Y., Chiou, J. M., and Wang, N. (2006), "Efficient Semiparametric Estimator for Heteroscedastic Partially Linear Models," *Biometrika*, 93, 75–84.

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

Newey, W. K., Hsieh, F., and Robins, J. M. (2004), "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica*, 72, 947–962.

Powell, J. L., and Stoker, T. M. (1996), "Optimal Bandwidth Choice for Density-Weighted Averages," *Journal of Econometrics*, 75, 291–316.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270; Corrigenda (1996), 91, 1380.

Sepanski, J. H., Knickerbocker, R., and Carroll, R. J. (1994), "A Semiparametric Correction for Attenuation," *Journal of the American Statistical Association*, 89, 1366–1373.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasilikelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Severini, T. A., and Wong, W. H. (1992), "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001), "Comparative Validation of the Block, Willett and National Cancer Institute Food Frequency Questionnaires: The Eating at America's Table Study," *American Journal of Epidemiology*, 154, 1089–1099.

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002), "Analysis of Repeated Measures Data With Clumping at Zero," *Statistical Methods in Medical Research*, 11, 341–355.

Wang, Q., Linton, O., and Härdle, W. (2004), "Semiparametric Regression Analysis With Missing Response at Random," *Journal of the American Statistical Association*, 99, 334–345.

Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekens, C. H., and Speizer, F. E. (1985), "Reproducibility and Validity of a Semiquantitative Food Frequency Questionnaire," *American Journal of Epidemiology*, 122, 51–65.

Woteki, C. E. (2003), "Integrated NHANES: Uses in National Policy," *Journal of Nutrition*, 133, 582S–584S.