Imputing responses that are not missing

Ursula U. Müller
Universität Bremen

Anton Schick *
Binghamton University

Wolfgang Wefelmeyer
Universität Siegen

Abstract

We consider estimation of linear functionals of the joint law of regression models in which responses are missing at random. The usual approach is to work with the fully observed data, and to replace unobserved quantities by estimators of appropriate conditional expectations. Another approach is to replace all quantities by such estimators. We show that the second method is usually better than the first.

AMS 2000 subject classifications. Primary 62J02; Secondary 62N01, 62F12, 62G08, 62G20.

Key words and Phrases. Nonparametric regression, semiparametric regression, linear regression, empirical estimator, weighted empirical estimator, influence function, gradient, efficient estimator.

1. Introduction

Let (X, Y) be a random vector. We want to estimate the expectation $E[h(X, Y)]$ of some known square-integrable function h. If we are able to sample from (X, Y), we can use the empirical estimator $\frac{1}{n}\sum_{i=1}^n h(X_i, Y_i)$. If nothing is known about the distribution of (X, Y), this estimator is efficient. We are interested in the situation where we always observe X, but Y only if some indicator Z equals one. We assume that Z and Y are conditionally independent given X. Then one says that Y is *missing at random*. In this case the empirical estimator is not available unless all Z_i are one. Let $\pi(X) = E(Z \mid X) = P(Z = 1 \mid X)$. If π is known and positive, we could use the estimator $\frac{1}{n}\sum_{i=1}^n Z_i h(X_i, Y_i)/\pi(X_i)$. If π is unknown, one could replace π by an estimator $\hat{\pi}$, resulting in

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Z_i}{\hat{\pi}(X_i)}\, h(X_i, Y_i). \tag{1.1}$$

Surprisingly, even if π is known, replacing π by an estimator can decrease the asymptotic variance. Such an improvement is given by Schisterman and Rotnitzky (2001). A similar result,

*Supported in part by NSF Grant DMS 0072174

on average treatment effects, is in Hirano, Imbens and Ridder (2002). Another estimator for $E[h(X,Y)]$ is the *partially imputed estimator*

$$\frac{1}{n}\sum_{i=1}^{n}\Big(Z_i h(X_i,Y_i)+(1-Z_i)\hat{\chi}(X_i)\Big), \tag{1.2}$$

where $\hat{\chi}(X_i)$ is an estimator of the conditional expectation $\chi(X_i) = E(h(X_i,Y_i) \mid X_i)$. An alternative to the partially imputed estimator is the *fully imputed* estimator

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\chi}(X_i). \tag{1.3}$$

An extreme case would be that the conditional distribution of Y given X is *known*. It is easy to see that then the fully imputed estimator $\frac{1}{n}\sum_{i=1}^{n}\chi(X_i)$ is at least as good as the partially imputed estimator, and strictly better if and only if $Z(h(X,Y)-\chi(X))$ is not zero almost surely.

We show that the fully imputed estimator (1.3) is usually better than the partially imputed estimator (1.2). We restrict attention to the situation where π is bounded away from zero but otherwise completely unknown. We also impose no structural assumptions on the covariate distribution. We consider four different models for the conditional distribution of Y given X.

Suppose first that the conditional distribution $Q(X,dy)$ of Y given X is completely unknown. For the case $h(X,Y)=Y$, Cheng (1994) shows that the partially and fully imputed estimators are asymptotically equivalent, and obtains their asymptotic distribution. He estimates $E(Y \mid X)$ by a truncated kernel estimator. Wang and Rao (2002) obtain a similar result with a differently truncated kernel estimator. Cheng and Chu (1996) study estimation of the response distribution function and quantiles. We generalize Cheng's result to arbitrary functions h and prove efficiency.

Suppose now that we have a parametric model $Q_\vartheta(X,dy)$ for the conditional distribution of Y given X. In this case the conditional expectation is of the form $\chi_\vartheta(x) = \int h(x,y)\,Q_\vartheta(x,dy)$. This suggests estimating χ_ϑ by $\chi_{\hat{\vartheta}}$. The natural estimator for ϑ is the conditional maximum likelihood estimator. We show that the fully imputed estimator $\frac{1}{n}\sum_{i=1}^{n}\chi_{\hat{\vartheta}}(X_i)$ is efficient, and better than the corresponding partially imputed estimator except in degenerate cases. This is related to Tamhane (1978) who assumes a parametric model for the *joint* distribution of X and Y. Then $E[h(X,Y)]$ is a smooth function of ϑ; hence it can be estimated efficiently by plugging in an efficient estimator, such as the maximum likelihood estimator.

Next we consider a model between the fully nonparametric and parametric ones for Q, a linear regression model with covariates and errors independent. For simplicity we take $Y=\vartheta X+\varepsilon$. We do not assume that ε has mean zero but require X to have positive variance for identifiability. Here $Q(x,dy) = f(y-\vartheta x)\,dy$, where f is the (unknown) density of the errors. Then $\chi(x) = \int h(x,\vartheta x+u)f(u)\,du$. Exploiting this representation, we estimate $\chi(x)$ by

$\sum_{j=1}^n Z_j h(x, \hat\vartheta x + Y_j - \hat\vartheta X_j) / \sum_{j=1}^n Z_j$. We show that the corresponding fully imputed estimator is efficient if an efficient estimator for ϑ is used. Again the partially imputed estimator will not be efficient in general, even if an efficient estimator for ϑ is used.

Finally we consider a linear regression model *without* assuming independence between covariates and errors. For simplicity we take $Y = \vartheta X + \varepsilon$ with $E(\varepsilon \mid X) = 0$. This can be written as a constraint on the conditional distribution of Y given X, namely $\int y\, Q(X, dy) = \vartheta X$. For $h(X, Y) = Y$ this suggests the estimator $\hat\vartheta \frac{1}{n} \sum_{i=1}^n X_i$, which happens to be the fully imputed estimator. Matloff (1981) has shown that such an estimator improves upon the partially imputed estimator for his choice of $\hat\vartheta$. We show that the fully imputed estimator of $E[h(X,Y)]$ for general h is efficient if an appropriate estimator for χ is used. This requires an efficient estimator $\hat\vartheta$ for ϑ and a correction term to the nonparametric estimator of χ. An efficient estimator of ϑ can be obtained as a weighted least squares estimator with estimated optimal weights, based on the fully observed pairs. Efficient estimation of ϑ for more general regression models and various models for π has been studied in Robins, Rotnitzky and Zhao (1994), Robins and Rotnitzky (1995), and Rotnitzky and Robins (1995), among others. Efficient score functions for ϑ are calculated by Nan, Emond and Wellner (2004) and Yu and Nan (2003). The partially imputed estimator will not be efficient, in general. In view of this, partially imputed estimators such as the one by Wang, Härdle and Linton (2003) for $E[Y]$ in a partly linear model are not efficient.

The paper is organized as follows. In Section 2 we characterize efficient estimators for linear functionals of arbitrary regression models with responses missing at random; in particular for the four cases above. Our results show that the model is adaptive in the sense that we can estimate $E[h(X,Y)]$ as well not knowing π as knowing π. In Section 3 we construct efficient fully imputed estimators of $E[h(X,Y)]$ in these four models.

2. Efficient influence functions

In this section we calculate the efficient influence function for estimating the expected value $E[h(X,Y)]$ with observations (X, ZY, Z) as described in the Introduction. The joint distribution $P(dx, dy, dz)$ of the observations depends on the marginal distribution $G(dx)$ of X, the conditional probability $\pi(x)$ of $Z = 1$ given $X = x$, and the conditional distribution $Q(x, dy)$ of Y given $X = x$. More precisely, we have

$$P(dx, dy, dz) = G(dx) B_{\pi(x)}(dz) \big(z Q(x, dy) + (1 - z)\delta_0(dy) \big),$$

where $B_p = p\delta_1 + (1-p)\delta_0$ denotes the Bernoulli distribution with parameter p and δ_t the Dirac measure at t. Consider perturbations G_{nu}, Q_{nv} and π_{nw} of G, Q and π that are *Hellinger* *differentiable* in the following sense:

$$\int \Big(n^{1/2}\big(\sqrt{dG_{nu}} - \sqrt{dG}\big) - \frac{1}{2}u\sqrt{dG}\Big)^2 \to 0,$$
$$\iint \Big(n^{1/2}\big(\sqrt{dQ_{nv}(x,\cdot)} - \sqrt{dQ(x,\cdot)}\big) - \frac{1}{2}v(x,\cdot)\sqrt{dQ(x,\cdot)}\Big)^2 G(dx) \to 0,$$
$$\iint \Big(n^{1/2}\big(\sqrt{dB_{\pi_{nw}(x)}} - \sqrt{dB_{\pi(x)}}\big) - \frac{1}{2}(\cdot - \pi(x))w(x)\sqrt{dB_{\pi(x)}}\Big)^2 G(dx) \to 0.$$

This requires that u belongs to

$$L_{2,0}(G) = \{u \in L_2(G) : \int u\, dG = 0\};$$

that v belongs to

$$V_0 = \{v \in L_2(M) : \int v(x,y)\, Q(x,dy) = 0\}$$

with $M(dx,dy) = Q(x,dy)G(dx)$; and that w belongs to $L_2(G_\pi)$, where $G_\pi(dx) = \pi(x)(1 - \pi(x))\, G(dx)$.

We have *local asymptotic normality*: With P_{nuvw} denoting the joint distribution of the observations (X, ZY, Z) under the perturbed parameters G_{nu}, Q_{nv} and π_{nw},

$$\sum_{i=1}^n \log \frac{dP_{nuvw}}{dP}(X_i, Z_iY_i, Z_i) \quad = \quad n^{-1/2}\sum_{i=1}^n t_{uvw}(X_i, Z_iY_i, Z_i) - \frac{1}{2}E[t_{uvw}^2(X, ZY, Z)] + o_p(1),$$

where $t_{uvw}(X, ZY, Z) = u(X) + Zv(X,Y) + (Z - \pi(X))w(X)$ and

$$\begin{aligned} E[t_{uvw}^2(X, ZY, Z)] &= E[u^2(X)] + E[Zv^2(X,Y)] + E[(Z-\pi(X))^2 w^2(X)] \\ &= \int u^2\, dG + \iint \pi(x)v^2(x,y)\, Q(x,dy)G(dx) + \int w^2\, dG_\pi. \end{aligned}$$

If we have models for the parameters G, Q and π, then, in order for the perturbations G_{nu}, Q_{nv} and π_{nw} to be within these models, the functions u, v and w must be restricted to subsets U of $L_{2,0}(G)$, V of V_0, and W of $L_2(G_\pi)$. The choices $U = L_{2,0}(G)$ and $V = V_0$ correspond to fully nonparametric models for G and Q. Parametric models for G and Q result in finite-dimensional U and V. In what follows the spaces U, V and W will be assumed to be closed and linear.

Let now κ be a functional of G, Q and π. The functional is *differentiable* with *gradient* $g \in L_2(P)$ if, for all $u \in U$, $v \in V$ and $w \in W$,

$$n^{1/2}\big(\kappa(G_{nu}, Q_{nv}, \pi_{nw}) - \kappa(G, Q, \pi)\big) \to E[g(X, ZY, Z)t_{uvw}(X, ZY, Z)].$$

The gradient g is not unique. The *canonical gradient* is g_*, where $g_*(X, ZY, Z)$ is the projection of $g(X, ZY, Z)$ onto the *tangent space*

$$T = \{t_{uvw}(X, ZY, Z) : u \in U, v \in V, w \in W\}.$$

Since T is a sum of orthogonal spaces

$$\begin{aligned} T_1 &= \{u(X) : u \in U\}, \\ T_2 &= \{Zv(X,Y) : v \in V\}, \\ T_3 &= \{(Z - \pi(X))w(X) : w \in W\}, \end{aligned}$$

the random variable $g_*(X,ZY,Z)$ is the sum $u_*(X) + Zv_*(X,Y) + (Z - \pi(X))w_*(X)$, where $u_*(X)$, $Zv_*(X,Y)$ and $(Z-\pi(X))w_*(X)$ are the projections of $g(X,ZY,Z)$ onto T_1, T_2 and T_3, respectively. We assume that $E[g_*^2(X,ZY,Z)]$ is positive.

An estimator $\hat\kappa$ for κ is *regular* with *limit* L if L is a random variable such that, for all $u \in U$, $v \in V$ and $w \in W$,

$$n^{1/2}\big(\hat\kappa - \kappa(G_{nu}, Q_{nv}, \pi_{nw})\big) \Rightarrow L \quad \text{under } P_{nuvw}.$$

The Hájek–Le Cam convolution theorem says that L is distributed as the sum of a normal random variable with mean zero and variance $E[g_*^2(X,ZY,Z)]$ and some independent random variable. This justifies calling an estimator $\hat\kappa$ *efficient* if it is regular with limit such a normal random variable.

An estimator $\hat\kappa$ for κ is *asymptotically linear* with *influence function* $\psi \in L_{2,0}(P)$ if

$$n^{1/2}\big(\hat\kappa - \kappa(G,Q,\pi)\big) = n^{-1/2}\sum_{i=1}^n \psi(X_i, Z_iY_i, Z_i) + o_p(1).$$

As a consequence of the convolution theorem, a regular estimator is efficient if and only if it is asymptotically linear with influence function g_*. A reference for the convolution theorem and the characterization is Bickel, Klaassen, Ritov and Wellner (1998).

We are interested in estimating

$$\kappa(G,Q,\pi) = E[h(X,Y)] = \iint h(x,y)\,Q(x,dy)G(dx) = \int h\,dM.$$

Let $M_{nuv}(dx,dy) = Q_{nv}(x,dy)G_{nu}(dx)$. Then M_{nuv} is Hellinger differentiable in the following sense:

$$\int \Big(n^{1/2}\big(dM_{nuv}^{1/2} - dM^{1/2}\big) - \frac{1}{2}t\,dM^{1/2}\Big)^2 \to 0$$

with $t(x,y) = u(x) + v(x,y)$. If M_{nuv} satisfies $\limsup_n \int h^2\,dM_{nuv} < \infty$, then

$$n^{1/2}\Big(\int h\,dM_{nuv} - \int h\,dM\Big) \to E\big[h(X,Y)\big(u(X) + v(X,Y)\big)\big];$$

see e.g. Ibragimov and Has'minskiĭ (1981, p. 67, Lemma 7.2).

Thus the canonical gradient of $E[h(X,Y)]$ is determined by

$$\begin{aligned} &E[u_*(X)u(X)] + E[Zv_*(X,Y)v(X,Y)] + E[(Z-\pi(X))^2 w_*(X)w(X)] \\ &= E\big[h(X,Y)\big(u(X)+v(X,Y)\big)\big] \end{aligned}$$

for all $u \in U$, $v \in V$ and $w \in W$. Setting first $u = 0$ and $v = 0$, we see that $w_* = 0$. Setting $v = 0$, we see that $u_*(X)$ is the projection of $h(X,Y)$ onto T_1. Taking $u = 0$, we see that the projection of $Zv_*(X,Y)$ onto $\tilde{V} = \{v(X,Y) : v \in V\}$ must equal the projection of $h(X,Y)$ onto $\tilde{V}$.

We are mainly interested in a fully nonparametric model for G, for which $U = L_{2,0}(G)$. Then $u_*(X) = \chi(X) - E[\chi(X)]$. We now give explicit formulas for v_*, and hence for the canonical gradient of $E[h(X,Y)]$, in four cases: fully nonparametric conditional distribution, with $V = V_0$; parametric conditional distribution, with V finite-dimensional; and two semi-parametric models, namely linear regression with and without independence of covariate and error.

1. Nonparametric conditional distribution. If $V = V_0$, then the projections of $h(X,Y)$ and $Zv_*(X,Y)$ onto $\tilde{V}$ are $h(X,Y) - \chi(X)$ and $\pi(X)v_*(X,Y)$. Thus

$$v_*(X,Y) = \frac{h(X,Y) - \chi(X)}{\pi(X)}.$$

Hence, if $U = L_{2,0}(G)$, the canonical gradient of $E[h(X,Y)]$ is

$$\psi_{np}(X, ZY, Z) = \chi(X) - E[\chi(X)] + \frac{Z}{\pi(X)}\,(h(X,Y) - \chi(X)).$$

For the important special case $h(X,Y) = Y$ we obtain $\psi_{np}(X, ZY, Z) = E(Y \mid X) - E[Y] + Z(Y - E(Y \mid X))/\pi(X)$.

2. Parametric conditional distribution. Let $Q(x, dy) = q_\vartheta(x,y)\,dy$, where ϑ is an m-dimensional parameter. In this case, V will be the span of the components of the score function ℓ_ϑ, the Hellinger derivative of the parametric model q_ϑ at ϑ:

$$\iint \Big(\sqrt{q_{\vartheta+t}(x,y)} - \sqrt{q_\vartheta(x,y)} - \frac{1}{2}t^\top \ell_\vartheta(x,y)\sqrt{q_\vartheta(x,y)}\Big)^2\,dy\,G(dx) = o(t^2).$$

We also assume that $E[Z\ell_\vartheta(X,Y)\ell_\vartheta(X,Y)^\top]$ is positive definite. If q_ϑ is differentiable in ϑ, then $\ell_\vartheta = \dot{q}_\vartheta / q_\vartheta$, where $\dot{q}_\vartheta$ is the derivative of q_ϑ with respect to ϑ. If we set $L = \ell_\vartheta(X,Y)$, then $\tilde{V} = \{c^\top L : c \in \mathbb{R}^m\}$. Thus v_* is of the form $c_*^\top L$. Since the projections of $h(X,Y)$ and $Zv_*(X,Y)$ onto $\tilde{V}$ are $a^\top L$ and $b^\top L$ with $a = (E[LL^\top])^{-1}E[Lh(X,Y)]$ and $b = (E[LL^\top])^{-1}E[ZLL^\top]\,c_*$, we obtain $c_* = (E[ZLL^\top])^{-1}E[Lh(X,Y)]$. Thus, if $U = L_{2,0}(G)$, the canonical gradient of $E[h(X,Y)]$ is

$$\psi_p(X, ZY, Z) \;=\; \chi(X) - E[\chi(X)] + Zc_*^\top \ell_\vartheta(X,Y).$$

3. Linear regression with independence. We consider the linear regression model $Y = \vartheta X + \varepsilon$ with ε and X independent. We assume that ε has an unknown density f with finite Fisher information J for location and X has finite and positive variance. We do *not* assume that ε has mean zero. In this model, $Q(x, dy) = f(y - \vartheta x)\, dy$. Write F for the distribution function of f. As shown in Bickel (1982),

$$\tilde{V} = \{\alpha X \ell(\varepsilon) + \beta(\varepsilon) : \alpha \in \mathbb{R}, \beta \in L_{2,0}(F)\}.$$

Here ℓ denotes the score function $\ell(y) = -f'(y)/f(y)$ for location. The space $\tilde{V}$ can be written as the orthogonal sum of the spaces $\tilde{V}_1 = \{\alpha\xi : \alpha \in \mathbb{R}\}$ with

$$\xi = (X - E[X])\ell(\varepsilon),$$

and $\tilde{V}_2 = \{\beta(\varepsilon) : \beta \in L_{2,0}(F)\}$. The projection of $h(X,Y)$ onto $\tilde{V}_1$ is $c_h\xi/E[\xi^2]$ with $c_h = E[h(X,Y)\xi]$, and the projection of $h(X,Y)$ onto $\tilde{V}_2$ is $\bar{h}(\varepsilon) - E[\bar{h}(\varepsilon)]$ with $\bar{h}(\varepsilon) = E(h(X,Y) \mid \varepsilon)$. For $b \in L_2(F)$, the projection of $Zb(\varepsilon)$ onto $\tilde{V}_1$ is $c\xi/E[\xi^2]$ with $c = E[Zb(\varepsilon)\xi] = E[Z](E(X|Z = 1) - E[X])E[b(\varepsilon)\ell(\varepsilon)]$, and the projection of $Zb(\varepsilon)$ onto $\tilde{V}_2$ is $E[Z](b(\varepsilon) - E[b(\varepsilon)])$. Let

$$\xi_* = \big(X - E(X \mid Z = 1)\big)\ell(\varepsilon).$$

Then $Z\xi_*$ is orthogonal to $\tilde{V}_2$, and its projection onto $\tilde{V}_1$ is $a_*\xi/E[\xi^2]$ with $a_* = E[Z\xi_*\xi] = E[Z\xi_*^2]$. Since $c_h = E[h(X,Y)\xi] = E[h(X,Y)\xi_*] + (E(X|Z = 1) - E[X])E[h(X,Y)\ell(\varepsilon)]$, it follows that

$$v_*(X,Y) = \frac{E[h(X,Y)\xi_*]}{E[Z\xi_*^2]}\,\xi_* + \frac{1}{E[Z]}\big(\bar{h}(\varepsilon) - E[\bar{h}(\varepsilon)]\big).$$

Thus, if $U = L_{2,0}(G)$, the canonical gradient of $E[h(X,Y)]$ is

$$\psi_I(X, ZY, Z) = \chi(X) - E[\chi(X)] + Z\Big(\frac{E[h(X,Y)\xi_*]}{E[Z\xi_*^2]}\,\xi_* + \frac{1}{E[Z]}\big(\bar{h}(\varepsilon) - E[\bar{h}(\varepsilon)]\big)\Big).$$

For $h(X,Y) = Y$ we can use the identity $E[\varepsilon\ell(\varepsilon)] = 1$ to simplify the canonical gradient to

$$\vartheta(X - E[X]) + \frac{Z(E[X] - E(X|Z = 1))}{E[Z\xi_*^2]}\xi_* + \frac{Z(\varepsilon - E[\varepsilon])}{E[Z]}.$$

4. Linear regression without independence. Now we consider the linear regression model $Y = \vartheta X + \varepsilon$ with $E(\varepsilon \mid X) = 0$. We write $\sigma^2(X) = E(\varepsilon^2 \mid X)$ and $\rho_h(X) = E(h(X,Y)\varepsilon \mid X)$. In this model, we have only the constraint $\int y\, Q(x, dy) = \vartheta x$ on the transition distribution Q. In this case, the space $\tilde{V}$ is the sum of the two orthogonal spaces

$$\begin{aligned} \tilde{V}_1 &= \{a\sigma^{-2}(X)X\varepsilon : a \in \mathbb{R}\}, \\ \tilde{V}_2 &= \{v(X,Y) : v \in V_0, E(v(X,Y)\varepsilon \mid X) = 0\}. \end{aligned}$$

For details see Müller, Schick and Wefelmeyer (2003). The projection of $h(X,Y)$ onto $\tilde{V}_1$ is $a_h\sigma^{-2}(X)X\varepsilon$ with $a_h = E[h(X,Y)\sigma^{-2}(X)X\varepsilon]/E[\sigma^{-2}(X)X^2]$, while the projection onto

$\tilde{V}_2$ is $\tilde{h}_2 = h(X,Y) - \chi(X) - E[\rho_h(X)]\sigma^{-2}(X)\varepsilon$. It is now easy to check that $v_*(X,Y) = a_*\sigma^{-2}(X)X\varepsilon + \tilde{h}_2/\pi(X)$. Thus, if $U = L_{2,0}(G)$, the canonical gradient of $E[h(X,Y)]$ is

$$\psi_{II}(X,ZY,Z) = \chi(X) - E[\chi(X)] + \frac{Z}{\pi(X)}(h(X,Y) - \chi(X)) - \frac{Z\varepsilon}{\sigma^2(X)}\Big(\frac{\rho_h(X)}{\pi(X)} - a_*X\Big).$$

Note that $\psi_{II} = \psi_{np} - \psi^*_{II}$ with

$$\psi^*_{II}(X,ZY,Z) = \frac{Z\varepsilon}{\sigma^2(X)}\Big(\frac{\rho_h(X)}{\pi(X)} - a_*X\Big).$$

3. Efficient estimators

In this section we indicate that the fully imputed estimators are efficient in the four models discussed at the end of Section 2. Throughout we assume that we have no structural information on the covariate distribution G.

1. Nonparametric conditional distribution. In this model, Q is completely unspecified. The usual partially imputed estimators for $E[h(X,Y)]$ are of the form

$$\hat{H}_1 = \frac{1}{n}\sum_{i=1}^n \Big(Z_i h(X_i,Y_i) + (1-Z_i)\hat{\chi}(X_i)\Big),$$

where $\hat{\chi}$ is a nonparametric estimator for χ of the form

$$\hat{\chi}(X_i) = \sum_{j=1}^n W_{ij} Z_j h(X_j,Y_j)$$

with weights W_{ij} depending on $X_1,\dots,X_n,Z_1,\dots,Z_n$ only. This includes kernel-type estimators and linear smoothers. Under appropriate smoothness conditions on χ and π, and for properly chosen weights W_{ij}, the estimator $\hat{H}_1$ has the stochastic expansion

$$\hat{H}_1 = \frac{1}{n}\sum_{i=1}^n \chi(X_i) + \frac{1}{n}\sum_{i=1}^n \frac{Z_i}{\pi(X_i)}(h(X_i,Y_i) - \chi(X_i)) + o_p(n^{-1/2}). \tag{3.1}$$

In the case $h(X,Y) = Y$, such conditions are given by Cheng (1994) and Wang and Rao (2002). These authors use weights W_{ij} corresponding to truncated kernel estimators. Cheng (1994) also shows that $\hat{H}_1$ is asymptotically equivalent to the fully imputed $\hat{H}_2 = \frac{1}{n}\sum_{i=1}^n \hat{\chi}(X_i)$. It follows from (3.1) that $\hat{H}_1$ and $\hat{H}_2$ have influence function $\psi = \psi_{np}$ and are therefore efficient by Section 2.

2. Parametric conditional distribution. In this model, $Q = Q_\vartheta$, with ϑ an m-dimensional parameter. Then

$$\chi(x) = \chi_\vartheta(x) = \int h(x,y)\, Q_\vartheta(x,dy).$$

Here we use an estimator $\hat{\vartheta}$ of ϑ and obtain for $E[h(X,Y)]$ the partially and fully imputed estimators

$$\hat{H}_3 = \frac{1}{n}\sum_{i=1}^n \Big(Z_i h(X_i, Y_i) + (1 - Z_i)\chi_{\hat{\vartheta}}(X_i)\Big) \quad \text{and} \quad \hat{H}_4 = \frac{1}{n}\sum_{i=1}^n \chi_{\hat{\vartheta}}(X_i).$$

For the following discussion, we assume again Hellinger differentiability of Q_ϑ as in Section 2 and write ℓ_ϑ for the score function. A natural estimator for ϑ is the conditional maximum likelihood estimator, which solves $\frac{1}{n}\sum_{i=1}^n Z_i \ell_\vartheta(X_i, Y_i) = 0$. Under some additional regularity conditions, this estimator has the expansion

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

with $I_\vartheta = E\big[\pi(X)\ell_\vartheta(X,Y)\ell_\vartheta(X,Y)^\top\big]$. One can show that $\hat{\vartheta}$ is efficient for $\vartheta = \kappa(G, Q_\vartheta, \pi)$. Moreover, under regularity conditions, for any $n^{1/2}$-consistent $\hat{\vartheta}$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n Z_i \chi_{\hat{\vartheta}}(X_i) &= \frac{1}{n}\sum_{i=1}^n Z_i \chi_\vartheta(X_i) + D_1^\top(\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}),\\
\frac{1}{n}\sum_{i=1}^n (1 - Z_i)\chi_{\hat{\vartheta}}(X_i) &= \frac{1}{n}\sum_{i=1}^n (1 - Z_i)\chi_\vartheta(X_i) + D_0^\top(\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}),
\end{aligned}$$

where

$$D_1 = E[Zh(X,Y)\ell_\vartheta(X,Y)] \quad \text{and} \quad D_0 = E[(1 - Z)h(X,Y)\ell_\vartheta(X,Y)].$$

Thus, if we use the conditional maximum likelihood estimator for ϑ, we have the expansions

$$\begin{aligned}
\hat{H}_3 &= \frac{1}{n}\sum_{i=1}^n \Big(Z_i h(X_i, Y_i) + (1 - Z_i)\chi_\vartheta(X_i) + D_0^\top I_\vartheta^{-1} Z_i \ell_\vartheta(X_i, Y_i)\Big) + o_p(n^{-1/2}),\\
\hat{H}_4 &= \frac{1}{n}\sum_{i=1}^n \Big(\chi_\vartheta(X_i) + (D_0 + D_1)^\top I_\vartheta^{-1} Z_i \ell_\vartheta(X_i, Y_i)\Big) + o_p(n^{-1/2}).
\end{aligned}$$

Since $D_0 + D_1 = E[h(X,Y)\ell_\vartheta(X,Y)]$, we see that $\hat{H}_4$ has influence function $\psi = \psi_p$ and is therefore efficient. The difference between the estimators is

$$\hat{H}_3 - \hat{H}_4 = \frac{1}{n}\sum_{i=1}^n Z_i\Big(h(X_i, Y_i) - \chi_\vartheta(X_i) - D_1^\top I_\vartheta^{-1}\ell_\vartheta(X_i, Y_i)\Big) + o_p(n^{-1/2}).$$

Hence $\hat{H}_3$ is asymptotically equivalent to $\hat{H}_4$, and therefore also efficient, if and only if $Z\big(h(X,Y) - \chi_\vartheta(X) - D_1^\top I_\vartheta^{-1}\ell_\vartheta(X,Y)\big)$ is zero almost surely. Since this is usually not the case, the partially imputed estimator $\hat{H}_3$ is typically inefficient.

3. Linear regression with independence. In this model, $Q(x, dy) = Q_{\vartheta,f}(x, dy) = f(y - \vartheta x)\, dy$. We assume that f has finite Fisher information J for location and X has finite and positive variance. In this model,

$$\chi(x) = \chi(x, \vartheta, f) = \int h(x, \vartheta x + u) f(u)\, du.$$

This suggests the estimator

$$\hat{\chi}(x, \hat{\vartheta}) = \frac{\frac{1}{n}\sum_{j=1}^{n} Z_j h(x, \hat{\vartheta} x + Y_j - \hat{\vartheta} X_j)}{\bar{Z}},$$

where $\bar{Z} = \frac{1}{n}\sum_{j=1}^{n} Z_j$. Then the partially and fully imputed estimators for $E[h(X, Y]$ are

$$\hat{H}_5 = \frac{1}{n}\sum_{i=1}^{n}\Big(Z_i h(X_i, Y_i) + (1 - Z_i)\hat{\chi}(X_i, \hat{\vartheta})\Big) \quad \text{and} \quad \hat{H}_6 = \frac{1}{n}\sum_{i=1}^{n}\hat{\chi}(X_i, \hat{\vartheta}).$$

Let

$$S = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{Z_j}{E[Z]} h(X_i, \vartheta X_i + \varepsilon_j).$$

Then $E[S] = E[h(X, Y)] = \kappa$. By the Hoeffding decomposition,

$$S = \kappa + \frac{1}{n}\sum_{i=1}^{n}(\chi(X_i) - \kappa) + \frac{1}{n}\sum_{j=1}^{n}\Big(\frac{Z_j \bar{h}(\varepsilon_j)}{E[Z]} - \kappa\Big)$$

with $\bar{h}(\varepsilon) = E(h(X, Y) \mid \varepsilon)$. Using this we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\hat{\chi}(X_i, \vartheta) &= \frac{E[Z]}{\bar{Z}} S \;=\; S - \frac{\bar{Z} - E[Z]}{E[Z]}\kappa + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\chi(X_i) + \frac{1}{n}\sum_{j=1}^{n}\frac{Z_j}{E[Z]}(\bar{h}(\varepsilon_j) - \kappa) + o_p(n^{-1/2}).
\end{aligned}$$

Under additional assumptions,

$$\begin{aligned}
\hat{H}_6 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{Z_j}{\bar{Z}} h\big(X_i, \vartheta X_i + \varepsilon_j + (\hat{\vartheta} - \vartheta)(X_i - X_j)\big) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{Z_j}{\bar{Z}} h(X_i, \vartheta X_i + \varepsilon_j) + D(\hat{\vartheta} - \vartheta) + o_p(n^{-1/2})
\end{aligned}$$

with

$$D = \frac{1}{E[Z]} E\big[h(X_1, X_1 + \varepsilon_2) Z_2 (X_1 - X_2)\ell(\varepsilon_2)\big] = E\big[h(X, Y)(X - E(X|Z = 1))\ell(\varepsilon)\big].$$

In the linear regression model *without* missing responses, efficient estimators for ϑ have been constructed by Bickel (1982), Koul and Susarla (1983), and Schick (1987, 1993). Their influence function is $\xi/E[\xi^2]$ with $\xi = (X - E[X])\ell(\varepsilon)$. An analogous construction based on the observations (X_i, Y_i) with $Z_i = 1$ yields an estimator for ϑ with influence function $Z\xi_*/E[Z\xi_*^2]$ with $\xi_* = \big(X - E(X \mid Z = 1)\big)\,\ell(\varepsilon)$. One can show that $\hat{\vartheta}$ is efficient for $\vartheta = \kappa(G, Q_{\vartheta,f}, \pi)$. If we use an estimator $\hat{\vartheta}$ with this influence function, then $\hat{H}_6$ has the stochastic expansion

$$\hat{H}_6 = \frac{1}{n}\sum_{i=1}^{n}\Big(\chi(X_i) + \frac{Z_i}{E[Z]}(\bar{h}(\varepsilon_i) - \kappa) + \frac{D}{E[Z\xi_*^2]}\,Z_i\big(X_i - E(X \mid Z = 1)\big)\,\ell(\varepsilon_i)\Big) + o_p(n^{-1/2}).$$

Thus this estimator has influence function $\psi = \psi_I$ and is therefore efficient by Section 2. Note that in general the partially imputed estimator $\hat{H}_5$ is different from $\hat{H}_6$ and therefore inefficient. If $h(X,Y) = Y$, our estimator becomes $\hat{\vartheta}\bar{X} + \frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - \hat{\vartheta}X_i)/\bar{Z}$.

4. Linear regression without independence. In this model, Q satisfies the constraint $\int y\,Q(x,dy) = \vartheta x$. We estimate ϑ by a weighted least squares estimator based on (X_i, Y_i) with $Z_i = 1$,

$$\hat{\vartheta} = \frac{\sum_{i=1}^{n} Z_i\hat{\sigma}^{-2}(X_i)X_iY_i}{\sum_{i=1}^{n} Z_i\hat{\sigma}^{-2}(X_i)X_i^2},$$

with $\hat{\sigma}^2(x)$ an estimator of $\sigma^2(x) = E(\varepsilon^2 \mid X = x)$. Such estimators have been studied without missing responses by Carroll (1982), Müller and Stadtmüller (1987), Robinson (1987), and Schick (1987). In view of their results, we get under appropriate conditions that

$$\hat{\vartheta} = \vartheta + \frac{\frac{1}{n}\sum_{i=1}^{n} Z_i\sigma^{-2}(X_i)X_i\varepsilon_i}{E[Z\sigma^{-2}(X)X^2]} + o_p(n^{-1/2}).$$

This estimator can be shown to be efficient for ϑ.

A possible estimator for χ is the nonparametric estimator $\hat{\chi}$ introduced above for the nonparametric model. Here, however, we have the constraint $\int y\,Q(x,dy) = \vartheta x$ and use the estimator

$$\hat{\chi}_{II}(X_i) = \sum W_{ij}Z_jh(X_j, Y_j) - \hat{c}$$

with

$$\hat{c} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i\hat{\rho}_h(X_i)}{\hat{\pi}(X_i)\hat{\sigma}^2(X_i)}(Y_i - \hat{\vartheta}X_i),$$

where $\hat{\pi}(x)$ and $\hat{\rho}_h(x)$ are nonparametric estimators of $\pi(x)$ and $\rho_h(x) = E(h(X,Y)\varepsilon \mid X = x)$. Note that $\hat{c}$ is of order $n^{-1/2}$. Hence $\hat{\chi}_{II}(x)$ is asymptotically equivalent to the nonparametric estimator $\hat{\chi}$. Nevertheless, it leads to a better estimator for $E[h(X,Y)]$. Under appropriate assumptions, $\hat{c}$ has the expansion

$$\hat{c} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i\rho_h(X_i)}{\pi(X_i)\sigma^2(X_i)}\,\varepsilon_i - d(\hat{\vartheta} - \vartheta) + o_p(n^{-1/2})$$

with $d = E[Z\rho_h(X)X/\pi(X)\sigma^2(X)] = E[h(X,Y)\sigma^{-2}(X)X\varepsilon]$. Using the expansion for the weighted least squares estimator $\hat{\vartheta}$, we see that

$$\begin{aligned}\hat{c} &= \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i\varepsilon_i}{\sigma^2(X_i)}\Big(\frac{\rho_h(X_i)}{\pi(X_i)} - \frac{dX_i}{E[Z\sigma^{-2}(X)X^2]}\Big) + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\psi_{II}^*(X_i, Z_iY_i, Z_i) + o_p(n^{-1/2}).\end{aligned}$$

Using this and the stochastic expansion of the nonparametric estimator $\hat{\chi}$, we obtain that the estimators $\hat{H}_1 - \hat{c}$ and $\hat{H}_2 - \hat{c}$ have influence functions $\psi = \psi_{II}$ and are therefore efficient by Section 2. Of course, $\hat{H}_2 - \hat{c}$ is the fully imputed estimator based on $\hat{\chi}_{II}$. Both $\hat{H}_1 - \hat{c}$ and $\hat{H}_2 - \hat{c}$ are better than the partially imputed estimators $\hat{H}_1$ based on the estimator $\hat{\chi}$, and $\hat{H}_1 - (1-\bar{Z})\hat{c}$ based on the estimator $\hat{\chi}_{II}$.

Simpler estimators are possible for certain functions h, such as $h(x,y) = y$, which is usually treated in the literature. Since $E(Y \mid X) = \vartheta X$, we can use the fully imputed estimator $\hat{\vartheta}\bar{X}$, with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. As smooth function of the two efficient estimators $\hat{\vartheta}$ and $\bar{X}$, the estimator $\hat{\vartheta}\bar{X}$ is efficient for $E(Y \mid X)$. Matloff (1981) has recommended an estimator of this form, but with a simpler, in general inefficient, estimator for ϑ.

References

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* 10, 647–671.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer, New York.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* 10, 1224–1233.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.* 89, 81–87.

Cheng, P. E. and Chu, C. K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statist. Sinica* 6, 63–78.

Hirano, K., Imbens, G. W. and Ridder, G. (2002). Efficient estimation of average treatment effects using the estimated propensity score. Technical Report, Department of Economics, University of California, Berkeley. http://emlab.berkeley.edu/users/imbens/.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical Estimation. Asymptotic Theory.* Applications of Mathematics 16, Springer, New York.

Koul, H. L. and Susarla, V. (1983). Adaptive estimation in linear regression. *Statist. Decisions* 1, 379–400.

Matloff, N. S. (1981). Use of regression functions for improved estimation of means. *Biometrika* 68, 685–689.

Müller, H.-G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* 15, 610–625.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2003). Estimating functionals of the error distribution in parametric and nonparametric regression. To appear in: *J. Nonparametr. Statist.*

Nan, B., Emond, M. and Wellner, J. A. (2004). Information bounds for Cox regression models with missing data. To appear in: *Ann. Statist.* 32.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* 90, 122–129.

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.

Robinson, P. M. (1987) Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* 55, 875–891.

Rotnitzky, A. and Robins, J. M. (1995). Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scand. J. Statist.* 22, 323–333.

Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference* 16, 89–105.

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

Schisterman, E. and Rotnitzky, A. (2001). Estimation of the mean of a K-sample U-statistic with missing outcomes and auxiliaries. *Biometrika* 88, 713–725.

Tamhane, A. C. (1978). Inference based on regression estimator in double sampling. *Biometrika* 65, 419–427.

Wang, Q., Härdle, W. and Linton, O. (2003). Semiparametric regression analysis under imputation for missing response data. Technical Report, Department of Economics, London School of Economics. http://econ.lse.ac.uk/staff/olinton/research/.

Wang, Q. and Rao, J. N. K. (2001). Empirical likelihood for linear regression models under imputation for missing responses. *Canad. J. Statist.* 29, 597–608.

Wang, Q. and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.* 30, 896–924.

Yu, M. and Nan, B. (2003). Semiparametric regression models with missing data: the mathematics in the work of Robins et al. Technical Report, Department of Biostatistics, University of Michigan. http://www.sph.umich.edu/~bnan/.