Electronic Journal of Statistics
Vol. 6 (2012) 1200–1219
ISSN: 1935-7524
DOI: 10.1214/12-EJS708

Efficient parameter estimation in regression with missing responses

Ursula U. Müller

Department of Statistics
Texas A&M University
College Station, TX 77843-3143, USA
e-mail: uschi@stat.tamu.edu
url: www.stat.tamu.edu/~uschi

and

Ingrid Van Keilegom

Institut de statistique
Université catholique de Louvain
Voie du Roman Pays 20
B-1348 Louvain-la-Neuve, Belgium
e-mail: ingrid.vankeilegom@uclouvain.be
url: www.stat.ucl.ac.be/ISpersonnel/vankeile

Abstract: We discuss efficient estimation in regression models that are defined by a finite-dimensional parametric constraint. This includes a variety of regression models, in particular the basic nonlinear regression model and quasi-likelihood regression. We are interested in the case where responses are missing at random. This is a popular research topic and various methods have been proposed in the literature. However, many of them are complicated and are not shown to be efficient. The method presented here is, in contrast, very simple – we use an estimating equation that does not impute missing responses – and we also prove that it is efficient if an appropriate weight matrix is selected. Finally, we show that this weight matrix can be replaced by a consistent estimator without losing the efficiency property.

AMS 2010 subject classifications: Primary 62F12, 62G05; secondary 62J02.
Keywords and phrases: Efficiency, influence function, missing at random, nonlinear regression, nuisance function, parametric regression, quantile regression, quasi-likelihood regression.

Received December 2011.

Contents

1. Introduction

In this article we consider a general class of regression models that can be specified as a finite-dimensional parametric constraint,

$$E\{a_\vartheta(X,Y)|X\} = 0, \quad a_\vartheta = (a_{\vartheta 1}, \ldots, a_{\vartheta k})^\top, \tag{1.1}$$

with parameter ϑ belonging to the interior of some compact parameter space $\Theta \subset \mathbb{R}^p$. This means in particular that the parameter ϑ is defined as a solution of a system of equations. Since there can be more than one solution of (1.1), or no solution at all, we will assume in the following that a solution ϑ exists and that it is unique. The variables X and Y are multi-dimensional, and we allow that Y is not always observed. In this setting it is possible to derive efficient estimators of ϑ as solutions of an appropriately chosen set of estimating equations, which is what we pursue in this article.

The general model (1.1) covers the regression model given by

$$Y = r_\vartheta(X) + \varepsilon,$$

with $E(\varepsilon|X) = 0$, which we call a "nonlinear regression model"; see below for more explanations. But model (1.1) also covers more complicated models, such as the quasi-likelihood model which is specified by the two-dimensional conditional constraint

$$E\{a_\vartheta(X,Y)|X\} = 0, \quad a_\vartheta(X,Y) = \begin{bmatrix} Y - r_\vartheta(X) \\ \{Y - r_\vartheta(X)\}^2 - v_\vartheta(X) \end{bmatrix}, \tag{1.2}$$

and the quantile regression model, which is defined by

$$a_\vartheta(X,Y) = p - \mathbf{1}\{Y - r_\vartheta(X) < 0\}.$$

In these examples Y is a one-dimensional response variable and X a vector of covariates.

Let us first take a closer look at the simple but important case of a nonlinear regression model, which includes the linear regression model as a special case with $r_\vartheta(X) = \vartheta^\top X$. We should emphasize that we are considering models that are solely specified by a conditional constraint of the form (1.1). This means that for the nonlinear regression model we do *not* assume a parametric form for the distribution of the covariate vector X or the error variable $\varepsilon = Y - r_\vartheta(X)$. We also do *not* assume that X and ε are independent – we only assume that the errors are conditionally centered given the covariates, $E(\varepsilon|X) = 0$. Since this and the parametric form of the regression function is all the information given,

the nonlinear regression model can be described by the simple one-dimensional constraint

$$E\{Y - r_\vartheta(X)|X\} = 0, \tag{1.3}$$

which is indeed a special case of (1.1). It is also worth noting that it is not necessary here to introduce an error variable ε.

Efficient estimation of ϑ in the complete data case has been studied by various authors. We refer first of all to Chapter 4 of Tsiatis (2006 [19]), who studied the nonlinear model (1.3) in detail, including the derivation of the efficient score function, and the adaptive estimation of the weight in the estimating equation. Müller (2007 [8]) considers weighted least squares estimators in possibly misspecified regression models and derives as a special case an efficient estimator for ϑ in the regression model above. The characterization sketched in that paper is analogous to that obtained in Müller and Wefelmeyer (2002 [11]) for *autoregressive* models satisfying a parametric constraint. A (different) derivation of the asymptotic variance bound is sketched in Chamberlain (1987 [3]), with generalizations in Chamberlain (1992 [4]). Two review articles are Newey (1990 [12], 1993 [13]).

Estimating ϑ efficiently is quite complicated in the *classical* regression setting, which assumes that covariates and errors are independent. The independence assumption is a structural assumption about the model, and must be incorporated by constructing an *efficient* estimator. Efficient estimation of the parameter in the classical setting with a *linear* regression function has been studied by Bickel (1982 [1]), Koul and Susarla (1983 [7]), and Schick (1987 [17], 1993 [18]). Schick (1993 [18]) also considers general semiparametric regression models with independent covariates and errors. He uses a preliminary estimator of ϑ and an estimator of the efficient influence function to construct an efficient estimator for ϑ. A further approach, which requires weaker conditions, is in Forrester et al. (2003 [6]).

All the above articles study estimation of ϑ when no data are missing. We are interested in the case when responses are possibly missing, in particular when responses are missing at random (MAR). This means that we only observe Y in those cases where some indicator δ equals one, and the indicator δ is conditionally independent of Y given X. This assumption is useful when information in the form of covariate data is available to explain the missingness. In that case we can estimate the *propensity score* $\pi(X) = P(\delta = 1|X)$ and the missingness mechanism is called *ignorable*.

A considerable amount of work has been done on regression models with responses missing at random, but little has been done on efficient estimation. Robins et al. (1994 [15]), for example, assume a parametric model for $\pi(X)$ (or that $\pi(X)$ is known), and estimate the regression parameters efficiently by solving an inverse probability weighted estimating equation. Also in Robins et al. (1995 [16]) a parametric model for $\pi(X)$ is assumed, which is conceptually quite different from a nonparametric model for $\pi(X)$, which will be assumed in this paper. The authors allow the response and the covariates to be varying over time. On the other hand, they do not establish the efficiency of their estimator.

Efficient estimation of ϑ in model (1.3) above, with MAR responses and *with* independence of covariates and errors, is studied in Müller (2009 [9]). There the influence function of an efficient estimator for ϑ is derived and the construction of an efficient estimator is discussed. Perhaps surprisingly, this can be done in the same way as in the complete data case: by simply omitting the covariates associated with missing responses and by using only the data (X, Y) that are complete. We show in this paper that the same applies for our regression model where the independence assumption is not imposed: ϑ can be estimated efficiently by using a weighted least squares estimator which uses only the data pairs (X_i, Y_i) for which response values are at hand. More precisely, we will show that the solution $\hat{\vartheta}$ of the estimating equation

$$\sum_{i=1}^{n} \delta_i \dot{r}_\theta(X_i)^\top \sigma^{-2}(X_i)\{Y_i - r_\theta(X_i)\} = 0 \tag{1.4}$$

with respect to θ is efficient. Here $\dot{r}_\theta$ is the vector of partial derivatives with respect to the components of θ, θ is an arbitrary value in the parameter space Θ, and $\sigma^2(X)$ is the conditional error variance given the covariates, $\sigma^2(X) = E[\{Y - r_\theta(X)\}^2|X]$. The conditional variance function depends on θ, $\sigma^2(X) = \sigma^2_\theta(X)$, but since we do not model it parametrically we prefer to write it without the subscript θ. This will also be helpful to distinguish the conditional variance in the nonlinear regression model from the conditional variance in more complex models such as the quasi-likelihood model, where we also assume a parametric model $v_\theta(X)$ for the variance function, $\sigma^2(X) = v_\theta(X)$. Note that the estimating equation above is called *undetermined* since $\sigma^2(X)$ is unknown. Estimation of σ^2 is addressed in Section 3.

To our knowledge, there is no published work where efficiency of the above estimator is proved or where an efficient estimator is provided for the nonlinear regression model (1.3) with MAR responses. We will therefore pay particular attention to this model. This is also motivated by the fact that model (1.3) is a *fundamental* model and therefore important. Although Tsiatis (2006 [19]) studied model (1.3) in great detail for the case when all data are completely observed, and although one can argue that the consistency of his estimation method should remain valid with MAR responses, it is not at all clear whether the *efficiency* of his method can be carried over to the MAR case. This needs careful investigation.

The efficient estimator for ϑ in model (1.3) can also be used as a point of reference for related approaches in more complex models with MAR responses. Wang and Sun (2007 [21]), for example, compare three estimators for the regression function in a partly linear model, which coincides with model (1.3) if we assume that the unknown smooth part of the regression function is zero and if r_θ is linear, $r_\theta(X) = \theta^\top X$. Another example is Wang et al. (2010 [22]), who consider a single index model with regression function $g(\theta^\top X)$ which would be our model (1.3) with a linear regression function if g were known to be the identity.

The conditional constraint (1.1) implies that the unconditional constraint $E\{a_\vartheta(X, Y)\}$ is zero, which is the model considered by Zhou et al. (2008 [23])

and by Wang and Chen (2009 [20]). In both articles the proposed estimators are similar to our estimator in that they are solutions of an estimating equation – but more complex. In contrast to our approach, the 'missing' terms of the estimating equation are replaced by nonparametric estimators of the conditional expectation $E\{a_\vartheta(X,Y)|X\}$ (which estimates zero if our model is in fact true). The estimation of this conditional expectation requires the careful selection of a smoothing parameter. These procedures are therefore more complicated than our method. A general efficiency statement is not established, but possible variance reductions are discussed. Our method, in contrast, is very simple since it exploits the conditional constraint – which suggests a weighted estimating equation. Since our model class is characterized by a conditional constraint we cover many basic regression models, including the nonlinear regression model. Above all we show that our method is efficient if we work with an *optimal* weight matrix. Estimating these optimal weights may require the use of smoothing techniques, but choosing the smoothness parameter is less important here since only consistency (without a specific rate) is needed (see Section 3).

The paper is organized as follows. In the next section we define our estimator of ϑ and show its asymptotic normality. Section 3 discusses a number of special cases of the general theory and provides a small simulation study. The efficiency of our method is established in Section 4. Finally, Section 5 contains some concluding remarks and a discussion of open questions.

2. Estimation

The motivation for our estimating equation comes from the nonlinear regression example. A simple estimator for this model (modified for the missing response setting) is the least squares estimator, which is the minimizer of $\sum_{i=1}^n \delta_i\{Y_i - r_\theta(X_i)\}^2$ with respect to θ. It is obtained by solving the weighted estimating equation

$$\sum_{i=1}^n \delta_i \dot{r}_\theta(X_i)^\top \{Y_i - r_\theta(X_i)\} = 0 \tag{2.1}$$

with respect to θ, where the weight vector $\dot{r}_\theta(\cdot)^\top$ is the $p \times 1$ vector of partial derivatives of $r_\theta(\cdot)$ with respect to θ. Since the nonlinear regression model has a simple structure – in particular there is no form for the variance assumed – it is intuitively clear that more weight should be put on data points (X_i, Y_i) when the variance is small and less weight when the variance is large. It appears to make sense to improve the usual least squares estimator by choosing weights $W_\theta(X) = \dot{r}_\theta(X)^\top \sigma^{-2}(X)$, which now additionally involve the conditional variance $\sigma^2(X) = E[\{Y - r_\vartheta(X)\}^2|X]$. Both approaches incorporate the gradient $\dot{r}_\theta$ and can therefore be regarded as weighted least squares estimators, i.e. as solutions of

$$\sum_{i=1}^n \delta_i W_\theta(X_i)\{Y_i - r_\theta(X_i)\} = 0.$$

Our estimator for the parameter vector ϑ in the conditionally constrained model (1.1) is defined analogously as a solution $\hat{\vartheta}$ of

$$\sum_{i=1}^{n} \delta_i W_\theta(X_i) a_\theta(X_i, Y_i) = 0, \tag{2.2}$$

where W_θ is a $p \times k$ weight matrix. Sometimes the system of equations (2.2) does not have a solution. This is often the case for quantile regression or any other model leading to non-smooth criterion functions. In that case we replace (2.2) by the minimizer of

$$\Big\| \sum_{i=1}^{n} \delta_i W_\theta(X_i) a_\theta(X_i, Y_i) \Big\|$$

with respect to θ, where $\|\cdot\|$ is the Euclidean norm. In the nonlinear regression model (1.3) W_θ is just a vector, and in the quasi-likelihood model (1.2) W_θ is a $p \times 2$ matrix. The estimating equation is unbiased for any choice of $W_\theta(X)$ since it is easy to verify that $E\{\delta W_\vartheta(X) a_\vartheta(X, Y)\} = 0$: using the MAR assumption on the responses, which postulates that the indicators and the responses are conditionally independent given the covariates, we obtain

$$\begin{aligned} E\{\delta W_\vartheta(X) a_\vartheta(X, Y)\} &= E\big[W_\vartheta(X) E\{\delta a_\vartheta(X, Y)|X\}\big] \\ &= E\big[W_\vartheta(X) E(\delta|X) E\{a_\vartheta(X, Y)|X\}\big] = 0. \end{aligned}$$

Note that we explicitly use that $E\{a_\vartheta(X, Y)|X\} = 0$, which is the only model structure that we assume. This suggests that the above approach *could* yield an efficient estimator. In particular, it becomes evident that the preconditions for obtaining an appealing (simple *and* possibly efficient) estimator are ideal if a constrained model of the form (1.1) can be assumed, and if the missingness of the responses can be explained by covariates.

Whether a solution $\hat{\vartheta}$ of the above equation is efficient or not will depend on the choice of W_θ. Our approach to find the *optimal* weight matrix was to derive the efficient influence function first (see Section 4 on efficiency), which is

$$\delta I^{-1} \ell_\vartheta(X, Y)$$

with

$$\begin{aligned} &\ell_\theta(X, Y) = -W_\theta(X) a_\theta(X, Y), \qquad I = E\{\delta \ell_\vartheta(X, Y) \ell_\vartheta(X, Y)^\top\}, \\ &W_\theta(X) = \big[\frac{\partial}{\partial \theta} E\{a_\theta(X, Y)|X\}\big]^\top E\{a_\theta(X, Y) a_\theta(X, Y)^\top|X\}^{-1}, \end{aligned} \tag{2.3}$$

where $\partial/(\partial\theta) E\{a_\theta(X, Y)|X\}$ is of dimension $k \times p$. Here we only assume that the expectation is differentiable with respect to θ. In many models we can even assume that a_θ is differentiable. If this is the case we will write briefly $E\{\dot{a}_\theta(X, Y)|X\}$ instead of $\partial/(\partial\theta) E\{a_\theta(X, Y)|X\}$.

For reasons of clarity we set

$$L(\theta) = E\{\delta \ell_\theta(X, Y)\}, \qquad L_n(\theta) = n^{-1} \sum_{i=1}^{n} \delta_i \ell_\theta(X_i, Y_i).$$

We have shown that $L(\vartheta) = 0$ and therefore estimate ϑ by the solution $\hat{\vartheta}$ of the corresponding estimating equation

$$L_n(\theta) = 0, \tag{2.4}$$

with respect to θ, or, if (2.4) does not have a solution, estimate ϑ by

$$\hat{\vartheta} = \text{argmin}_{\theta \in \Theta} \|L_n(\theta)\|. \tag{2.5}$$

It should be pointed out that the resulting estimator $\hat{\vartheta}$ only uses completely observed pairs (X_i, Y_i) – in particular it discards information that is given in the form of (observed) covariates X_i.

It remains to be shown that the influence function of $\hat{\vartheta}$ is indeed of the required form, i.e. we have to derive the asymptotic expansion of the estimator. Since our estimator is the solution of an estimating equation, this is a standard result for M-estimators, and rests on a Taylor expansion. Here we provide the statement under fairly weak conditions, using Theorem 3.3 in Pakes and Pollard (1989 [14]). The conditions in this theorem include the case where the criterion function $L_n(\theta)$ is not smooth. It is also interesting to note that, regardless of the dimension of the original set of defining equations (namely of $E\{a_\vartheta(X, Y)|X\} = 0$), the dimension of the final estimating function $L_n(\theta)$ always equals p – the dimension of θ.

Theorem 2.1. *Suppose that*

(i) $\hat{\vartheta} - \vartheta = o_p(1)$.
(ii) ϑ *is the unique solution of* $L(\theta) = 0$.
(iii) $L(\theta)$ *is differentiable at* $\theta = \vartheta$*; the matrix* I *is of full rank and, for almost every* x*, the matrix* $E\{a_\vartheta(X, Y)a_\vartheta(X, Y)^\top | X = x\}$ *is also of full rank.*
(iv) For all $j = 1, \ldots, p$*,* $\delta \ell_{\theta,j}(X, Y)$ *is locally uniformly* L_2*-continuous with respect to* θ *in the sense that*

$$E\Big[\sup_{\theta_2 : \|\theta_1 - \theta_2\| < \alpha} \delta\{\ell_{\theta_1,j}(X, Y) - \ell_{\theta_2,j}(X, Y)\}^2\Big] \le K_j \alpha^{2s_j}$$

for all $\theta_1 \in \Theta$*, for all* $\alpha = o(1)$*, and for some constants* $s_j \in (0, 1]$*,* $K_j > 0$.

Then

(a) the estimator $\hat{\vartheta}$ *has the stochastic expansion*

$$n^{1/2}(\hat{\vartheta} - \vartheta) = I^{-1} n^{-1/2} \sum_{i=1}^{n} \delta_i \ell_\vartheta(X_i, Y_i) + o_p(1) \tag{2.6}$$

and is asymptotically normally distributed with covariance matrix

$$E[I^{-1}\delta \ell_\vartheta(X, Y)\{I^{-1}\delta \ell_\vartheta(X, Y)\}^\top] = I^{-1},$$

(b) the estimator $\hat{\vartheta}$ is efficient for estimating ϑ, provided the joint distribution of (X, Y) satisfies the mild regularity conditions stated in Section 4.

Part (b) is important: it shows that efficiency can be obtained without using complicated procedures to replace the missing responses with estimators. Our method, which completely discards the missing observations, is easy to compute and is efficient if the weight matrix is suitably chosen.

Remark 1. Condition *(i)* can be easily shown using standard results (see e.g. Theorem 3.1 or Corollary 3.2 in Pakes and Pollard, 1989 [14]), whereas condition *(ii)* is needed for identifiability reasons. The differentiability condition in *(iii)* is imposed on the function $L(\theta)$, which will in many cases be smooth even if the function ℓ_θ is not smooth in θ. Finally, note that condition *(iv)* also allows for discontinuous functions ℓ_θ such as sign and indicator functions. In the smooth case, *(iv)* can be replaced by the following more direct condition:

(iv)' For all $j = 1, \ldots, p$, the function $(\delta, x, y) \to \delta \ell_{\theta,j}(x, y)$ is Hölder continuous with respect to θ in the sense that

$$\delta |\ell_{\theta_1,j}(x, y) - \ell_{\theta_2,j}(x, y)| \le b_j(\delta, x, y) \|\theta_1 - \theta_2\|^{s_j}$$

for some constant $s_j \in (0, 1]$ and a measurable function b_j with finite second moment $E[b_j^2(\delta, X, Y)]$.

Remark 2. By part (b) of Theorem 2.1, an efficient estimator $\hat{\vartheta}$ of ϑ satisfies (2.6), i.e. it has influence function $I^{-1}\delta\ell_\vartheta(X, Y)$. The classical approach to constructing an efficient estimator is to start with an initial inefficient estimator of ϑ and to improve it by adding an estimator of the influence function, with appropriate estimators for I and ℓ (see, for example, Bickel et al., 1998 [2]). This construction does not, however, take advantage of the special feature of our model and is not recommended: our method only requires solving (2.4), or, more generally, (2.5). In particular we do not need to estimate I.

Proof of Theorem 2.1. We have to verify that the stochastic expansion (2.6) in part (a) holds true. The proof of (b) is in Section 4 where we show that $I^{-1}\delta\ell_\vartheta(X, Y)$ is the efficient influence function for estimating ϑ (see the characterization at the end of Section 4). In Section 4 we work with some additional notation for the (rather technical) derivation, to keep the presentation clear. For example we write Q_x for the conditional expectation given $X = x$. It is easy to verify that $\ell_\theta(x, y) = -\{\partial/(\partial\theta) Q_x(a_\theta)\}^\top Q_x(a_\theta a_\theta^\top)^{-1} a_\theta(x, y)$ from Section 4 and $\ell_\theta(x, y) = -W_\theta(x) a_\theta(x, y)$ from above (with W_θ given in (2.3)) are identical.

We prove (2.6) by showing that the conditions of Theorem 3.3 in Pakes and Pollard (1989 [14]) are satisfied. Here the criterion function is $\delta\ell_\theta(X, Y)$. It can quickly be verified that these conditions hold true, provided that:

(1) Our matrix I and Pakes and Pollard's matrix $-\Gamma$ are the same, where

$$\Gamma = \frac{\partial}{\partial\theta} E\{\delta\ell_\theta(X, Y)\}\Big|_{\theta=\vartheta}.$$

Hence we must show that

$$E\{\delta\ell_\vartheta(X,Y)\ell_\vartheta(X,Y)^\top\} = -\frac{\partial}{\partial\theta}E\{\delta\ell_\theta(X,Y)\}\Big|_{\theta=\vartheta}. \tag{2.7}$$

(2) Condition *(iv)* above implies condition *(iii)* of Theorem 3.3 in Pakes and Pollard (1989 [14]).

Let us begin with the matrix I on the left-hand side of (2.7). For reasons of clarity we use some notation from Section 4 and set $\dot{Q}_x(a_\vartheta) = \partial/(\partial\theta)Q_x(a_\theta)|_{\theta=\vartheta}$. This lets us avoid writing $\partial/(\partial\theta)Q_x(a_\vartheta)$ for the gradient which could be confusing since the conditional constraint $Q_x(a_\vartheta)$ is zero. We have

$$\begin{aligned}
I &= E\{\delta W_\vartheta(X)a_\vartheta(X,Y)a_\vartheta(X,Y)^\top W_\vartheta(X)^\top\} \\
&= E\big[\delta\dot{Q}_X(a_\vartheta)^\top E\{a_\vartheta(X,Y)a_\vartheta(X,Y)^\top|X\}^{-1}a_\vartheta(X,Y)a_\vartheta(X,Y)^\top \\
&\qquad \times E\{a_\vartheta(X,Y)a_\vartheta(X,Y)^\top|X\}^{-1}\dot{Q}_X(a_\vartheta)\big] \\
&= E\big[\dot{Q}_X(a_\vartheta)^\top E\{a_\vartheta(X,Y)a_\vartheta(X,Y)^\top|X\}^{-1}E\{\delta a_\vartheta(X,Y)a_\vartheta(X,Y)^\top|X\} \\
&\qquad \times E\{a_\vartheta(X,Y)a_\vartheta(X,Y)^\top|X\}^{-1}\dot{Q}_X(a_\vartheta)\big] \\
&= E\{W_\vartheta(X)E(\delta|X)\dot{Q}_X(a_\vartheta)\}.
\end{aligned}$$

Here we have used

$$E\{\delta a_\vartheta(X,Y)a_\vartheta(X,Y)^\top|X\} = E(\delta|X)E\{a_\vartheta(X,Y)a_\vartheta(X,Y)^\top|X\},$$

which follows from the MAR assumption. Handling the matrix on the right-hand side of (2.7) is notationally cumbersome. We therefore consider just a single entry of the matrix. Write $W_{\theta,i}$ for the i-th row of W_θ. Again using the MAR assumption, and the fact that $E\{a_\vartheta(X,Y)|X\} = 0$, the (i,j)-th entry computes as follows:

$$\begin{aligned}
&\frac{\partial}{\partial\theta_j}E\big\{\delta W_{\theta,i}(X)a_\theta(X,Y)\big\}\big|_{\theta=\vartheta} \\
&= \frac{\partial}{\partial\theta_j}E\big[E(\delta|X)W_{\theta,i}(X)E\{a_\theta(X,Y)|X\}\big]\big|_{\theta=\vartheta} \\
&= E\Big(E(\delta|X)\big[W_{\theta,i}(X)\frac{\partial}{\partial\theta_j}E\{a_\theta(X,Y)|X\} + E\{a_\theta(X,Y)|X\}\frac{\partial}{\partial\theta_j}W_{\theta,i}(X)\big]\Big)\Big|_{\theta=\vartheta} \\
&= E\big[E(\delta|X)W_{\theta,i}(X)\frac{\partial}{\partial\theta_j}E\{a_\theta(X,Y)|X\}\big]\big|_{\theta=\vartheta}.
\end{aligned}$$

Comparing this with the above calculation for I it is now apparent that the entries of I and $-\Gamma$ are the same. Hence, (2.7) is satisfied. It remains to prove condition (2) above. This follows from Theorem 3 in Chen et al. (2003 [5]) (discarding the nonparametric nuisance function h which is present in that theorem). □

3. Discussion and examples

3.1. Estimation of the weight matrix

As pointed out in the introduction, the estimating equation will in general be undetermined (and therefore of no use for applications) since the weights depend on unknown features of the distribution, for example on the conditional variance $\sigma^2(X)$ in nonlinear regression. This is not a problem: the unknown quantities can usually be estimated consistently by some simple nonparametric approach. This will not change the asymptotic variance of the resulting estimator. In particular, it will still be efficient. The estimator of $W_\theta(\cdot)$ does not need to converge to $W_\theta(\cdot)$ at a certain specific rate: simple (uniform) consistency is sufficient.

To show that the asymptotic variance does not change, one can use the results from Chen et al. (2003 [5]). They give high-level conditions under which a parameter estimator defined by the solution of a set of equations depending on a nonparametric estimator is consistent and asymptotically normal. These results extend Pakes and Pollard's (1989 [14]) article to the case of semiparametric estimators, and cover as a special case our model when the unknown weight matrix is replaced by a nonparametric estimator. Consider Theorem 2 in Chen et al. (2003 [5]), which states the asymptotic normality of $\hat{\vartheta}$. Most of the high-level conditions under which this result is valid are straightforward to verify. Two points, however, need closer attention:

(1) we need to calculate the asymptotic variance of $\hat{\vartheta}$, in order to confirm that it is not affected by using an estimator $\hat{W}_\theta$ for the weight matrix W_θ;
(2) we need to show that the required conditions on $\hat{W}_\theta$ are satisfied.

Let us address (1) first. According to Theorem 2 in Chen et al. (2003 [5]) the formula for the asymptotic variance of $\hat{\vartheta}$ depends on the matrices Γ_1 and V_1 given in conditions (2.2) and (2.6) of that paper. The matrix Γ_1 is constant and therefore not affected by using estimated weights. The matrix V_1 must be inspected more carefully: it is the asymptotic variance of an expression which involves the Gâteaux derivative of $M(\theta, W_\theta) := E\{\delta W_\theta(X) a_\theta(X, Y)\}$ in the direction $\hat{W}_\theta - W_\theta$ (with $\hat{W}_\theta(X)$ a consistent estimator of $W_\theta(X)$), evaluated at $\theta = \vartheta$. The Gâteaux derivative is defined by

$$
\begin{aligned}
&\Gamma_2(\theta, W_\theta)(\hat{W}_\theta - W_\theta) \\
&= \lim_{\tau \downarrow 0} \frac{1}{\tau} \big[M\{\theta, W_\theta + \tau(\hat{W}_\theta - W_\theta)\} - M(\theta, W_\theta) \big] \\
&= \lim_{\tau \downarrow 0} \frac{1}{\tau} \big\{ E\big(\delta[W_\theta(X) + \tau\{\hat{W}_\theta(X) - W_\theta(X)\}] a_\theta(X, Y)\big) - E\{\delta W_\theta(X) a_\theta(X, Y)\} \big\} \\
&= E[\delta\{\hat{W}_\theta(X) - W_\theta(X)\} a_\theta(X, Y)].
\end{aligned}
$$

Note that the expected value is calculated in accordance with the definition of the vector $M(\theta, W_\theta)$, namely with respect to (δ, X, Y), i.e. the stochastic nature of $\hat{W}_\theta$ is not taken into account. Writing the last expectation in the above display as an iterated expectation (conditional on X) yields $\Gamma_2(\vartheta, W_\vartheta)(\hat{W}_\vartheta - W_\vartheta) = 0$.

In other words, the contribution to the asymptotic variance which comes from using estimated weights is zero. The matrix V_1 is the same as in the case with known weights.

For (2), note that the main requirement on $\hat{W}_\theta$ is condition (2.4) in Chen et al. (2003 [5]), which requires that $\sup_{\theta\in\Theta}\sup_x|\hat{W}_\theta(x)-W_\theta(x)|=o_p(n^{-1/4})$. However, a closer look at the proof of Theorem 2 in that paper reveals that the rate $o_p(n^{-1/4})$ can be weakened to $o_p(1)$ if $M(\theta,W_\theta)$ depends on W_θ in a linear way (or, equivalently, if $\Gamma_2(\theta,W_\theta)(\hat{W}_\theta-W_\theta)=M(\theta,\hat{W}_\theta)-M(\theta,W_\theta)$), which is the case here. Hence, all we need is an estimator $\hat{W}_\theta(x)$ that is uniformly consistent (in θ and x).

This sketches the main steps of the proof that the estimation of the weight matrix does not impair the efficiency property of $\hat{\vartheta}$.

3.2. Conditional versus unconditional constraints

Our focus here is on inference for parameters defined via conditional equations. A related topic is inference for parameters defined via unconditional constraints of the form $E\{a_\vartheta(X,Y)\}=0$, see e.g. Zhou et al. (2008 [23]) and Wang and Chen (2009 [20]) for references on this type of models. Let us explain the relationship between the two classes of models. The conditional model $E\{a_\vartheta(X,Y)|X\}=0$ a.s. is equivalent to $E\{W(X)a_\vartheta(X,Y)\}=0$ for all possible functions $W(\cdot)$. Indeed, the former equation clearly implies the latter one. On the other hand, the latter set of equations yields that $E[E\{a_\vartheta(X,Y)|X\}^2]=0$ by choosing $W(X)=E\{a_\vartheta(X,Y)|X\}$. This implies that $E\{a_\vartheta(X,Y)|X\}=0$ a.s.. This means that the conditional constraint is equivalent to an *infinite collection* of unconditional constraints, one of which (namely the one corresponding to $W=W_\theta$ given in (2.3)) is efficient. So the approach with the conditional constraint makes it possible to select the weight matrix that leads to an efficient estimator. On the other hand, an unconditional constraint corresponds to *one single* equation, or equivalently one single weight matrix.

3.3. Illustration: Linear and nonlinear regression

The estimating equation for the nonlinear regression model (which includes linear regression as a special case) is given in the introduction (1.4). Let us check that it is indeed a special case of the general estimating equation (2.4), i.e. of $\sum_{i=1}^n\delta_i\ell_\theta(X_i,Y_i)=-\sum_{i=1}^n\delta_iW_\theta(X_i)a_\theta(X_i,Y_i)=0$. Here the vector a_ϑ is one-dimensional, $a_\theta(X,Y)=Y-r_\theta(X)$, which yields that the matrix $E\{a_\theta(X,Y)a_\theta(X,Y)^\top|X\}$ is one-dimensional as well, $E\{a_\theta(X,Y)^2|X\}=\sigma^2(X)$, where $\sigma^2(X)$ is the conditional variance of Y given X. Assuming that r_θ is differentiable in θ we also have that $E\{\dot{a}_\theta(X,Y)|X\}=-\dot{r}_\theta(X)$. This yields

$$\ell_\theta(x,y)=\dot{r}_\theta(x)^\top\sigma^{-2}(x)\{y-r_\theta(x)\}$$

as postulated. A simple consistent nonparametric estimator of $\sigma^2(x)$ is

$$\hat{\sigma}^2(x) = \sum_{i=1}^{n} \frac{\delta_i k_b(x - X_i)}{\sum_{i=1}^{n} \delta_i k_b(x - X_i)} \{Y_i - r_{\hat{\vartheta}_0}(X_i)\}^2 \tag{3.1}$$

(which can be regarded as a ratio of Nadaraya-Watson estimators), where $k_b(x)$ is a d-dimensional kernel with bandwidth b, $k_b(x) = k(x_1/b, \ldots, x_d/b)/b^d$. Here d is the dimension of X and $\hat{\vartheta}_0$ is some consistent estimator for ϑ, e.g. the ordinary least squares estimator (OLS) which uses weights $W_\theta(x) = \dot{r}_\theta(x)^\top$. In the general case a consistent estimator of the *optimal* weight matrix W_θ may similarly involve a preliminary consistent estimator $\hat{\vartheta}_0$ of ϑ. Such an estimator can be obtained as a solution of equation (2.2), i.e. of $\sum_{i=1}^{n} \delta_i W_\theta(X_i) a_\theta(X_i, Y_i) = 0$, now with an *arbitrary* (feasible) $p \times k$ weight matrix W_θ (which does not need to depend on θ) such that the system of equations has a unique solution $\hat{\vartheta}_0$ (see the discussion in Section 2).

As an illustration of the method we performed a small simulation study using R and compared three different approaches: the efficient estimator, the OLS which solves (2.1), and a weighted least squares estimator that uses the propensity score, $W_\theta(X) = W(X) = \pi(X)^{-1} = E(\delta|X)^{-1}$. The latter choice of weights is suitable for the larger model defined by the unconditional constraint $E\{a_\vartheta(X, Y)\} = 0$, since the corresponding estimating equation is unbiased in that model,

$$\begin{aligned} E\{\delta\pi(X)^{-1} a_\vartheta(X, Y)\} &= E\big[\pi(X)^{-1} E(\delta|X) E\{a_\vartheta(X, Y)|X\}\big] \\ &= E\{a_\vartheta(X, Y)\} = 0. \end{aligned}$$

For the simulations we chose an increasing propensity score $\pi(x) = 1/(1 + e^{-x})$. The covariate X is generated from a uniform distribution on $(-1, 1)$, and the error variable is of the form $\varepsilon = \sigma(X)Z$, where Z is standard normal and independent of X. We studied a linear regression function, $r_\vartheta(X) = \vartheta X$, and a nonlinear regression function, $r_\vartheta(X) = \cos(\vartheta X)$. In both cases $\vartheta = 2$. The conditional variance $\sigma^2(x)$ is linear or parabolic, and estimated by $\hat{\sigma}^2(x)$ from equation (3.1), with $\hat{\vartheta}_0$ the OLS estimator. We studied five bandwidths b between 0.1 and 0.5, and an automatically selected bandwidth $b = b_{cv}$ using the cross-validation method for fitting a smooth curve into the completely observed 'data' pairs $(X, \widetilde{Y})$, where $\widetilde{Y} = \{Y - r_{\hat{\vartheta}_0}(X)\}^2$. Table 1 lists the simulated mean squared errors based on 5,000 repetitions for the case of a linear regression function. The results for the cosine function are given in Table 2. The propensity score $\pi(X)$ increases from 0.27 to 0.73 on $(-1, 1)$ so that around 50% of the responses are missing. Hence, if $n = 50$, we are essentially only working with about 25 data points and the R routine "nls" (nonlinear least squares), which we used for the cosine function, does not always converge (the simulations 'crashed'). For this reason Table 2 only includes the results for $n = 100$ and $n = 200$. In the linear case (Table 1) the estimator can be calculated with an explicit formula and the simulations ran without any problems, allowing us to include results for $n = 50$ as well.

TABLE 1
Simulated MSEs of estimators of ϑ with $r_\vartheta(X) = \vartheta X$ ($\vartheta = 2$)

$\sigma^2(X)$	n	OLS	PS	$\sigma^2(x)$	0.1	0.2	0.3	0.4	0.5	b_{cv}
(a)	50	0.058	0.103	0.034	0.044	0.043	0.042	0.041	0.042	0.044
	100	0.028	0.052	0.016	0.019	0.018	0.018	0.019	0.019	0.020
	200	0.014	0.026	0.008	0.008	0.008	0.008	0.009	0.009	0.009
(b)	50	0.080	0.149	0.037	0.049	0.046	0.046	0.047	0.049	0.050
	100	0.039	0.075	0.018	0.021	0.020	0.021	0.022	0.023	0.023
	200	0.019	0.038	0.009	0.009	0.010	0.010	0.010	0.011	0.010

The first two results columns give the mean squared errors (MSE) for the OLS and the propensity score weighted estimator (PS). For simplicity, PS uses the true $\pi(X)$. The third column shows the results for the efficient estimator that uses the true conditional variance $\sigma^2(x)$ with (a) $\sigma^2(x) = 0.6 - 0.5x$ in the upper panel, and (b) $\sigma^2(x) = (x - 0.4)^2 + 0.1$ in the lower panel. The six columns on the right-hand side refer to the efficient estimator based on the kernel estimator (3.1) for $\sigma^2(x)$, for five different fixed bandwidths b, and for $b = b_{cv}$ obtained by cross-validation.

TABLE 2
Simulated MSEs of estimators of ϑ with $r_\vartheta(X) = \cos(\vartheta X)$ ($\vartheta = 2$)

$\sigma^2(X)$	n	OLS	PS	$\sigma^2(x)$	0.1	0.2	0.3	0.4	0.5	b_{cv}
(a)	100	0.021	0.035	0.011	0.016	0.013	0.013	0.14	0.014	0.015
	200	0.010	0.017	0.005	0.006	0.006	0.006	0.006	0.007	0.007
(b)	100	0.031	0.055	0.013	0.020	0.016	0.016	0.017	0.018	0.018
	200	0.015	0.027	0.007	0.008	0.007	0.008	0.008	0.009	0.008

The entries are simulated mean squared errors of various estimators of ϑ as in Table 1, now with a nonlinear regression function, $r_\vartheta(X) = \cos(\vartheta X)$.

We observe that the efficient estimator that uses the variance estimator $\hat{\sigma}^2(x)$ always performs better than both the OLS and the propensity score weighted estimator (PS), for all choices of b, and also for the automatic bandwidth b_{cv} selected by cross-validation. Note that our estimator $\hat{\sigma}^2(x)$ uses a normal kernel k, without adjusting for boundary bias, which is probably one reason why the estimator that uses the true variance function is better when n is small, e.g. $n = 50$ in Table 1.

A reasonable next step, with view towards small sample performance, would be to develop a better estimator of the conditional variance. If the variance function is constant, ordinary least squares and the efficient estimator are asymptotically equivalent and we recommend the OLS estimator since it is easier to use. The same applies if the variance function does not show much variation and if the sample size is small so that the estimated variance function is nearly constant.

3.4. Further examples

Quasi-likelihood model Now consider the quasi-likelihood model where we assume parametric models for both the regression function and the conditional variance function. Here it is also straightforward to calculate $\ell_\theta(x, y)$: assuming

that both r_θ and v_θ are differentiable in θ we obtain

$$\ell_\theta(x,y) = \left[\dot{r}_\theta(x)^\top \dot{v}_\theta(x)^\top\right] \begin{bmatrix} v_\theta(x) & \mu_3(x) \\ \mu_3(x) & \mu_4(x) - v_\theta^2(x) \end{bmatrix}^{-1} \begin{bmatrix} y - r_\theta(x) \\ \{y - r_\theta(x)\}^2 - v_\theta(x) \end{bmatrix}$$

where $\mu_k(x) = E(Y^k|X = x)$, $k \in \mathbb{N}$, is the k-conditional moment of the distribution of Y given $X = x$. A simple consistent estimator for $\mu_k(x)$ is the estimator

$$\hat{\mu}_k(x) = \sum_{i=1}^n \frac{\delta_i k_b(x - X_i)}{\sum_{i=1}^n \delta_i k_b(x - X_i)} Y_i^k,$$

similar to that in the previous example.

Multi-response model In the two examples above the response variable was assumed to be univariate. Our method also applies if the responses are multivariate, i.e. if we assume a multi-response model. Again it would be straightforward to specify the estimating equation.

Quantile regression The situation is different if $a_\theta(x,y)$ involves indicators and cannot be differentiated with respect to θ. An important class of applications are quantile regression models. Suppose that the conditional p-th quantile of Y given X is specified by a parametric model $r_\vartheta(X)$. This can be expressed as a conditional constraint, namely as

$$E\{a_\vartheta(X,Y)|X\} = 0, \quad \text{with } a_\theta(X,Y) = p - \mathbf{1}\{Y - r_\theta(X) < 0\}, \quad \theta \in \Theta.$$

A simple calculation shows that $E\{a_\theta(X,Y)^2|X\} = p^2 + (1-2p)F_{Y|X}\{r_\theta(X)\}$ for any $\theta \in \Theta$, where $F_{Y|X}\{r_\theta(X)\} = P\{Y - r_\theta(X) < 0|X\}$. Thus the weights of the estimating equation reduce to

$$\begin{aligned} W_\theta(X) &= \left[p^2 + (1-2p)F_{Y|X}\{r_\theta(X)\}\right]^{-1} \frac{\partial}{\partial\theta} E\{a_\theta(X,Y)|X\} \\ &= -\frac{\frac{\partial}{\partial\theta} F_{Y|X}\{r_\theta(X)\}}{\left[p^2 + (1-2p)F_{Y|X}\{r_\theta(X)\}\right]}. \end{aligned}$$

The conditional probability $F_{Y|X}\{r_\theta(X)\}$ must be estimated with a smooth estimator to ensure that the partial derivatives can be calculated. One option is to use a kernel smoother of the form

$$\hat{F}_{Y|X=x}(y) = \sum_{i=1}^n \frac{\delta_i k_b(x - X_i)}{\sum_{i=1}^n \delta_i k_b(x - X_i)} K\Big(\frac{y - Y_i}{h}\Big),$$

where, as before, k is a kernel function and b and h are appropriate smoothing parameters, and where K is a smooth distribution function, e.g. the cumulative integral of a suitable kernel density function.

4. Efficiency

In order to derive the canonical gradient of ϑ (which characterizes the efficient influence function) one can build on results from Müller (2007 [8]) on estimating ϑ when all data are observed. We will also rely on results by Müller et al. (2006 [10]) on efficient estimation of expectations $Eh(X,Y)$ in regression models (not covering our model) with responses missing at random, that is, with observations $(X,\delta Y,\delta)$ as here. The characterization of an efficient estimator $\hat{\vartheta}$ of ϑ is given at the end of this section.

We begin with the characterization of the influence function of an arbitrary differentiable functional κ of the joint distribution of (X,Y) which is derived in that article. The joint distribution $P(dx,dy,dz)$ of the observations $(X,\delta Y,\delta)$ can be written as

$$P(dx,dy,dz) = M(dx)B_{\pi(x)}(dz)\{zQ(x,dy) + (1-z)D_0(dy)\}.$$

Here $M(dx)$ is the marginal distribution of X, $Q(x,dy)$ is the conditional distribution of Y given $X=x$, and $\pi(x) = P(\delta = 1|X=x)$. Further, for any $0 \le p \le 1$, $B_p = pD_1 + (1-p)D_0$ is the Bernoulli distribution with parameter p, where D_t denotes the Dirac measure at t. As shown in Müller et al. (2006 [10]), a gradient γ for κ is characterized by

$$\lim_{n\to\infty} n^{1/2}\{\kappa(M_{nu},Q_{nv}) - \kappa(M,Q)\} = E[\gamma(X,\delta Y,\delta)\{u(X) + \delta v(X,Y)\}]$$

for all $u \in U$ and $v \in V$, where

$$U = L_{2,0}(M) = \Big\{u \in L_2(M) : M(u) = \int u\,dM = 0\Big\},$$

and

$$V = \Big\{v \in L_2(M\cdot Q) : Q_x(v) = \int v(x,y)Q(x,dy) = 0 \text{ for all } x\Big\}.$$

Here M_{nu} and Q_{nv} are Hellinger differentiable perturbations of M and Q,

$$\begin{aligned} M_{nu}(dx) &= M(dx)\{1 + n^{-1/2}u(x)\} + o(n^{-1/2}),\\ Q_{nv}(x,dy) &= Q(x,dy)\{1 + n^{-1/2}v(x,y)\} + o(n^{-1/2}). \end{aligned}$$

The perturbed distributions M_{nu} and Q_{nv} must both be probability distributions, i.e. integrate to one, which explains the form of U and V. Write T for the tangent space relevant for estimating κ (i.e. for functionals of M and Q),

$$T = \big\{u(X) : u \in U\big\} \oplus \big\{\delta v(X,Y) : v \in V\big\},$$

where the orthogonality follows from the missing at random assumption. It contains the canonical gradient, which is defined as a gradient that is also an element of the tangent space, i.e. it is of the form $\gamma_*(X,\delta Y,\delta) = u_*(X) +$

$\delta v_*(X,Y)$ with the terms of the sum being projections onto the tangent space. As a gradient of κ, γ_* must satisfy the above characterization which now becomes

$$\begin{aligned}&\lim_{n\to\infty} n^{1/2}\{\kappa(M_{nu},Q_{nv})-\kappa(M,Q)\}\\&= E\{u_*(X)u(X)\}+E\{\delta v_*(X,Y)v(X,Y)\}.\end{aligned}\tag{4.1}$$

For a full specification of the tangent space see Müller et al. (2006 [10]). That larger tangent space has to be considered when the goal is to estimate functionals of the full joint distribution, i.e. of functionals that also involve the conditional distribution $\pi(x)$ of the indicator variable δ given x.

After these general considerations we now also take the structure of our model (1.1) into account, which is defined by a parametric constraint,

$$0=E\{a_\vartheta(X,Y)|X=x\}=Q_x(a_\vartheta)=\int a_\vartheta(x,y)Q(x,dy).$$

The perturbed distribution must satisfy a perturbed constraint, $Q_{xnv}(a_{\vartheta_{nt}})=0$ for some ϑ_{nt} close to ϑ, say $\vartheta_{nt}=\vartheta+n^{-1/2}t$ with t in $\mathbb{R}^p$. Using $Q_x(a_\vartheta)=0$ and $Q_x(v)=0$ we obtain

$$\begin{aligned}0&=Q_{xnv}(a_{\vartheta_{nt}})\\&=Q_x\{(1+n^{-1/2}v)a_{\vartheta_{nt}}\}+o(n^{-1/2})\\&=Q_x(a_{\vartheta_{nt}})+n^{-1/2}Q_x(va_{\vartheta_{nt}})+o(n^{-1/2})\\&=Q_x(a_\vartheta)+n^{-1/2}t\frac{\partial}{\partial\theta}Q_x(a_\theta)\big|_{\theta=\vartheta}+n^{-1/2}Q_x(va_\vartheta)+o(n^{-1/2})\\&=n^{-1/2}Q_x\{va_\vartheta+t\dot{Q}_x(a_\vartheta)\}+o(n^{-1/2}),\end{aligned}$$

with $\dot{Q}_x(a_\vartheta)=\frac{\partial}{\partial\theta}Q_x(a_\theta)\big|_{\theta=\vartheta}$. This leads to a constraint $Q_x(va_\vartheta)=-\dot{Q}_x(a_\vartheta)t$ on v in V, which can be written in the form $Q_x(va_\vartheta)=-Q_x(\dot{a}_\vartheta)t$ if a_θ is differentiable in θ. For fixed $t\in\mathbb{R}^p$ we write H_t for the solution space of this equation,

$$H_t=\{v\in V: Q_x(va_\vartheta)=-\dot{Q}_x(a_\vartheta)t\},$$

and H_* for the union of all affine spaces H_t, $t\in\mathbb{R}^p$. In order to determine v_* we find it convenient to go further and decompose H_* into the space H_0 of solutions of the homogeneous equation, $H_0=\{v\in V: Q_x(va_\vartheta)=0\}$, and into the solution space of the inhomogeneous equation given above. This space can be written as a linear span, analogously to Müller (2007 [8]). The idea is to solve the equation for the standard basis vectors $t=e_j$, $j=1,\dots,p$. Call the solutions ℓ_j. Then the solution space of the inhomogeneous equation is the linear span $[\ell]$ of the solutions $\ell_1,\dots,\ell_p$, where $\ell=(\ell_1,\dots,\ell_p)^\top$ has the form

$$\ell(x,y) \;=\; -\dot{Q}_x(a_\vartheta^\top)Q_x(a_\vartheta a_\vartheta^\top)^{-1}a_\vartheta(x,y).$$

Simple calculations show that ℓ indeed satisfies $Q_x(a_\vartheta\ell^\top)=-\dot{Q}_x(a_\vartheta)$ and that ℓ is orthogonal to H_0, i.e. $H_*=H_0\oplus[\ell]$. The tangent space of the constrained model is now specified,

$$T=\big\{u(X):u\in U\big\}\oplus\big\{\delta v(X,Y):v\in H_0\oplus[\ell]\big\}.$$

From now on we focus on estimating ϑ and write it as a functional of P by setting $\kappa(P) = \vartheta$ if $Q_x(a_\vartheta) = 0$. The left-hand side of characterization (4.1) of the canonical gradient now involves $t \in \mathbb{R}^p$ and simplifies to $n^{1/2}(\vartheta_{nt} - \vartheta) = t$. The canonical gradient $\gamma_*(X, \delta Y, \delta) = u_*(X) + \delta v_*(X, Y)$ is therefore determined by

$$E\{u_*(X)u(X)\} + E\{\delta v_*(X, Y)v(X, Y)\} = t \quad \text{for all } t \in \mathbb{R}^p,\ u \in U,\ v \in H_0 \oplus [\ell].$$

Setting $v = 0$ and $t = 0$ we see that u_* must be zero. Further we know that $v_* \in H_0 \oplus [\ell]$ where $[\ell]$ comes from ϑ being unknown. We can therefore assume that v_* is of the form $J\ell$, where J is a $p \times p$ matrix to be determined and where ℓ functions as the score function. This yields

$$E\{\delta J\ell(X, Y)v(X, Y)\} = t \quad \text{for all } t \in \mathbb{R}^p,\ v \in H_0 \oplus [\ell],$$

where

$$\begin{aligned} E\{\delta J\ell(X, Y)v(X, Y)\} &= -JE\{\delta \dot{Q}_X(a_\vartheta^\top) Q_X(a_\vartheta a_\vartheta^\top)^{-1} a_\vartheta(X, Y)v(X, Y)\} \\ &= -JE[\dot{Q}_X(a_\vartheta^\top) Q_X(a_\vartheta a_\vartheta^\top)^{-1} E\{\delta a_\vartheta(X, Y)v(X, Y)|X\}] \end{aligned}$$

with

$$\begin{aligned} &E\{\delta a_\vartheta(X, Y)v(X, Y)|X = x\} = E\{\delta a_\vartheta(x, Y)v(x, Y)|X = x\} \\ &= E\{\delta|X = x\}E\{a_\vartheta(x, Y)v(x, Y)|X = x\} = Q_x(\delta)Q_x(a_\vartheta v) = -Q_x(\delta)\dot{Q}_x(a_\vartheta)t. \end{aligned}$$

Here we have used the MAR assumption and the conditional constraint $Q_x(va_\vartheta) = -\dot{Q}_x(a_\vartheta)t$ on v in V. Inserting this in the above gives

$$E\{\delta J\ell(X, Y)v(X, Y)\} \quad = \quad JE[\delta \dot{Q}_X(a_\vartheta^\top) Q_X(a_\vartheta a_\vartheta^\top)^{-1} \dot{Q}_X(a_\vartheta)t].$$

This equals t if $J = I^{-1}$ with

$$I = E\{\delta \dot{Q}_X(a_\vartheta^\top) Q_X(a_\vartheta a_\vartheta^\top)^{-1} \dot{Q}_X(a_\vartheta)\} = E\{\delta\ell(X, Y)\ell(X, Y)^\top\}.$$

Our canonical gradient for estimating ϑ is determined: it is $\gamma_*(X, \delta Y, \delta) = \delta v_*(X, Y) = \delta I^{-1}\ell(X, Y)$.

Characterization of the efficient estimator By the characterization of efficient estimators, an estimator $\hat{\vartheta}$ is efficient for ϑ if it is asymptotically linear with influence function equal to the canonical gradient. The efficient influence function for estimating ϑ is $I^{-1}\delta\ell(X, Y)$ in our model (1.1). An estimator $\hat{\vartheta}$ is therefore efficient for ϑ if it satisfies

$$n^{1/2}(\hat{\vartheta} - \vartheta) = I^{-1} \sum_{i=1}^{n} \delta_i \ell_\vartheta(X_i, Y_i) + o_p(n^{-1/2})$$

with $\ell_\vartheta(x, y) = -\dot{Q}_x(a_\vartheta^\top) Q_x(a_\vartheta a_\vartheta^\top)^{-1} a_\vartheta(x, y)$ and $I = E\{\delta\ell_\vartheta(X, Y)\ell_\vartheta(X, Y)^\top\}$.

5. Concluding remarks and future research

We have derived asymptotically efficient estimators for the parameter vector ϑ for the large class of regression models that can be specified by a conditional constraint of the form $E\{a_\vartheta(X,Y)|X\} = 0$. We focus on the situation when responses are missing at random, but this also covers the case when no data are missing, namely when $\pi(X) = P(\delta = 1|X) = 1$ and all indicators are equal to one. The proposed method is not only efficient, it is also simple: we estimate ϑ by solving a weighted estimating equation which only incorporates completely observed cases (X,Y), and discard those cases that contain missing values. Although this requires estimating the weights, we only need consistency (without a rate). It is certainly remarkable that an efficient estimator may be based only on the observations for which both the regressors and responses are available. However, the final efficient estimator does not necessarily have to be of this type: a consistent estimator of the weight matrix *can* be obtained by discarding the data for which the response is missing, but other consistent estimators of this weight matrix are allowed as well. For instance, one could use imputation of the missing responses if one is in favor of the imputation principle, although we do not recommend doing so because the estimators can become quite involved, as explained in the introduction.

There are several open questions for future research. For example, our class of models does not include regression models where the regression function itself contains a nonparametric part, such as partially linear models which are defined by the conditional constraint $E\{Y-\vartheta^\top X_1+\eta(X_2)|X_1,X_2\}=0$. This constraint additionally involves the infinite-dimensional nuisance parameter η.

It would also be interesting to see whether the methodology developed in this paper can be extended to other missingness schemes. Clearly, the results apply to the MCAR (missing completely at random) mechanism, i.e. when $\pi(\cdot) \equiv \pi$ is constant. On the other hand, when the missingness is not at random (NMAR), the present methodology cannot be applied: the equality $E\{\delta a_\vartheta(X,Y)|X\} = E(\delta|X)E\{a_\vartheta(X,Y)|X\}$, which relies on the MAR assumption, is crucial for the development of an efficient (optimally weighted) estimator since it guarantees unbiasedness of the estimating equation (2.2). Of interest is also the situation when both covariates and responses are missing, or when only covariates are missing with the missingness explained by the response variable.

So far we have only studied estimation of the parameter vector, but it would also be interesting to derive estimators for expectations $Eh(X,Y)$, with the mean response EY as an important special case. Although the mean response has been well studied, it is not yet clear how to estimate expectations in our model efficiently. To our knowledge, this has not even been considered in the nonlinear regression model which is specified by $E\{a_\vartheta(X,Y)|X\} = E\{Y - r_\vartheta(X)|X\} = 0$. In this model we expect that, similar to the model with *independent* covariates and errors (Müller, 2009 [9]), the estimator $n^{-1}\sum_{i=1}^n r_{\hat\vartheta}(X_i)$, now with our efficient estimator from equation (2.5) plugged in, will be efficient for EY. This is in agreement with the linear regression model, $r_\vartheta(X) = \vartheta^\top X$. Here $n^{-1}\sum_{i=1}^n r_{\hat\vartheta}(X_i) = n^{-1}\sum_{i=1}^n \hat\vartheta^\top X_i = \hat\vartheta^\top \bar X$, which is a smooth function

of two efficient estimators and therefore efficient. Since efficient estimators are asymptotically normally distributed with the asymptotic variance specified by the length of the canonical gradient, the construction of (approximative) normal confidence intervals for moments of the response variable, and for more general expectations, would be straightforward.

In applications it is often necessary to work with more complex models. We would expect interesting and useful results in the field of generalized linear models, for certain change point models, for models with censored/truncated data (in addition to missing data), and for models used in case-control studies in the field of biostatistics. For each of these models one would need to specify the function $a_\theta(X, Y)$, from which the formula of the weight matrix and its estimator can be obtained.

Acknowledgment

U.U. Müller's research was supported by National Science Foundation grant DMS-0907014. I. Van Keilegom acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650. The authors would like to thank two referees for their helpful comments, which they believe have improved the paper.

References

[1] Bickel P.J. (1982). On adaptive estimation. *Ann. Statist.*, 10, 647-671. MR0663424

[2] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer. MR1623559

[3] Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics*, 34, 305-334. MR0888070

[4] Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, 60, 567-596. MR1162999

[5] Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71, 1591-1608. MR2000259

[6] Forrester, J., Hooper, W., Peng, H. and Schick, A. (2003). On the construction of efficient estimators in semiparametric models. *Statist. Decisions*, 21, 109-138. MR2000666

[7] Koul, H.L. and Susarla, V. (1983). Adaptive estimation in linear regression. *Statist. Decisions*, 1, 379-400. MR0736109

[8] Müller, U.U. (2007). Weighted least squares estimators in possibly misspecified nonlinear regression. *Metrika*, 66, 39-59. MR2306376

[9] Müller, U.U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Statist.*, 37, 2245-2277. MR2543691

[10] Müller, U.U., Schick, A. and Wefelmeyer, W. (2006). Imputing responses that are not missing. In: *Probability, Statistics and Modelling in Public Health* (M. Nikulin, D. Commenges and C. Huber, eds.), 350-363, Springer. MR2230741

[11] Müller, U.U. and Wefelmeyer, W. (2002). Autoregression, estimating functions, and optimality criteria. In: *Advances in Statistics, Combinatorics and Related Areas* (C. Gulati, Y.-X. Lin, J. Rayner and S. Mishra, eds.), 180-195, World Scientific Publishing, Singapore. MR2063849

[12] Newey, W.K. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics*, 5, 99-135.

[13] Newey, W.K. (1993). Efficient estimation of models with conditional moment restrictions. In: *Handbook of Statistics 11: Econometrics* (G. S. Maddala, C. R. Rao and H. D. Vinod, eds.), 419-454. Elsevier, Amsterdam. MR1247253

[14] Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57, 1027-1057. MR1014540

[15] Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, 89, 846-866. MR1294730

[16] Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.*, 90, 106-121. MR1325118

[17] Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference*, 16, 89-105. MR0887419

[18] Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.*, 21, 1486-1521. Correction and addendum: 23 (1995), 1862-1863. MR1241276

[19] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. Springer. MR2233926

[20] Wang, D. and Chen, S.X. (2009). Empirical likelihood for estimating equations with missing values. *Ann. Statist.*, 37, 490-517. MR2488360

[21] Wang, Q. and Sun, Z. (2007). Estimation in partially linear models with missing response at random. *J. Multivariate Anal.*, 98, 1470-1493. MR2364130

[22] Wang, Y., Shen, J., He, S. and Wang, Q. (2010). Estimation of single index model with missing response at random. *J. Statist. Plann. Inference*, 140, 1671-1690. MR2606708

[23] Zhou, Y., Wan, A.T.K. and Wang, X. (2008). Estimating equations inference with missing data. *J. Amer. Statist. Assoc.*, 103, 1187-1199. MR2462892