Bernoulli **23**(4A), 2017, 2693–2719
DOI: 10.3150/16-BEJ824

Efficiency transfer for regression models with responses missing at random

URSULA U. MÜLLER[1] and ANTON SCHICK[2]

[1]*Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.*
E-mail: uschi@stat.tamu.edu
[2]*Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, USA.*
E-mail: anton@math.binghamton.edu

We consider independent observations on a random pair (X, Y), where the response Y is allowed to be missing at random but the covariate vector X is always observed. We demonstrate that characteristics of the conditional distribution of Y given X can be estimated efficiently using complete case analysis, that is, one can simply omit incomplete cases and work with an appropriate efficient estimator which remains efficient. This means in particular that we do not have to use imputation or work with inverse probability weights. Those approaches will never be better (asymptotically) than the above complete case method.

This *efficiency transfer* is a general result and holds true for all regression models for which the distribution of Y given X and the marginal distribution of X do not share common parameters. We apply it to the general homoscedastic semiparametric regression model. This includes models where the conditional expectation is modeled by a complex semiparametric regression function, as well as all basic models such as linear regression and nonparametric regression. We discuss estimation of various functionals of the conditional distribution, for example, of regression parameters and of the error distribution.

Keywords: complete case analysis; efficient estimation; efficient influence function; linear and nonlinear regression; nonparametric regression; partially linear regression; random coefficient model; tangent space; transfer principle

1. Introduction

Missing values present a challenge in many applications. In practice, popular methods of handling missing data are single value imputation, multiple imputation, maximum likelihood estimation and complete case analysis (or "listwise deletion"), which simply discards incomplete cases. The last method carries the risks of bias and of losing valuable information, and is usually not recommended. There are, however, many applications where complete case analysis is indeed appropriate. A well-known example where complete case analysis is the accepted approach is maximum likelihood estimation of a parameter when the distribution of a sample $Z_1, \dots, Z_n$ is modeled by a parametric density f_θ, and when observations are *missing at random* (MAR) in the sense of [28]. This means that the missingness mechanism depends only on the subvector Z_{obs} of the data that contains the complete observations. The likelihood then factorizes in such a way that only the "observed-data" likelihood (based on Z_{obs}) depends on θ; see, for example, the recent book by Kim and Shao ([9]; Chapter 2). For an overview of common methods of handling missing data, see the book by [14].

1350-7265

In this article, we consider independent copies $(X_1, \delta_1 Y_1, \delta_1), \ldots, (X_n, \delta_n Y_n, \delta_n)$ of a base observation $(X, \delta Y, \delta)$, where δ is an indicator which equals 1 if the response Y is observed, and 0 otherwise. If δY is 0, then either Y is an observed numerical zero (and $\delta = 1$), or Y is missing (and $\delta = 0$), that is, the indicator helps us to distinguish a missing response from an observation with value 0. We assume that the covariate vector X is always observed and that Y is (strongly ignorable) missing at random in the sense of [27], who also consider a regression setting. This means that the probability that Y is observed depends only on the covariate vector X, that is,

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = \pi(X).$$

It implies that Y and δ are conditionally independent given X. An important special case which is also covered in this paper is the model with responses missing *completely* at random, in which π is a constant. The MAR assumption is reasonable in many applications. It has the advantage that the missingness depends only on the observed data – in our case on the covariates – and can therefore be estimated from the data.

We will show the general result that *arbitrary* (differentiable) functionals of the conditional distribution of Y given X (without assuming a specific regression structure) can be estimated efficiently by a complete case version of an efficient estimator, that is, by a statistic that uses only the observations that are completely observed. This means that we can ignore the incomplete cases and work with a familiar efficient estimator of choice without losing consistency and optimality. We call this property *efficiency transfer*.

Our article generalizes Koul, Müller and Schick's finding that a complete case version of an efficient estimator of the finite-dimensional parameter in a "full" partially linear model (where no data are missing) remains efficient in the corresponding "MAR model", that is, in the model with responses missing at random [12]. Koul *et al.* prove efficiency of the parameter estimator by direct means. That proof is now obsolete: since the regression parameter is a functional of the conditional distribution, it is simply a consequence of the general result to be presented in this article.

The efficiency transfer applies to all regression models. We focus on the homoscedastic semiparametric regression model with unknown error distribution to illustrate its usefulness. This model assumes that the conditional expectation of Y given X depends on a finite-dimensional parameter ϑ and an infinite-dimensional parameter ξ,

$$Y = r(X, \vartheta, \xi) + \varepsilon, \tag{1.1}$$

and that the centered error variable ε is independent of the covariate X. Model (1.1) includes the basic regression models, that is, linear, nonlinear and nonparametric regression, and also more elaborate models, for example, the partially linear model, the single index model, and models with random coefficients. Here the efficiency transfer applies to the estimation of functionals of the regression parameters ϑ, ξ and the error distribution F. Our results are valid for any model of the covariate distribution G, as long as the model does not link the covariate distribution to the regression parameters or the error distribution. It should be noted that the efficiency transfer does not apply to functionals of the *joint* distribution that also involve the marginal distribution of X. Consider, for example, estimation of $Eh(X, Y)$, where h is a given function. The complete case version of the empirical estimator is $\sum_{j=1}^n \delta_j h(X_j, Y_j)/N$ with $N = \sum_{j=1}^n \delta_j$ the number of complete cases. It cannot be recommended because it estimates the conditional expectation

$E[h(X,Y)|\delta=1]$. In those cases other methods should be used. An established approach to correct the bias is to add estimated inverse probability weights $\hat{\pi}(X_j)^{-1}$; see, for example, [26] and other papers by these authors and their collaborators. Müller [15] provides an efficient imputation estimator for $Eh(X,Y)$ in nonlinear regression, which does not require an estimator of the function π.

Most of the literature on regression with MAR responses studies estimation of the mean response $E[Y]$; for example, [3]. Articles that study functionals of the conditional distribution typically treat the regression function. Complete case analysis has only recently received increased attention. Efromovich [5] proposes a nonparametric complete case estimator of a nonparametric regression function and demonstrates an asymptotic minimax property. Müller [15] shows efficiency of a complete case estimator of the regression parameter in nonlinear regression with independent errors and covariates. Müller and Van Keilegom [23] do not require independent error variables: they consider regression models defined by constraints on the conditional distribution, and prove that complete case estimators are efficient in this large class of models. This is related to [26], who also consider a conditionally constrained model. They estimate the regression parameters by solving an inverse probability weighted estimating equation. This requires a parametric model for $\pi(X)$, which makes the setting conceptually different. There are also articles on estimating parameters where the efficiency transfer does not apply. Wang and Chen [32], for example, study estimation of parameters that are defined by *unconditional* constraints, which form a model for the joint distribution. They suggest an empirical likelihood approach where missing variables are imputed using nonparametric methods. Chown and Mstudy efficient estimation of the error distribution function in homoscedastic nonparametric regression using complete cases. González-Manteiga and Pérez-González [7] and Li [13] propose imputation to derive lack-of-fit tests based on suitable estimators of the error distribution function.

This article is organized as follows. In Section 2, we discuss estimating a functional of the conditional distribution Q of Y given X in the MAR model and characterize efficient estimators by deriving their influence function. This result is of independent interest. In Section 3, it is combined with the transfer principle of asymptotically linear estimators by [12] to yield our main result, the efficiency transfer, which states that the complete case version of an efficient estimator for the full model is efficient in the MAR model under a mild assumption. In our application, this assumption is typically implied by the requirement that π is bounded away from zero. The efficiency transfer is formulated for independent copies of a base observation $(\delta, X, \delta Y)$, with Y missing at random, but without assuming a *specific* regression model. Doing so would impose additional structure and therefore limit the generality of our statement. In Sections 4 and 5, we focus on the semiparametric regression model (1.1), with independent covariates and errors, and consider several important special cases and applications. In Theorem 4.1 in Section 4.1, we present the efficient influence function for a general functional of the conditional distribution for this family of models. In the same section, we study four types of functionals, the finite-dimensional regression parameter, linear functionals of the regression function, functionals of the infinite-dimensional regression parameter, and linear functionals of the error distribution such as the error variance and the error distribution function (see Examples 1–4). In Sections 4.2 and 4.3, we discuss the model class where the regression function only depends on ϑ and the model class where it only depends on ξ, that is, the special cases where the regression function is parametric or nonparametric. In Section 5, we illustrate our results in three specific regression models: the linear regression model as a special parametric regression model, the classical nonparametric regression models which only assumes a smooth regression function, and the partially linear random coefficient model (as an example of a more complex semiparametric regression model). The paper concludes (in Section 6) with a proof of Theorem 4.1.

2. Efficient influence functions

In this section, we derive the efficient influence function for estimating a general functional τ of the conditional distribution Q of Y given X in the MAR model and in the full model; see equations (2.4) and (2.6) below. We follow the approach outlined on page 58 in [2] as it is most suitable for estimating general functionals. It consists of two steps. First, one derives the tangent space of the model and then obtains the efficient influence function of a differentiable functional as the orthogonal projection of any gradient of the functional onto the tangent space. When estimating the finite-dimensional parameter in a semiparametric model, this approach reduces to the more familiar approach of projecting the score function of this parameter onto the tangent space of the nuisance parameter. The efficient influence function is then obtained as the inverse of the dispersion matrix of the efficient score function times the efficient score function, which is the difference of the score function and its projection. These approaches are illustrated in examples 2 and 3 of [2], pages 144–147, for estimating regression coefficients when covariates are missing completely at random; see also [31], who uses the familiar second approach to construct estimators for finite-dimensional parameters in semiparametric models with fully observed, missing and coarsened data.

We follow the calculations in [19] and consider a general missing data problem, with base observation $(\delta, X, \delta Y)$, where Y is missing at random, and where X and Y do not have to follow a regression model. Müller *et al.* expressed the joint distribution P of $(X, \delta Y, \delta)$ via

$$P(dx, dy, dz) = G(dx) B_{\pi(x)}(dz)\big(zQ(x, dy) + (1-z)\Delta_0(dy)\big)$$

in terms of the distribution G of X, the conditional probability $\pi(x)$ of $\delta = 1$ given $X = x$ (which comes in through the MAR assumption), and the conditional distribution $Q(x, dy)$ of Y given $X = x$. Here B_p denotes the Bernoulli distribution with parameter p and Δ_t the Dirac measure at t. We exclude the degenerate case that no responses are observed by assuming $E[\delta] > 0$.

The parameter for the above model is (G, π, Q). As parameter set we take the product $\mathcal{G} \times \mathcal{P} \times \mathcal{Q}$, where $\mathcal{G}$ is a model for the distribution G, $\mathcal{P}$ is a model for the propensity π, and $\mathcal{Q}$ is a model for the conditional distribution Q. This means that the parameters are not linked. The case that responses are missing completely at random can be modeled by taking the propensity to be constant and setting $\mathcal{P}$ to be the interval (0, 1]. The full model, that is, when responses are not missing, is also captured by taking $\pi = 1$ and $\mathcal{P} = \{1\}$.

The tangent space is the set of all perturbations of P. As in [19], we write this set as the sum of the orthogonal spaces

$$T_1 = \big\{u(X) : u \in \mathcal{U}\big\},$$

$$T_2 = \big\{\delta v(X, Y) : v \in \mathcal{V}(G_1)\big\},$$

$$T_3 = \big\{\big(\delta - \pi(X)\big)w(X) : w \in \mathcal{W}\big\}.$$

The set $\mathcal{U}$ consists of all real-valued functions u satisfying $\int u\,dG = 0$, $\int u^2\,dG < \infty$ and for which there is a sequence G_{nu} in $\mathcal{G}$ satisfying

$$\int \Big(n^{1/2}\big(dG_{nu}^{1/2} - dG^{1/2}\big) - \frac{1}{2} u\,dG^{1/2} \Big)^2 \to 0.$$

The set $\mathcal{W}$ consists of real-valued functions w with the property $\int w^2 \pi(1-\pi)\,dG < \infty$ for which there is a sequence π_{nw} in $\mathcal{P}$ such that

$$\int \sum_{z=0}^{1} \Big(n^{1/2}\big(dB_{\pi_{nw}(x)}^{1/2}(z) - dB_{\pi(x)}^{1/2}(z)\big) - \frac{1}{2}\big(z - \pi(x)\big) w(x)\,dB_{\pi(x)}^{1/2}(z) \Big)^2 G(dx) \to 0.$$

Finally, the set $\mathcal{V}(G_1)$ consists of functions v with the properties $\int v(x,y)Q(x,dy) = 0$ and $\int v^2(x,y)G_1(dx)Q(x,dy) < \infty$, and for which there is a sequence Q_{nv} in $\mathcal{Q}$ such that

$$\iint \Big(n^{1/2}\big(dQ_{nv}^{1/2}(x,\cdot) - dQ^{1/2}(x,\cdot)\big) - \frac{1}{2} v(x,\cdot)\,dQ^{1/2}(x,\cdot) \Big)^2 G_1(dx) \to 0. \tag{2.1}$$

Here G_1 is the conditional distribution of X given $\delta = 1$. It has density $\pi / \int \pi\,dG$ with respect to G. If π is bounded away from 0, then our formulation is equivalent to that in [19], who worked with G instead of G_1 in (2.1). Since we do not yet want to assume that π is bounded away from zero, we work with G_1 instead of G. Note that $\mathcal{V}(G_1)$ is the tangent set of the model $\mathcal{M}(G_1)$ of distributions $G_1 \otimes Q(dx,dy) = G_1(dx)Q(x,dy)$ with G_1 held fixed. Think of $G_1 \otimes Q$ as the distribution of a pair $(\tilde{X}, \tilde{Y})$ where $\tilde{X}$ has distribution G_1 and the conditional distribution of $\tilde{Y}$ given $\tilde{X}$ is Q.

We assume from now on that $\mathcal{V}(G_1)$ is a closed linear subspace of $L_2(G_1 \otimes Q)$. We are interested in estimating a characteristic of the conditional distribution Q, more formally a functional of the form

$$\kappa(G, \pi, Q) = \tau(Q).$$

For this, we assume that τ is differentiable with gradient $\gamma(\cdot\,; G_1)$ in $L_2(G_1 \otimes Q)$. This means that

$$n^{1/2}\big(\tau(Q_{nv}) - \tau(Q)\big) \to \int \gamma(x,y;G_1) v(x,y) G_1(dx) Q(x,dy) \tag{2.2}$$

holds for all $v \in \mathcal{V}(G_1)$ and with Q_{nv} as above. Let $\gamma_*(\cdot\,; G_1)$ denote the canonical gradient, i.e., the projection of $\gamma(\cdot\,; G_1)$ onto $\mathcal{V}(G_1)$ in $L_2(G_1 \otimes Q)$. Then $\gamma_*(\cdot\,; G_1)$ is the efficient influence function for estimating $\tau(Q)$ in the model $\mathcal{M}(G_1)$. Now set

$$\psi(\delta, X, \delta Y) = \frac{\delta}{E[\delta]} \gamma_*(X, Y; G_1). \tag{2.3}$$

Then $\psi(\delta, X, \delta Y)$ belongs to T_2. Thus by the orthogonality of the spaces T_1, T_2 and T_3, we have

$$E\big[\psi(\delta, X, \delta Y) u(X)\big] = 0, \qquad u \in \mathcal{U},$$
$$E\big[\psi(\delta, X, \delta Y)\big(\delta - \pi(X)\big) w(X)\big] = 0, \qquad w \in \mathcal{W}.$$

We calculate

$$E\big[\psi(\delta, X, \delta Y)\delta v(X, Y)\big] = \int \gamma_*(x, y; G_1)v(x, y)G_1(dx)Q(x, dy), \qquad v \in \mathcal{V}(G_1).$$

This shows that ψ is the canonical gradient for estimating $\kappa(G, \pi, Q) = \tau(Q)$ in our general missing data problem and hence the efficient influence function. This means an efficient estimator $\hat{\tau}_{\mathrm{MAR}}$ of $\kappa(G, \pi, Q) = \tau(Q)$ must satisfy the expansion

$$\hat{\tau}_{\mathrm{MAR}} = \tau(Q) + \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j}{E[\delta]}\gamma_*(X_j, Y_j; G_1) + o_P\big(n^{-1/2}\big). \tag{2.4}$$

We have just seen that the efficient influence function for our MAR model is the product of the influence function in the model $\mathcal{M}(G_1)$ and the factor $\delta/E[\delta]$. So in order to determine the efficient influence function ψ for a specific application, we only need to find the canonical gradient in the corresponding model $\mathcal{M}(G_1)$. In some cases, these canonical gradients are already available in the literature.

To derive the efficient influence function in the full model, consider the above with $\delta = 1$ (i.e., $\pi = 1$). Then T_2 becomes $\{v(X, Y) : v \in \mathcal{V}(G)\}$ and T_3 becomes $\{0\}$. The differentiability (2.2) of τ now needs to hold with G_1 replaced by G, i.e.,

$$n^{1/2}\big(\tau(Q_{nv}) - \tau(Q)\big) \to \int \gamma(x, y; G)v(x, y)G(dx)Q(x, dy). \tag{2.5}$$

The canonical gradient $\gamma_*(\cdot; G)$ is now the projection of $\gamma(\cdot; G)$ onto $\mathcal{V}(G)$ in $L_2(G \otimes Q)$ and the role of $\psi(\delta, X, \delta Y)$ is now played by $\gamma_*(X, Y; G)$. Thus, an efficient estimator $\hat{\tau}_{\mathrm{FULL}}$ of τ must satisfy

$$\hat{\tau}_{\mathrm{FULL}} = \tau(Q) + \frac{1}{n}\sum_{j=1}^{n} \gamma_*(X_j, Y_j; G) + o_P\big(n^{-1/2}\big). \tag{2.6}$$

Note that (2.2) implies (2.5) with $\gamma(X, Y; G) = \gamma(X, Y; G_1)\pi(X)/\int \pi\, dG$. If π is bounded away from zero, then (2.5) implies (2.2) with

$$\gamma(X, Y; G_1) = \gamma(X, Y; G)\int \pi\, dG/\pi(X).$$

3. Preservation of efficiency

In the previous section, we derived the efficient influence function for the full model and the MAR model. We now use this information to show that complete case versions of efficient estimators for the full model are efficient in the MAR model. For this, we rely on the transfer principle by [12] for asymptotically linear statistics. We state this principle first.

Let $(\delta_1, X_1, Y_1), \dots, (\delta_n, X_n, Y_n)$ be independent copies of (δ, X, Y). Consider a statistic

$$T_n = t_n(X_1, Y_1, \dots, X_n, Y_n)$$

and its complete case version

$$T_{n,c} = t_N(X_{i_1}, Y_{i_1}, \dots, X_{i_N}, Y_{i_N}),$$

where $N = \sum_{j=1}^n \delta_j$ denotes the number of complete observations and $i_1, \dots, i_N$ denote the indices of the complete observations. Let $\mathcal{M}$ denote a model of joint distributions of (X, Y) and T denote a function from $\mathcal{M}$ to $\mathbb{R}$. We assume that the original statistic has an influence function ϕ. More precisely, the following holds for each M in $\mathcal{M}$. If (X, Y) has distribution M, then the expansion

$$T_n = T(M) + \frac{1}{n}\sum_{j=1}^n \phi(X_j, Y_j; M) + o_P\bigl(n^{-1/2}\bigr)$$

holds with $E[\phi(X, Y; M)] = 0$ and $E[\phi^2(X, Y; M)]$ finite. The transfer principle for asymptotically linear estimators then gives the following result for the complete case version. If the conditional distribution M_1 of (X, Y) given $\delta = 1$ belongs to the model $\mathcal{M}$, then the complete case version obeys the expansion

$$T_{n,c} = T(M_1) + \frac{1}{n}\sum_{j=1}^n \frac{\delta_j}{E[\delta]}\phi(X_j, Y_j; M_1) + o_P\bigl(n^{-1/2}\bigr).$$

This shows that the complete case version of T_n is an estimator of $T(M_1)$ rather than $T(M)$. Note that this does not require the MAR assumption. Under our MAR assumption, if $M = G \otimes Q$, then $M_1 = G_1 \otimes M$ with G_1 the conditional distribution of X given $\delta = 1$. Thus, for estimating a functional $\tau(Q)$ of the conditional distribution Q, we have

$$T(M) = T(G \otimes Q) = \tau(Q) = T(G_1 \otimes Q) = T(M_1),$$

that is, in this case the original statistic and its complete case version are both consistent estimators of $\tau(Q)$. In particular, a complete case version of an efficient statistic will in general also be efficient in the MAR model, which we now present as the key result of this article.

Efficiency transfer

If the original statistic is efficient in the full model, then the function $\phi(X, Y; G \otimes Q)$ equals the efficient influence function $\gamma_*(X, Y; G)$ from the previous section, and we have

$$T_n = \tau(Q) + \frac{1}{n}\sum_{j=1}^n \gamma_*(X_j, Y_j; G) + o_P\bigl(n^{-1/2}\bigr)$$

and

$$T_{n,c} = \tau(Q) + \frac{1}{n}\sum_{j=1}^n \frac{\delta_j}{E[\delta]}\gamma_*(X_j, Y_j; G_1) + o_P\bigl(n^{-1/2}\bigr),$$

provided $G_1 \otimes Q$ belongs to $\mathcal{M}$. This shows that the complete case version of an efficient estimator is efficient in the model with missing data under the mild assumption that $G_1 \otimes Q$ belongs also to the model. We refer to this result as efficiency transfer.

Let us illustrate our findings with a simple example. Suppose we have a parametric model for the conditional distribution Q of Y given X and want to estimate a linear functional of the parameter. More precisely, we assume that $Q = Q_\vartheta$ for some m-dimensional parameter ϑ and that

$$\iint \big(dQ^{1/2}_{\vartheta+t}(x,dy) - dQ^{1/2}_{\vartheta}(x,dy) - (1/2)t^\top v_\vartheta(x,y)\,dQ^{1/2}_{\vartheta}(x,dy)\big)^2\,dG(x) = o\big(\|t\|^2\big)$$

holds with $\int v_\vartheta(x,y)\,dQ_\vartheta(x,dy) = 0$ and $\int \|v_\vartheta(x,y)\|^2 G(dx)\,dQ_\vartheta(x,dy) < \infty$. Thus the set $\mathcal{V}(G)$ equals the linear span $\{t^\top v_\vartheta : t \in \mathbb{R}^m\}$ of v_ϑ. We assume that the matrix

$$W(\vartheta, G) = \int v_\vartheta(x,y)v_\vartheta(x,y)^\top G(dx)\,dQ_\vartheta(x,dy)$$

is positive definite so that $\mathcal{V}(G)$ has dimension m. Then an efficient estimator in the full model for a linear functional $\tau(Q_\vartheta) = a^\top\vartheta$ of ϑ has efficient influence function $a^\top\gamma(X,Y;G)$ where $\gamma(X,Y;G) = W(\vartheta,G)^{-1}v_\vartheta(X,Y)$. Indeed, we have

$$n^{1/2}\big(a^\top\big(\vartheta + n^{1/2}t\big) - a^\top\vartheta\big) \to a^\top t = a^\top \iint \gamma(x,y;G)\big(t^\top v_\vartheta(x,y)\big)G(dx)Q_\vartheta(x,dy).$$

It is easy to see that G_1 satisfies the same assumptions as G as long as $W(\vartheta, G_1)$ is positive definite. The complete case version of an efficient estimator of $\tau(Q_\vartheta)$ is therefore efficient in the MAR model if this condition is met. Its influence function is $(\delta/E[\delta])a^\top\gamma(X,Y;G_1) = (\delta/E[\delta])a^\top W(\vartheta,G_1)^{-1}v_\vartheta(X,Y)$.

The above shows that an efficient estimator of ϑ in the full model has influence function $\gamma(X,Y;G)$. Typically, the maximum likelihood estimator is efficient in the full model, and its complete case version is then efficient in the MAR model.

Remark 3.1. Consider the case where, in addition to Y, X is also missing with indicator γ. Here the base observation is the quadruple $(\gamma,\delta,\gamma X,\delta Y)$. Suppose now that an analogue to the MAR condition holds:

$$P(\gamma = i, \delta = j | X, Y) = P(\gamma = i, \delta = j | X), \qquad i = 0,1,\ j = 0,1.$$

In this case, a complete case analysis is still valid in the sense of leading to (asymptotically) unbiased estimation of characteristics of the conditional distribution Q. In particular, complete case versions of $n^{1/2}$-consistent estimators will preserve this property. The efficiency transfer, however, typically does not carry over to this more general setting, that is, complete case versions of estimators efficient in the full model will no longer be automatically efficient. This follows from the fact that on the event $\{(\gamma,\delta) = (0,1)\}$ one still observes Y, but omits it from the analysis. The conditional distribution of Y given $(\gamma,\delta) = (0,1)$, however, depends on Q and thus carries information about Q, which typically cannot be ignored for efficiency purposes.

4. Application to regression

In this section, we specialize the previous results to the large class of homoscedastic regression models that have the form (1.1), that is,

$$Y = r(X, \vartheta, \xi) + \varepsilon.$$

We assume that the mean zero error ε is independent of the covariate X, with finite variance $\sigma^2 > 0$, distribution function F, and a density f with finite Fisher information for location. The latter means that f is absolutely continuous and that the score function $\ell_f = -f'/f$ has a finite second moment $J_f = \int \ell_f^2 \, dF$. The regression function r is assumed to depend (smoothly) on a p-dimensional parameter ϑ and some infinite-dimensional parameter ξ. We begin with the general semiparametric regression model and then discuss special cases where the regression function does not depend on ξ or ϑ.

4.1. Semiparametric regression function

In the general homoscedastic semiparametric regression model, we have

$$Q(x, dy) = Q_{\vartheta,\xi,f}(x, dy) = f\bigl(y - r(x, \vartheta, \xi)\bigr) \, dy.$$

To find the efficient influence function for functionals of Q with missing data, we derive the efficient influence function for model $\mathcal{M}(G_1)$, which is the regression model

$$\tilde{Y} = r(\tilde{X}, \vartheta, \xi) + \tilde{\varepsilon},$$

where the error $\tilde{\varepsilon}$ is independent of the covariate $\tilde{X}$, $\tilde{X}$ has distribution G_1 and $\tilde{\varepsilon}$ has the same distribution as ε. The results of Section 2 immediately provide the efficient influence function for the MAR model; see (2.3). The efficient influence functions for model $\mathcal{M}(G_1)$ and the MAR model are given later in this section in Theorem 4.1.

As shown in [30], the tangent set $\mathcal{V}(G_1)$ consists of the functions

$$v(\tilde{X}, \tilde{Y}) = \bigl\{a^\top h(\tilde{X}) + b(\tilde{X})\bigr\}\ell_f(\tilde{\varepsilon}) + c(\tilde{\varepsilon}), \tag{4.1}$$

where a belongs to $\mathbb{R}^p$, h to $L_2^p(G_1)$, b to some closed linear subspace B of $L_2(G_1)$, and c is a member of $\mathcal{C}$, where

$$\mathcal{C} = \left\{c \in L_2(F) : \int c(y) f(y) \, dy = \int y c(y) f(y) \, dy = 0\right\}.$$

This requires that for each b in B there is a sequence ξ_{nb} such that

$$\int \Bigl(n^{1/2}\bigl(r\bigl(x, \vartheta + n^{-1/2}a, \xi_{nb}\bigr) - r(x, \vartheta, \xi)\bigr) - a^\top h(X) - b(x)\Bigr)^2 \, dG_1(x) = o(1) \tag{4.2}$$

for all $a \in \mathbb{R}^p$. Here h is the $L_2(G_1)$-derivative of $t \mapsto r(\cdot, t, \xi)$ at ϑ. Note that $\mathcal{C}$ is the tangent space for the error densities with zero mean, finite variance and finite Fisher information for location. For each $c \in \mathcal{C}$, there is a sequence f_{nc} of such densities such that

$$\int \bigl(n^{1/2}\bigl(f_{nc}^{1/2}(y) - f^{1/2}(y)\bigr) - (1/2)c(y)f^{1/2}(y)\bigr)^2\,dy = o(1). \tag{4.3}$$

We then have (2.1) if we take $Q_{nv} = Q_{\vartheta+n^{-1/2}a,\xi_{nb},f_{nc}}$ and $v(\tilde{X}, \tilde{Y})$ as in (4.1).

We are interested in estimating a functional

$$\tau(Q_{\vartheta,\xi,f}) = \tau_0(\vartheta, \xi, f)$$

of the regression parameters ϑ and ξ and the error density f. We assume that the sequences ξ_{nb} and f_{nc} can be chosen such that, in addition to (4.2) and (4.3),

$$n^{1/2}\bigl(\tau_0\bigl(\vartheta + n^{-1/2}a, \xi_{nb}, f_{nc}\bigr) - \tau_0(\vartheta, \xi, f)\bigr) \to a_*^\top a + \int b_* b\,dG_1 + \int c_* c\,dF \tag{4.4}$$

holds for all $a \in \mathbb{R}^p$, $b \in B$ and $c \in \mathcal{C}$ and for some $a_* \in \mathbb{R}^p$, $b_* \in B$ and $c_* \in \mathcal{C}$. To describe the efficient influence function we need to introduce some additional notation.

For a closed linear subspace L of $L_2(G_1)$, let Π_L denote the projection operator onto L in $L_2(G_1)$, and let $d_L = \Pi_L(1)$ denote the projection of 1 onto L. We introduce the constants

$$\Delta = \frac{J_f - 1/\sigma^2}{J_f} = 1 - \frac{1}{J_f\sigma^2} \quad \text{and} \quad \rho_L = \frac{1}{1 - \Delta \int d_L\,dG_1}.$$

Now set

$$h_* = \bigl(h_1 - \Pi_B(h_1), \ldots, h_p - \Pi_B(h_p)\bigr)^\top$$

and introduce the matrix

$$H_* = \int h_* h_*^\top\,dG_1.$$

Note that the space $K = \{a^\top h + b : a \in \mathbb{R}^p, b \in B\}$ can be expressed as the sum $A + B$ of the orthogonal spaces $A = \{a^\top h_* : a \in \mathbb{R}^p\}$ and B. This implies

$$d_K = d_A + d_B.$$

Finally, for χ in $L_2^m(G_1)$ and a closed linear subspace L of $L_2(G_1)$, we write

$$\bar{\chi} = \int \chi\,dG_1, \qquad M\chi = \chi - \Delta\bar{\chi}, \qquad \Gamma_L\chi = \chi + \Delta\rho_L d_L\bar{\chi},$$

and

$$D\chi(\tilde{X}, \tilde{Y}) = \bigl(\chi(\tilde{X}) - \bar{\chi}\bigr)\ell_f(\tilde{\varepsilon}) + \bar{\chi}\frac{\tilde{\varepsilon}}{\sigma^2}.$$

Using the identity $1 + \Delta \bar{d}_L \rho_L = \rho_L$, we obtain

$$M(\Gamma_L \chi) = \chi - \Delta \rho_L (1 - d_L)\bar{\chi} \quad \text{and} \quad \Gamma_L d_L = \rho_L d_L.$$

Theorem 4.1. *Suppose the differentiability conditions* (4.2)–(4.4) *hold and the matrix* H_* *is positive definite. Then the efficient influence function for estimating* $\tau_0(\vartheta, \xi, f)$ *in model* $\mathcal{M}(G_1)$ *is*

$$\gamma_*(\tilde{X}, \tilde{Y}; G_1) = c_*(\tilde{\varepsilon}) + \frac{1}{J_f} D\big[(a_* - \alpha)^\top H_*^{-1} \Gamma_K h_* + \Gamma_B b_* - \tilde{c}_* \Gamma_K d_K\big](\tilde{X}, \tilde{Y})$$

with $a_* \in \mathbb{R}^p$, $b_* \in B$ *and* $c_* \in \mathcal{C}$ *as in* (4.4),

$$\alpha = \int M(\Gamma_B b_*) h \, dG_1 \quad \textit{and} \quad \tilde{c}_* = \int c_* \ell_f \, dF.$$

The efficient influence function in the MAR model is therefore

$$\frac{\delta}{E[\delta]}\bigg[c_*(\varepsilon) + \frac{1}{J_f} D\big[(a_* - \alpha)^\top H_*^{-1} \Gamma_K h_* + \Gamma_B b_* - \tilde{c}_* \Gamma_K d_K\big](X, Y)\bigg].$$

The proof of Theorem 4.1 is deferred to Section 6. Straightforward calculations show that the constant α can be expressed as

$$\alpha = \int b_* h \, dG_1 - \Delta \rho_B \bar{b}_* \int (1 - d_B) h \, dG_1$$

and simplifies to

$$\alpha_* = \int b_* h \, dG_1$$

if d_B equals 1. The latter happens if and only if B contains the constant functions.

We now use the theorem to describe efficient influence functions for some important functionals. We will derive influence functions for efficient estimators of ϑ, of functionals of ξ, and of the error distribution F.

Example 1 (Estimating the finite-dimensional parameter). For $a_0 \in \mathbb{R}^p$, the functional $\tau_0(\vartheta, \xi, f) = a_0^\top \vartheta$ satisfies (4.4) with $a_* = a_0$, $b_* = 0$ and $c_* = 0$. Hence, using Theorem 4.1, the corresponding efficient influence function in model $\mathcal{M}(G_1)$ is given by $a_0^\top (J_f H_*)^{-1} D h_\#(\tilde{X}, \tilde{Y})$ with $h_\# = \Gamma_K h_*$. This implies that the efficient influence function for the finite dimensional parameter ϑ is

$$(J_f H_*)^{-1} D h_\#(\tilde{X}, \tilde{Y}) = (J_f H_*)^{-1} \big[\big(h_\#(\tilde{X}) - \bar{h}_\#\big)\ell_f(\tilde{\varepsilon}) + \bar{h}_\# \tilde{\varepsilon}/\sigma^2\big].$$

It simplifies to

$$(J_f H_*)^{-1} h_*(\tilde{X}) \ell_f(\tilde{\varepsilon})$$

if B contains the constants in which case we have $\bar{h}_* = 0$ and $h_\# = h_*$. These results can also be found in [30]; his w corresponds to our $H_*^{-1} h_\#$. The efficient influence function for estimating ϑ with missing data is

$$\frac{\delta}{E[\delta]}(J_f H_*)^{-1}\big[\big(h_\#(X) - \bar{h}_\#\big)\ell_f(\varepsilon) + \bar{h}_\# \varepsilon/\sigma^2\big].$$

It simplifies to

$$\frac{\delta}{E[\delta]}(J_f H_*)^{-1} h_*(X)\ell_f(\varepsilon)$$

if B contains the constant functions. For the construction of efficient estimators in the full model we refer to [30] and [6]. Thus, complete case versions of these estimators will be efficient with missing data under mild assumptions.

Example 2 (Estimating a linear functional of the regression function). Let us consider the functional

$$\tau_0(\vartheta, \xi, f) = \tau_1(\vartheta, \xi) = \int w(x) r(x, \vartheta, \xi)\, dx$$

for some measurable function w. If w is an indicator of a set, this functional represents the area under the regression curve over this set. We assume that G_1 has a density g_1 and that w/g_1 belongs to $L_2(G_1)$. Then we have

$$\begin{aligned} &n^{1/2}\big(\tau_1\big(\vartheta + n^{-1/2}a, \xi_{nb}\big) - \tau_1(\vartheta, \xi)\big) \\ &\quad = \int \frac{w(x)}{g_1(x)} n^{1/2}\big(r\big(x, \vartheta + n^{-1/2}a, \xi_{nb}\big) - r(x, \vartheta, \xi)\big)\, dG_1(x) \\ &\quad \to \int \frac{w}{g_1}\big(a^\top h + b\big)\, dG_1, \qquad a \in \mathbb{R}^p, b \in B, \end{aligned}$$

which gives us differentiability with $c_* = 0$, $b_* = \Pi_B(w/g_1)$ and $a_* = \int w(x)h(x)\, dx$. Thus the efficient influence function for estimating $\tau_1(\vartheta, \xi)$ with missing data is

$$\frac{\delta}{E[\delta]}\left[\frac{1}{J_f} D\big[(a_* - \alpha)^\top H_*^{-1} \Gamma_K h_* + \Gamma_B b_*\big](X, Y)\right].$$

This simplifies to

$$\frac{\delta}{E[\delta]}\left[\frac{1}{J_f} D\big[(a_* - \alpha_*)^\top H_*^{-1} h_* + \Gamma_B b_*\big](X, Y)\right]$$

if B contains the constant functions.

Example 3 (Estimating a functional of the infinite-dimensional parameter). Now consider estimating a functional

$$\tau_0(\vartheta, \xi, f) = \tau_2(\xi)$$

of the infinite dimensional parameter ξ only. We assume that there is a b_* in B such that

$$n^{1/2}\big(\tau_2(\xi_{nb}) - \tau_2(\xi)\big) \to \int b_* b \, dG_1$$

holds for all b in B. This yields (4.4) with $a_* = 0$ and $c_* = 0$. The efficient influence function for estimating $\tau_2(\xi)$ with missing data is thus

$$\frac{\delta}{E[\delta] J_f} D\big[\Gamma_B b_* - \alpha^\top H_*^{-1} h_\#\big](X, Y),$$

where $h_\#$ equals $\Gamma_K h$ as in Example 1. This simplifies to

$$\frac{\delta}{E[\delta]}\bigg[\big(b_*(X) - \bar{b}_* - \alpha_*^\top H_*^{-1} h_*(X)\big)\frac{\ell_f(\varepsilon)}{J_f} + \bar{b}_* \varepsilon\bigg]$$

if B contains the constant functions.

Example 4 (Estimating functionals of the error distribution). Now we look at estimating a linear functional of the error distribution,

$$\tau_0(\vartheta, \xi, f) = \tau_3(f) = \int \phi(x) f(x) \, dx,$$

for some measurable function ϕ. This includes estimating the error variance σ^2 by taking $\phi(x) = x^2$ and estimating $F(y)$, the error distribution function at a fixed point y, by taking $\phi(x) = \mathbf{1}[x \le y]$. We assume that $\int \phi^2 \, dF$ is finite. For each $c \in \mathcal{C}$, we can choose f_{nc} such that (4.3) and

$$n^{1/2} \int \phi(x)\big(f_{nc}(x) - f(x)\big) \, dx \to \int \phi c \, dF$$

hold. Hence, we have (4.4) with $a_* = 0$, $b_* = 0$ and $c_* = \phi_*$, where ϕ_* is the projection of ϕ onto $\mathcal{C}$. We have

$$\phi_*(\tilde{\varepsilon}) = \phi(\tilde{\varepsilon}) - \int \phi \, dF - \int \phi(x) x f(x) \, dx \frac{\tilde{\varepsilon}}{\sigma^2}.$$

The efficient influence function for estimating $\int \phi \, dF$ with missing data is therefore

$$\frac{\delta}{E[\delta]}\bigg[\phi_*(\varepsilon) - \int \phi_* \ell_f \, dF \, T_K(X, Y)\bigg], \tag{4.5}$$

where

$$\begin{aligned} T_K(X, Y) &= \frac{1}{J_f} D \Gamma_K \, d_K(X, Y) = \frac{\rho_K}{J_f} D \, d_K(X, Y) \\ &= \rho_K \bigg[\big(d_K(X) - \bar{d}_K\big)\frac{\ell_f(\varepsilon)}{J_f} + (1 - \Delta)\bar{d}_K \varepsilon\bigg]. \end{aligned}$$

If K contains the constant functions, then we have $d_K = 1$ and thus $T_K(X, Y) = \varepsilon$, and the efficient influence function simplifies to

$$\frac{\delta}{E[\delta]}\left[\phi(\varepsilon) - \int \phi \, dF - \int \phi \ell_f \, dF \varepsilon\right].$$

This result was derived for classical nonparametric regression without missing data in [17].

Let us mention two special cases. The efficient influence function for estimating the error variance σ^2 is

$$\frac{\delta}{E[\delta]}\left[\varepsilon^2 - \sigma^2 - \rho\varepsilon + \rho T_K(X, Y)\right]$$

with $\rho = \int x^3 \, dF(x)/\sigma^2$. This requires the error distribution to have a finite fourth moment and uses the identity $\int x^2 \ell_f(x) \, dF(x) = 0$. The efficient influence function for estimating $F(y)$ is

$$\frac{\delta}{E[\delta]}\left[\mathbf{1}[\varepsilon \le y] - F(y) - \nu(y)\varepsilon + \big(f(y) + \nu(y)\big)T_K(X, Y)\right]$$

with $\nu(y) = \int_{-\infty}^{y} x \, dF(x)/\sigma^2$. If K contains the constant functions, then the above influence functions simplify to

$$\frac{\delta}{E[\delta]}\left(\varepsilon^2 - \sigma^2\right)$$

and

$$\frac{\delta}{E[\delta]}\left[\mathbf{1}[\varepsilon \le y] - F(y) + f(y)\varepsilon\right].$$

The latter was derived directly by [4] for classical nonparametric regression with missing data. Müller, Schick and Wefelmeyer [20] obtained an analogous result for the full partly linear regression model.

In the following two subsections, we discuss modifications to cases when either the role of ξ or the role of ϑ is void.

4.2. Parametric regression function

Consider the parametric regression model $Y = r_\vartheta(X) + \varepsilon$ where ε and X are as before. One typically assumes that $r_t(x)$ is differentiable in t with gradient $\dot{r}_t(x)$. We also assume that

$$\int \left(r_{\vartheta+a} - r_\vartheta - a^\top \dot{r}_\vartheta\right)^2 dG_1 = o\big(|a|^2\big)$$

and that the matrix

$$H = \int \dot{r}_\vartheta \dot{r}_\vartheta^\top \, dG_1$$

is positive definite. This model does not involve ξ. Hence, we have

$$Q(x,dy) = Q_{\vartheta,f}(x,dy) = f\big(y - r_\vartheta(x)\big)\,dy,$$

and the functional of interest is

$$\tau(Q_{\vartheta,f}) = \tau_0(\vartheta, f).$$

We assume τ_0 to be differentiable in the sense that

$$n^{1/2}\big(\tau_0\big(\vartheta + n^{-1/2}a, f_{nc}\big) - \tau_0(\vartheta, f)\big) \to a_*^\top a + \int c_* c\, dF$$

for all $a \in \mathbb{R}^p$, $c \in \mathcal{C}$ and some $a_* \in \mathbb{R}^p$, $c_* \in \mathcal{C}$. The tangent associated with the perturbed version $Q_{\vartheta+n^{-1/2},f_{nc}}$ of $Q_{\vartheta,f}$ is

$$a^\top \dot r_\vartheta(\tilde X)\ell_f(\varepsilon) + c(\tilde\varepsilon).$$

Here a belongs to $\mathbb{R}^p$ and c to $\mathcal{C}$. So we have $h = \dot r_\vartheta$, $B = \{0\}$ and $b_* = 0$ and have $d_B = 0$, $d_K = \bar h^\top H^{-1} h$ and

$$\begin{aligned} H^{-1}\Gamma_K h &= \big(I + \Delta\rho_K H^{-1}\bar h\bar h^\top\big)H^{-1}h \\ &= \big(I - \Delta H^{-1}\bar h\bar h^\top\big)^{-1}H^{-1}h \\ &= J_f\big(J_f\big(H - \bar h\bar h^\top\big) + \big(1/\sigma^2\big)\bar h\bar h^\top\big)^{-1}h. \end{aligned}$$

The efficient influence function for estimating $\tau_0(\vartheta, f)$ in the MAR model is

$$\frac{\delta}{E[\delta]}\big[c_*(\varepsilon) + v_*^\top Dh(X,Y)\big]$$

with

$$v_* = a_*^\top\big(J_f\big(H - \bar h\bar h^\top\big) + \big(1/\sigma^2\big)\bar h\bar h^\top\big)^{-1} - \frac{\tilde c_*}{J_f(1 - \Delta\bar h H^{-1}\bar h)}\bar h^\top H^{-1}.$$

The efficient influence function for estimating ϑ is

$$\frac{\delta}{E[\delta]}\big\{J_f\big(H - \bar h\bar h^\top\big) + \big(1/\sigma^2\big)\bar h\bar h^\top\big\}^{-1}\bigg[\big\{h(X) - \bar h\big\}\ell_f(\varepsilon) + \bar h\frac{\varepsilon}{\sigma^2}\bigg]. \tag{4.6}$$

This was derived directly in [15].

The efficient influence function for estimating $\int \phi\, dF$ with $\int \phi^2\, dF < \infty$ is as in (4.5) with $d_K = \bar h^\top H^{-1} h$. Thus, the formulas for the efficient influence functions for estimating σ^2 and $F(y)$ given in Example 4 remain valid with the present d_K.

Here a natural model for the covariate distribution is the set $\mathcal{G}$ of all distributions G such that $\int (r_{\vartheta+a} - r_\vartheta - a^\top \dot r_\vartheta)^2\, dG = o(|a|^2)$ and the matrix $\int \dot r_\vartheta \dot r_\vartheta^\top\, dG$ is positive definite. If π is bounded away from zero, then G in $\mathcal{G}$ implies G_1 in $\mathcal{G}$. The efficiency transfer is therefore valid if π is bounded away from zero.

4.3. Generalized nonparametric regression function

We now treat the case where there is no finite-dimensional parameter: the model is $Y = r(X, \xi) + \varepsilon$ with ε and X as before. This covers the classical nonparametric model in which $r(x, \xi) = \xi(x)$ and ξ is a smooth function, additive regression in which $r(x, \xi) = \xi_1(x_1) + \cdots + \xi_p(x_p)$ with smooth univariate functions $\xi_1, \ldots, \xi_p$, and random coefficient models in which $r(x, \xi) = \xi_1(x_p)x_1 + \cdots + \xi_{p-1}(x_p)x_{p-1}$ with smooth univariate functions $\xi_1, \ldots, \xi_{p-1}$.

In the present setting, the conditional distribution Q is of the form

$$Q(x, dy) = Q_{\xi,f}(x, dy) = f\big(y - r(x, \xi)\big)\, dy.$$

The functional of interest is

$$\tau(Q_{\xi,f}) = \tau_0(\xi, f).$$

The analogues of (4.2) and (4.4) are

$$\int \big(n^{1/2}\big(r(x, \xi_{nb}) - r(x, \xi)\big) - b(x)\big)^2\, dG_1 = o(1)$$

and

$$n^{1/2}\big(\tau_0(\xi_{nb}, f_{nc}) - \tau_0(\vartheta, f)\big) \to \int b_* b\, dG_1 + \int c_* c\, dF.$$

The tangent generated by the perturbed version $Q_{\xi_{nb}, f_{nc}}$ is

$$b(\tilde{X})\ell_f(\tilde{\varepsilon}) + c(\tilde{\varepsilon}).$$

This is essentially the case $h = 0$ and $a_* = 0$. The efficient influence function for estimating $\tau_0(\vartheta, f)$ with missing data is therefore

$$\frac{\delta}{E[\delta]}\bigg[c_*(\varepsilon) + \frac{1}{J_f} D[\varGamma_B b_* - \tilde{c}_* \varGamma_B d_B](X, Y)\bigg].$$

If B contains the constant functions, this simplifies to

$$\frac{\delta}{E[\delta]}\bigg[c_*(\varepsilon) - \tilde{c}_*\varepsilon + \big(b_*(X) - \bar{b}_*\big)\frac{\ell_f(\varepsilon)}{J_f} + b_*\varepsilon\bigg].$$

Consider estimating

$$\tau_0(\xi, f) = \tau_1(\xi) = \int w(x) r(x, \xi)\, dx$$

for some measurable function w. Suppose that G_1 has a density g_1 and that w/g_1 belongs to $L_2(G_1)$. Then we have

$$n^{1/2}\big(\tau_1(\xi_{nb}) - \tau_1(\xi)\big) \to \int \frac{w}{g_1} b\, dG_1$$

and thus differentiability with $c_* = 0$ and $b_* = \Pi_B(w/g_1)$. Therefore, the efficient influence function for estimating $\tau_1(\xi)$ is

$$\frac{\delta}{E[\delta]}\left[\frac{1}{J_f} D[\Gamma_B b_*](X,Y)\right].$$

This simplifies to

$$\frac{\delta}{E[\delta]}\left[\big(b_*(X) - \bar{b}_*\big)\frac{\ell_f(\varepsilon)}{J_f} + b_*\varepsilon\right]$$

if B contains the constant functions.

The efficient influence function for estimating $\int \phi\, dF$ with $\int \phi^2\, dF$ finite is given by

$$\frac{\delta}{E[\delta]}\left[\phi_*(\varepsilon) - \frac{\int \phi_* \ell_f\, dF}{(1-\Delta \bar{d}_B)J_f} D\, d_B(X,Y)\right].$$

Thus, the efficient influence function is as in Example 4 with $d_K = d_B$. The efficient influence function simplifies to

$$\frac{\delta}{E[\delta]}\left[\phi(\varepsilon) - \int \phi\, dF - \int \phi \ell_f\, dF \varepsilon\right]$$

if B contains the constants. This holds for the classical nonparametric regression model and for the additive regression model, but typically not for the random coefficient model.

5. Examples of regression models

In this section, we discuss three specific regression models. We begin with the fundamental linear regression model as a particular parametric regression model (see Section 4.2). We treat this model in detail to demonstrate how the results from the previous section translate to a specific application. The second model is the classical nonparametric regression model with a regression function that is only assumed to be smooth. This illustrates the results from Section 4.3. Our third model is the partially linear random coefficient model, as an example with a more complex semiparametric regression function. This model covers the partially linear model as a special case.

5.1. Linear regression

A special case of parametric regression is *linear* regression,

$$Y = \vartheta^\top h(X) + \varepsilon$$

for some measurable function h. We assume that $E[\|h(X)\|^2] = \int \|h\|^2\, dG$ is finite and that the matrix $H_G = \int hh^\top\, dG$ is positive definite. For the efficiency transfer to hold, we have to assume that these assumptions are met by G_1. It is easy to see that this amounts to requiring that $H_{G_1} = \int hh^\top \, dG_1$ is positive definite.

Let us set $\bar{h}_G = \int h \, dG$. The influence function of an efficient estimator of ϑ in the full model is

$$\lambda(X, Y; G) = \big[\big(H_G - \bar{h}_G \bar{h}_G^\top\big) J_f + \big(1/\sigma^2\big) \bar{h}_G \bar{h}_G^\top\big]^{-1} \big[\big(h(X) - \bar{h}_G\big) \ell_f(\varepsilon) + \bar{h}_G \varepsilon/\sigma^2\big].$$

By the efficiency transfer, its complete case version will be efficient with influence function $\delta\lambda(X, Y; G_1)/E[\delta]$. This holds for the efficient estimators given by [30].

Let us briefly look at an important special case. In *simple* linear regression, X is one-dimensional and $h(X) = (1, X)^\top$. Then the matrix H_G is positive definite if and only if $\mathrm{Var}_G(X)$ is positive. It is easy to check that the influence function above simplifies to $(\varepsilon - \mu_G \chi(X, Y; G), \chi(X, Y; G))^\top$, where

$$\chi(X, Y; G) = \frac{(X - \mu_G)\ell_f(\varepsilon)}{\mathrm{Var}_G(X) J_f}$$

is the efficient influence function for the slope. Here $\mu_G = \int x \, dG(x)$ is the mean of G. The construction of efficient estimators of the slope has been addressed in the literature. The usual approach is to estimate the score function ℓ_f. Bickel [1] uses sample splitting and kernel density estimators; [29] avoids sample splitting. Jin [8] uses splines to estimate the score function. In a recent preprint, [16] propose a different approach that does not require estimating the score function. They show that a maximum empirical likelihood estimator with an increasing number of random constraints is efficient. The complete case versions of these estimators are therefore efficient for missing responses provided $\mathrm{Var}_{G_1}(X)$ is positive. Müller *et al.* use the transfer principle and the characterization of efficient estimators in [15] to obtain this result. The assumption on $\mathrm{Var}_{G_1}(X)$ rules out that G_1 is concentrated at a single point. For example, if G is discrete, then π needs to be positive at a minimum of two points in the support of G, but can be zero at the other support points.

We now return to the general case and address estimation of linear functionals of the error density, namely the error variance and the error distribution function at a fixed point y. We first look at the case when $a_h^\top h = 1$ for some a_h in $\mathbb{R}^p$. This condition is met if the first coordinate of h is 1 so that the model contains an intercept as in the simple linear regression model. Note that the vector a_h must equal $H_G^{-1} \bar{h}_G$.

A commonly used estimator of ϑ is the least squares estimator

$$\hat{\vartheta}_L = \left(\frac{1}{n} \sum_{j=1}^n h(X_j) h(X_j)^\top\right)^{-1} \frac{1}{n} \sum_{j=1}^n h(X_j) Y_j.$$

This estimator has influence function $H_G^{-1} h(X)\varepsilon$ which coincides with the efficient influence function if and only if $\ell_f(\varepsilon)$ equals ε/σ^2, which is the case only if the error density is a centered normal density. Thus, the least squares estimator is efficient only if the errors happen to be normal.

With the least squares estimator, we associate the residuals $\hat{\varepsilon}_{L,j} = Y_j - \hat{\vartheta}_L^\top h(X_j)$. This suggests the following estimator of the error variance

$$\hat{\sigma}_L^2 = \frac{1}{n}\sum_{j=1}^n \hat{\varepsilon}_{L,j}^2.$$

It is easy to confirm that this estimator obeys the expansion

$$\hat{\sigma}_L^2 = \frac{1}{n}\sum_{j=1}^n \varepsilon_j^2 + O_p\big(n^{-1}\big).$$

This shows that $\hat{\sigma}_L^2$ is efficient if the errors have a finite fourth moment. Indeed, the influence function of $\hat{\sigma}_L^2$ is $\varepsilon^2 - \sigma^2$ and coincides with the efficient influence function in view of $a_h^\top h = 1$ (see Example 4 in the previous section).

The residual-based empirical distribution function

$$\hat{\mathbb{F}}_L(y) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}[\hat{\varepsilon}_{L,j} \le y], \qquad y \in \mathbb{R},$$

is an estimator of the error distribution function F. Since the error density f is uniformly continuous we have, for every root-n consistent estimator $\hat{\vartheta}$ of ϑ and corresponding residuals $\hat{\varepsilon}_j = Y_j - \hat{\vartheta}^\top H(X_j)$,

$$\sup_{y\in\mathbb{R}}\left|\frac{1}{n}\sum_{j=1}^n \mathbf{1}[\hat{\varepsilon}_j \le y] - \frac{1}{n}\sum_{j=1}^n \mathbf{1}[\varepsilon_j \le y] - f(y)\bar{h}_G^\top(\hat{\vartheta} - \vartheta)\right| = o_p\big(n^{-1/2}\big); \tag{5.1}$$

see Müller, Schick and Wefelmeyer ([20,21]) and the earlier work by [11]. Applying this with $\hat{\vartheta} = \hat{\vartheta}_L$ and observing that $\bar{h}_G^\top \hat{\vartheta}_L$ has influence function $\bar{h}_G^\top H_G^{-1} h(X)\varepsilon = a_h^\top h(X)\varepsilon = \varepsilon$, we obtain

$$\sup_{y\in\mathbb{R}}\left|\hat{\mathbb{F}}_L(y) - \frac{1}{n}\sum_{j=1}^n \big(\mathbf{1}[\varepsilon_j \le y] + f(y)\varepsilon_j\big)\right| = o_p\big(n^{-1/2}\big).$$

This shows that $\hat{\mathbb{F}}_L(y)$ is an efficient estimator of $F(y)$ for each $y \in \mathbb{R}$.

From now on, we no longer require that there be a vector a_h such that $a_h^\top h = 1$. In this case, efficient estimation of σ^2 and $F(y)$ becomes more complicated. By the results in Section 4.2 and Example 4, the efficient influence function for estimating σ^2 is

$$\varepsilon^2 - \sigma^2 - \rho\varepsilon + \rho T(X, Y; G),$$

while the efficient influence function for estimating $F(y)$ is

$$\mathbf{1}[\varepsilon \le y] - F(y) - \nu(y)\varepsilon + \big(f(y) + \nu(y)\big)T(X, Y; G).$$

Here ρ and $\nu(y)$ are as in Example 4 and $T(X, Y; G)$ is given by

$$T(X, Y; G) = \frac{1}{1 - \Delta E[d(X; G)]} \left[\big(d(X; G) - E\big[d(X; G)\big] \big) \frac{\ell_f(\varepsilon)}{J_f} + (1 - \Delta) E\big[d(X; G)\big] \varepsilon \right]$$

with $d(X; G) = \bar{h}_G H_G^{-1} h(X)$. In what follows, we shall need the fact that $T(X, Y; G)$ is the influence function of $\bar{h}_G^\top \hat{\vartheta}$ when $\hat{\vartheta}$ is efficient. This follows using the calculations in Section 4.2.

We now work with the residuals $\hat{\varepsilon}_j = Y_j - \hat{\vartheta}^\top h(X_j)$, where $\hat{\vartheta}$ for the moment is a root-n consistent estimator of ϑ. Even for the least squares estimator these residuals are no longer guaranteed to sum to zero. That property captured the information that the error distribution has zero in the previous setting. The equivalent here is to use a weighted residual distribution function where the weights are the maximizers of the empirical likelihood

$$\sup \left\{ \prod_{j=1}^n n\pi_j : \pi_1 \geq 0, \dots, \pi_n \geq 0, \sum_{j=1}^n \pi_j = 1, \sum_{j=1}^n \pi_j \hat{\varepsilon}_j = 0 \right\},$$

which imposes this constraint. It follows from Owen ([24,25]) that the maximizers are of the form $\hat{\pi}_j = 1/(1 + \zeta \hat{\varepsilon}_j)$ where the Lagrange multiplier ζ is a random variable such that $1 + \zeta \hat{\varepsilon}_1, \dots, 1 + \zeta \hat{\varepsilon}_n$ are positive and

$$\frac{1}{n} \sum_{j=1}^n \frac{\hat{\varepsilon}_j}{1 + \zeta \hat{\varepsilon}_j} = 0$$

holds. Such a random variable exists except on an event whose probability tends to zero. This idea was used by Müller, Schick and Wefelmeyer ([18,19]) in the context of time series models.

We estimate σ^2 by

$$\hat{\sigma}_W^2 = \frac{1}{n} \sum_{j=1}^n \frac{\hat{\varepsilon}_j^2}{1 + \zeta \hat{\varepsilon}_j}.$$

A standard argument yields

$$\zeta = \frac{1}{n} \sum_{j=1}^n \frac{\hat{\varepsilon}_j}{\sigma^2} + o_p\big(n^{-1/2}\big) = \frac{1}{n} \sum_{j=1}^n \frac{\varepsilon_j}{\sigma^2} - \frac{1}{\sigma^2} \bar{h}_G^\top (\hat{\vartheta} - \vartheta) + o_p\big(n^{-1/2}\big); \tag{5.2}$$

a version of this was used by [18] in an autoregressive setting. If ε has a finite fourth moment, we obtain the expansion

$$\begin{aligned} \hat{\sigma}_W^2 &= \frac{1}{n} \sum_{j=1}^n \big[\hat{\varepsilon}_j^2 - \hat{\varepsilon}_j^3 \zeta\big] + o_p\big(n^{-1/2}\big) \\ &= \frac{1}{n} \sum_{j=1}^n \varepsilon_j^2 - \frac{E[\varepsilon^3]}{\sigma^2} \left[\frac{1}{n} \sum_{j=1}^n \varepsilon_j - \bar{h}_G^\top (\hat{\vartheta} - \vartheta) \right] + o_p\big(n^{-1/2}\big) \end{aligned}$$

by a standard argument. This shows that $\hat{\sigma}_W^2$ is an efficient estimator of σ^2 if $\hat{\vartheta}$ is an efficient estimator of ϑ which implies that $\bar{h}_G^\top \hat{\vartheta}$ has influence function $T(X, Y; G)$.

One can also show that

$$\hat{\mathbb{F}}_W(y) = \frac{1}{n} \sum_{j=1}^{n} \frac{\mathbf{1}[\hat{\varepsilon}_j \le y]}{1 + \zeta \hat{\varepsilon}_j} \tag{5.3}$$

is an efficient estimator of $F(y)$ provided $\hat{\vartheta}$ is efficient for ϑ. Here one verifies

$$\sup_{y\in\mathbb{R}} \left| \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\mathbf{1}[\hat{\varepsilon}_j \le y]}{1 + \zeta \hat{\varepsilon}_j} - \mathbf{1}[\hat{\varepsilon}_j \le y] + \zeta \hat{\varepsilon}_j \mathbf{1}[\hat{\varepsilon}_j \le y] \right) \right| = o_p\big(n^{-1/2}\big),$$

$$\sup_{y\in\mathbb{R}} \left| \frac{1}{n} \sum_{j=1}^{n} \hat{\varepsilon}_j \mathbf{1}[\hat{\varepsilon}_j \le y] - E\big[\varepsilon \mathbf{1}[\varepsilon \le y]\big] \right| = o_p(1),$$

and then uses (5.1) and (5.2) to conclude

$$\sup_{y\in\mathbb{R}} \left| \hat{\mathbb{F}}_W(y) - \frac{1}{n} \sum_{j=1}^{n} \big(\mathbf{1}[\varepsilon_j \le y] - \nu(y)\varepsilon_j\big) - \big[f(y) + \nu(y)\big] \bar{h}_G^\top (\hat{\vartheta} - \vartheta) \right| = o_p\big(n^{-1/2}\big).$$

Thus, $\hat{\mathbb{F}}_W(y)$ has the desired influence function if $\hat{\vartheta}$ is efficient.

Recall that the efficiency transfer requires that H_{G_1} is positive definite. Thus, under this assumption, complete case versions of the above estimators of σ^2 and $F(y)$ are efficient in the MAR model.

5.2. Nonparametric regression

Now consider the conventional nonparametric regression model $Y = \xi(X) + \varepsilon$, where ξ is a smooth but otherwise unknown function. This model is important for applications where the functional relationship between response and covariate cannot be predetermined, but can be approximated using data. Popular methods to carry this out involve kernel estimators, local polynomials, splines and wavelets.

Let ξ be a twice continuously differentiable function. We assume that X is *quasi-uniform* on the interval [0, 1]. This means that X has a density g that vanishes off [0, 1] and is bounded and bounded away from zero on [0, 1]. For the transfer principle to work, the distribution G_1 of $\tilde{X}$ has to be quasi-uniform on [0, 1] as well. In view of the formula $G_1(dx) = \pi(x)\, dG(x)/E[\delta]$ (cf. Section 2), it is easy to see that quasi-uniformity of G_1 is equivalent to π being bounded away from zero on the support of X. If this holds true, then B equals $L_2(G_1)$ as the twice continuously differentiable functions are dense in $L_2(H)$ for each quasi-uniform distribution H on [0, 1].

We briefly address estimation of the error distribution function and a linear functional of ξ. We first look at the case where no responses are missing, which corresponds to $\delta \equiv 1$. Then G_1 equals G and B equals $L_2(G)$. Let $\hat{\xi}$ denote a nonparametric estimator of ξ, and $\hat{\varepsilon}_j = Y - \hat{\xi}(X_j)$,

$j = 1, \dots, n$ the corresponding (nonparametric) residuals. Müller *et al.* [20] showed that the uniform expansion

$$\sup_{y\in\mathbb{R}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}[\hat{\varepsilon}_j \le y] - \mathbf{1}[\varepsilon_j \le y] - f(y)\varepsilon_j\right| = o_p\big(n^{-1/2}\big)$$

holds for an undersmoothed local linear smoother under an additional moment assumption on the errors. The Hölder condition required in their result is met, as the error density has finite Fisher information and is therefore Hölder with exponent 1/2. The regression function can alternatively be estimated by a series estimator; see [22] who estimate an additive regression function by a sum of series estimators. The same expansion was obtained by [10], who propose a weighted version that takes additional model information into account using the empirical likelihood method.

From this and Example 4, we conclude that the residual-based empirical distribution function is an efficient estimator of the error distribution function. This implies that its complete case version is efficient with missing responses as long as π is bounded away from zero on [0, 1]. This was proved by [4].

The assumption on π to be bounded away from zero on [0, 1] is crucial here because of our assumption that X is quasi-uniform on [0, 1]. We can, however, relax this assumption and require that X is quasi-uniform on some compact (unknown) interval of positive length. Then π does not have to be bounded away from zero on this interval; it suffices to require that π is bounded away from zero on a compact subinterval of positive length and be zero outside this interval. In this setting, the choices $\delta = \mathbf{1}[X \le v]$, $\delta = \mathbf{1}[X \ge u]$ and $\delta = \mathbf{1}[u \le X \le v]$ would be allowed as long as u is less than the right endpoint and v is greater than the left endpoint of the support of X. Such choices are of interest in medical applications. A treatment might only be performed if the covariate falls into a safety zone, for example.

When using local polynomial smoothers, quasi-uniformity is typically essential, but not the knowledge of the compact interval. Thus, the efficient estimator of [20] will work under the relaxed assumptions and for the choices of δ mentioned above.

When working with kernel estimators one typically requires for technical reasons, in addition to quasi-uniformity, smoothness properties for the density g on its support. Then one needs the same smoothness of πg, and this translates into smoothness assumptions on π.

Next, we look at estimating $\int w(x)\xi(x)\,dx$ for some known bounded measurable function w that vanishes outside the interval [0, 1]. The efficient influence function for this quantity is $(b_*(X) - E[b_*(X)])\ell_f(\varepsilon)/J_f + E[b_*(X)]\varepsilon$ with $b_* = w/g$. A candidate for an efficient estimator is

$$\int_0^1 \hat{b}_*(x)\hat{\xi}(x)\,dx + \frac{1}{n}\sum_{j=1}^{n}\left[\big(\hat{b}_*(X_j) - \hat{\mu}\big)\frac{\hat{\ell}(\hat{\varepsilon}_j)}{\hat{J}} + \hat{\mu}\varepsilon_j\right]$$

with $\hat{b}_* = w/\hat{g}$ for a kernel estimator of $\hat{g}$ of g and $\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n}\hat{b}_*(X_j)$. This can be verified using the work of [30] with $\hat{\xi}$ an undersmoothed local linear smoother of ξ and appropriate selection of bandwidth. Thus the complete case version will be efficient with missing observations whenever π is bounded away from zero on [0, 1]. The assumption that X is quasi-uniform on

[0, 1] can again be relaxed to X being quasi-uniform on an (unknown) compact interval containing [0, 1]. The efficiency transfer is then valid as long as π is bounded away from zero on a compact subinterval containing [0, 1] and is zero outside this interval.

5.3. Partially linear random coefficient model

Now we consider a partially linear random coefficient model

$$Y = \vartheta^\top U + S\xi(T) + \varepsilon,$$

where $\|U\|$ has a finite second moment, T is quasi-uniform on [0, 1], $E[S^2|T = t]$ is bounded and bounded away from zero for $t \in [0, 1]$, and ξ is twice continuously differentiable. For the real parameter ϑ to be identifiable, we also require that the matrix

$$H_G = E\big[\big(U - S\mu_G(T)\big)\big(U - S\mu_G(T)\big)^\top\big]$$

is positive definite, where

$$\mu_G(T) = E[SU|T]/E\big[S^2|T\big].$$

Here the covariate vector X equals $(S, T, U^\top)^\top$. We also require that π is bounded away from zero. This implies that the efficiency transfer applies. For example, the positive definiteness of H_{G_1} follows from that of H_G in view of the inequality

$$v^\top H_{G_1} v \geq \eta v^\top H_G v, \qquad v \in \mathbb{R}^p,$$

which is valid for some positive η. We can take η to be a lower bound on the density $\pi/E[\delta]$ of G_1 with respect to G. Indeed, using $E[(U - S\mu_G(T))S|T] = 0$, we calculate

$$\begin{aligned} v^\top H_{G_1} v &\geq \eta E\big(v^\top U - S v^\top \mu_{G_1}(T)\big)^2] \\ &= \eta E\big[\big(v^\top U - S v^\top \mu_G(T)\big)^2\big] + \eta E\big[S^2 v^\top \big(\mu_G(T) - \mu_{G_1}(T)\big)^2\big]. \end{aligned}$$

In the full model, we have (4.2) with G_1 replaced by G, $h(X) = U$ and $b(X) = Sb_0(T)$ for each b_0 in $L_2(G_T)$, where G_T is the distribution of T under G. This follows from the fact that the twice differentiable functions are dense in $L_2(G_T)$. The role of B is now played by $B(G) = \{b \in L_2(G) : b(X) = Sa(T), a \in L_2(G_T)\}$. The projection operator on this set is given by

$$\Pi_{B(G)} k(X) = S\frac{E[Sk(X)|T]}{E[S^2|T]}, \qquad k \in L_2(G).$$

The roles of h_* and d_K are now played by h_G and e_G where

$$h_G(X) = U - S\mu_G(T)$$

and

$$e_G(X) = E\big[h_G(X)\big]H_G^{-1}h_G(X) + S\frac{E[S|T]}{E[S^2|T]}.$$

The efficient influence function for estimating ϑ without missing responses is thus

$$(J_f H_G)^{-1}\Bigg[\big(h_G(X) - E\big[h_G(X)\big]\big)\ell_f(\varepsilon) + E\big[h_G(X)\big]\frac{\varepsilon}{\sigma^2}$$
$$+ \frac{\Delta}{1-\Delta E[e_G(X)]}E\big[h_G(X)\big]\Big(\big(e_G(X) - E\big[e_G(X)\big]\big)\ell_f(\varepsilon) + E\big[e_G(X)\big]\frac{\varepsilon}{\sigma^2}\Big)\Bigg].$$

An efficient estimator can be constructed along the lines outlined in [30]. This requires a root-n consistent estimator of ϑ and appropriate estimators of μ_G and ξ.

Next, we look at estimating $F(y)$ for some y. The efficient influence function is

$$\mathbf{1}[\varepsilon \le y] - F(y) - \nu(y)\varepsilon$$
$$+ \frac{\nu(y) + f(y)}{(1-\Delta E[e_G(X)])J_F}\Big(\big(e_G(X) - E\big[e_G(X)\big]\big)\ell_f(\varepsilon) + E\big[e_G(X)\big]\frac{\varepsilon}{\sigma^2}\Big).$$

We expect that the weighted residual-based empirical distribution function $\hat{\mathbb{F}}_W(y)$, defined as in (5.3) but with semiparametric residuals

$$\hat{\varepsilon}_j = Y_j - \hat{\vartheta}^\top U_j - S_j\hat{\xi}(T_j),$$

is efficient if $\hat{\vartheta}$ is an efficient estimator of ϑ and $\hat{\xi}$ is an appropriate estimator of ξ.

If $S = 1$, then the above model reduces to the partially linear model for which efficient estimators of ϑ are available, see [30] and [6], who propose a direct estimator of the influence function. As pointed out in [12], the complete case versions of these estimators are efficient in the missing data case as long as π is bounded away from zero. Efficient estimators of the error distribution function were obtained in [20], who use local linear smoothers to construct a residual-based empirical distribution function. Again, their complete case versions are efficient with missing responses as long as π is bounded away from zero.

6. Proof of Theorem 4.1

Using the identity

$$\int y\ell_f(y)\,dF(y) = 1$$

we see that the function $y \mapsto \ell_f(y) - y/\sigma^2$ belongs to $\mathcal{C}$. Let χ belong to K. Then $D\chi(\tilde{X}, \tilde{Y})$ is a tangent and satisfies

$$E\big[D\chi(\tilde{X}, \tilde{Y})c(\tilde{\varepsilon})\big] = 0, \qquad c \in \mathcal{C},$$

and

$$E\big[D\chi(\tilde{X},\tilde{Y})k(\tilde{X})\ell_f(\tilde{\varepsilon})\big] = J_f \int kM\chi\, dG_1, \qquad k \in K.$$

Using the formula for $M\Gamma_L$ given prior to Theorem 4.1 and the fact that $1 - d_L$ is the projection of 1 onto the orthocomplement of L, we find

$$\int kM\Gamma_L g\, dG_1 = \int kg\, dG_1 - \frac{\Delta\bar{g}}{1-\Delta\bar{d}_L}\int (k - \Pi_L k)\, dG_1, \qquad k, g \in L_2(G_1).$$

Note that the last integral is zero if k belongs to L. For $k = a^\top h + b$ in K and β in $\mathbb{R}^p$, we find

$$\begin{aligned}
\int kM\Gamma_K d_K\, dG_1 &= \int kd_K\, dG_1 = \bar{k},\\
\int kM\Gamma_K\big(\beta^\top h_*\big)\, dG_1 &= \int k\beta^\top h_*\, dG_1 = \int \big(a^\top h_*\big)\big(\beta^\top h_*\big)\, dG_1 \beta = \beta^\top H_* a,\\
\int kM\Gamma_B b_*\, dG_1 &= a^\top\alpha + \int bb_*\, dG_1.
\end{aligned}$$

Here we used the fact that d_K is the projection of 1 onto K, $a^\top h_*$ is the projection of k onto A, and the definition of α. Now let us take

$$\chi = \big[(a_* - \alpha)^\top H_*^{-1}\Gamma_K h_* + \Gamma_B b_* - \tilde{c}_*\Gamma_K d_K\big]/J_f.$$

Then we have the identity

$$\gamma_*(\tilde{X},\tilde{Y};G_1) = c_*(\tilde{\varepsilon}) + D\chi(\tilde{X},\tilde{Y}).$$

Since χ belongs to K and c_* to $\mathcal{C}$, we see that $\gamma_*(\tilde{X},\tilde{Y};G_1)$ is a tangent. Thus, it suffices to show that

$$E\big[\gamma_*(\tilde{X},\tilde{Y};G_1)v(\tilde{X},\tilde{Y})\big] = a_*^\top a + \int b_* b\, dG_1 + \int c_* c\, dF$$

holds for all tangents $v(\tilde{X},\tilde{Y})$ as in (4.1). By the above, we have

$$E\big[D\chi(\tilde{X},\tilde{Y})v(\tilde{X},\tilde{Y})\big] = (a_* - \alpha)^\top a + \alpha^\top a + \int b_* b\, dG_1 - \tilde{c}_*\big[a^\top\bar{h} + \bar{b}\big]$$

and

$$E\big[c_*(\tilde{\varepsilon})v(\tilde{X},\tilde{Y})\big] = \tilde{c}_*\big[a^\top\bar{h} + \bar{b}\big] + \int c_* c\, dF,$$

and the desired result follows.

Acknowledgements

The authors thank two referees for helpful comments which have improved the presentation of 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*. New York: Springer. MR1623559
[3] Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.* **89** 81–87.
[4] Chown, J. and Müller, U.U. (2013). Efficiently estimating the error distribution in nonparametric regression with responses missing at random. *J. Nonparametr. Stat.* **25** 665–677. MR3174290
[5] Efromovich, S. (2011). Nonparametric regression with responses missing at random. *J. Statist. Plann. Inference* **141** 3744–3752. MR2823645
[6] Forrester, J., Hooper, W., Peng, H. and Schick, A. (2003). On the construction of efficient estimators in semiparametric models. *Statist. Decisions* **21** 109–137. MR2000666
[7] González-Manteiga, W. and Pérez-González, A. (2006). Goodness-of-fit tests for linear regression models with missing response data. *Canad. J. Statist.* **34** 149–170. MR2280923
[8] Jin, K. (1992). Empirical smoothing parameter selection in adaptive estimation. *Ann. Statist.* **20** 1844–1874. MR1193315
[9] Kim, K.K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. London: Chapman & Hall/CRC.
[10] Kiwitt, S., Nagel, E.-R. and Neumeyer, N. (2008). Empirical likelihood estimators for the error distribution in nonparametric regression models. *Math. Methods Statist.* **17** 241–260. MR2448949
[11] Koul, H.L. (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. *Ann. Math. Stat.* **40** 1950–1979. MR0260126
[12] Koul, H.L., Müller, U.U. and Schick, A. (2012). The transfer principle: A tool for complete case analysis. *Ann. Statist.* **40** 3031–3049. MR3097968
[13] Li, X. (2012). Lack-of-fit testing of a regression model with response missing at random. *J. Statist. Plann. Inference* **142** 155–170. MR2827138
[14] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Hoboken, NJ: Wiley. MR1925014
[15] Müller, U.U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Statist.* **37** 2245–2277. MR2543691
[16] Müller, U.U., Peng, H. and Schick, A. (2015). Inference about the slope in linear regression with missing responses: An empirical likelihood approach. Unpublished manuscript.
[17] Müller, U.U., Schick, A. and Wefelmeyer, W. (2004). Estimating linear functionals of the error distribution in nonparametric regression. *J. Statist. Plann. Inference* **119** 75–93. MR2018451
[18] Müller, U.U., Schick, A. and Wefelmeyer, W. (2005). Weighted residual-based density estimators for nonlinear autoregressive models. *Statist. Sinica* **15** 177–195. MR2125727
[19] 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. New York: Springer. MR2230741
[20] Müller, U.U., Schick, A. and Wefelmeyer, W. (2007). Estimating the error distribution function in semiparametric regression. *Statist. Decisions* **25** 1–18. MR2370101

[21] Müller, U.U., Schick, A. and Wefelmeyer, W. (2009). Estimating the error distribution function in nonparametric regression with multivariate covariates. *Statist*. *Probab. Lett*. **79** 957–964. MR2509488
[22] Müller, U.U., Schick, A. and Wefelmeyer, W. (2012). Estimating the error distribution function in semiparametric additive regression models. *J*. *Statist*. *Plann*. *Inference* **142** 552–566. MR2843057
[23] Müller, U.U. and Van Keilegom, I. (2012). Efficient parameter estimation in regression with missing responses. *Electron*. *J*. *Stat*. **6** 1200–1219. MR2988444
[24] Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049
[25] Owen, A.B. (2001). *Empirical Likelihood*. *Monographs on Statistics and Applied Probability* **92**. London: Chapman & Hall.
[26] Robins, J.M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J*. *Amer*. *Statist*. *Assoc*. **90** 122–129. MR1325119
[27] Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974
[28] Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196
[29] Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J*. *Statist*. *Plann*. *Inference* **16** 89–105. Correction **22** (1989) 269–270.
[30] Schick, A. (1993). On efficient estimation in regression models. *Ann*. *Statist*. **21** 1486–1521. MR1241276
[31] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer. MR2233926
[32] Wang, D. and Chen, S.X. (2009). Empirical likelihood for estimating equations with missing values. *Ann*. *Statist*. **37** 490–517. MR2488360

Received January 2015 and revised September 2015