RESEARCH ARTICLE

Efficiently estimating the error distribution in nonparametric regression with responses missing at random

Justin Chown[a*] and Ursula U. Müller[b*]

[ab] *Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA*

(*Received ; final version received*)

This article considers nonparametric regression models with multivariate covariates and with responses missing at random. We estimate the regression function with a local polynomial smoother. The residual-based empirical distribution function that only uses complete cases, i.e. residuals that can actually be constructed from the data, is shown to be efficient in the sense of Hájek and Le Cam. In the proofs we derive, more generally, the efficient influence function for estimating an arbitrary linear functional of the error distribution; this covers the distribution function as a special case. We also show that the complete case residual-based empirical distribution function admits a functional central limit theorem. This is done by applying the transfer principle for complete case statistics developed by Koul et al. (2012), which makes it possible to adapt known results for fully observed data to the missing data case. The article concludes with a small simulation study investigating the performance of the complete case residual-based empirical distribution function.

Keywords: nonparametric regression, local polynomial smoother, transfer principle, efficient estimator, empirical distribution function, missing at random, test for normal errors, martingale transform

2010 AMS Subject Classifications: Primary: 62G05; Secondary: 62G08, 62G20

*Correspondence should be addressed to Justin Chown (email: jchown@stat.tamu.edu) or to Ursula U. Müller (email: uschi@stat.tamu.edu)

1. Introduction and main result

An important tool for making decisions about goodness-of-fit and lack-of-fit is the residual-based empirical distribution function. This has been studied in many articles. Stute (1997) and Khmaladze and Koul (2004, 2009), for example, test parametric hypotheses about the regression function in nonparametric models. Neumeyer and Van Keilegom (2010) study additivity tests in heteroskedastic nonparametric regression. Müller, Schick and Wefelmeyer (2012) test for normal errors.

In this article we study the nonparametric regression model

$$Y = r(X) + \varepsilon,$$

with the error ε independent of the covariate vector X. Nonparametric models are particularly useful for residual-based inference because residuals constructed from them are usually consistent. We are interested in the case where responses Y are missing, i.e. we observe the sample $(X_1, \delta_1 Y_1, \delta_1), \ldots, (X_n, \delta_n Y_n, \delta_n)$, where δ is an indicator variable which equals one if Y is observed and zero otherwise. In practical applications most datasets contain missing responses, so it is important to choose statistical methods that ensure conclusions are not biased. We make the assumption that responses are *missing at random* (MAR). This means that the probability that Y is observed depends only on the covariates,

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

We will refer to the model with responses missing at random as the *MAR model*. MAR is a common assumption and is reasonable in many situations (see Little & Rubin, 2002, Chapter 1). As an example, consider missing responses to a survey question about income. If additional data (X) about medical conditions were available, we might see that the response probabilities (π) are smaller for subjects diagnosed with depression. In this case the missing mechanism is ignorable since π depends only on fully observed data X, i.e. it can be estimated from the data. More examples of missing data can be found in Tsiatis (2006), in Liang, Wang and Carroll (2007), in Molenberghs and Kenward (2007), and in Efromovich (2011a, 2011b).

We show in this article that the residual-based empirical distribution function $\hat{\mathbb{F}}_c$ given in equation (1.2) below is an efficient estimator of the unknown error distribution function F. This estimator uses only the complete data pairs (X, Y), i.e. the available residuals $\hat{\varepsilon}_{j,c} = Y_j - \hat{r}_c(X_j)$, where $\hat{r}_c$ is a suitable complete case estimator of the regression function. Demonstrating this requires two steps. First we show that $\hat{\mathbb{F}}_c$ satisfies the uniform stochastic expansion

$$\sup_{t\in\mathbb{R}} \left| \hat{\mathbb{F}}_c(t) - \frac{1}{N}\sum_{j=1}^{n} \delta_j \mathbf{1}(\varepsilon_j \le t) - f(t)\frac{1}{N}\sum_{j=1}^{n} \delta_j \varepsilon_j \right| = o_p(n^{-1/2}). \tag{1.1}$$

Here f is the error density and $N = \sum_{j=1}^{n} \delta_j$ is the number of complete cases. Then we show that an estimator of F that admits this expansion is asymptotically efficient in the sense of Hájek and Le Cam. In Section 2 we derive, more generally, the efficient influence function for estimating an arbitrary linear functional $E\{h(\varepsilon)\}$. This covers $F(t) = E\{1(\varepsilon \le t)\}$ as a special case. We conclude that an estimator $\hat{\mathbb{F}}_c$ with expansion

(1.1) is indeed efficient for F.

The first part can be dealt with easily using the *transfer principle* for complete case statistics in Koul, Schick and Müller (2012). This principle makes it possible to adapt results for the model where all data are fully observed, the *full model*, to missing data models. In particular, we can use the complete case version $\hat{r}_c$ of the estimator $\hat{r}$ proposed by Müller, Schick and Wefelmeyer (2009). They obtain expansion (1.1) for the full model (with all indicators equal to one) using a local polynomial smoother to estimate r. See also Neumeyer and Van Keilegom (2010), who consider heteroskedastic nonparametric regression.

In order to summarize the main result by Müller et al. (2009) (Theorem 1.1 below) we introduce some notation. Let $i = (i_1, \dots, i_m)$ be a multi-index and write $I(k)$ for the set of multi-indices that satisfy $i_1 + \cdots + i_m \le k$. Müller et al. (2009) estimate r by a local polynomial smoother $\hat{r}$ of degree d. It is defined as the component $\hat{\beta}_0$ corresponding to the multi-index $0 = (0, \dots, 0)$ of a minimizer

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta=(\beta_i)_{i\in I(d)}} \sum_{j=1}^{n} \Big\{ Y_j - \sum_{i\in I(d)} \beta_i \psi_i\Big(\frac{X_j - x}{c_n}\Big) \Big\}^2 w\Big(\frac{X_j - x}{c_n}\Big),$$

where

$$\psi_i(x) = \frac{x_1^{i_1}}{i_1!} \cdots \frac{x_m^{i_m}}{i_m!}, \quad x = (x_1, \dots, x_m) \in \mathbb{R}^m,$$

$w(x) = w_1(x_1) \cdots w_m(x_m)$ is a product of densities, and c_n is a bandwidth.

The estimator $\hat{r}$ permits the desired expansion if the assumptions of Theorem 1.1 (below) are satisfied. This requires, in particular, that the regression function r belongs to the Hölder space $H(d, \gamma)$, i.e. it has continuous partial derivatives of order d (or higher), and that the partial derivatives of order d are Hölder with exponent γ. The choice of the degree d of the local polynomial smoother will also depend on smoothness and moment conditions on the error density, and on the dimension of the covariate vector. In our simulation study in Section 3 we consider an infinitely differentiable regression function r and a one-dimensional covariate X. This allows us to use a locally linear smoother.

Theorem 1 from Müller et al. (2009) is proved under the following assumption on the covariate distribution.

Assumption (G) The covariate vector X is quasi-uniform on the cube $[0,1]^m$, i.e. X has a density which is bounded and bounded away from zero on $[0,1]^m$.

Theorem 1.1 : (Müller, Schick and Wefelmeyer, 2009, Theorem 1)
Let assumption (G) be satisfied. Suppose that the regression function r belongs to $H(d, \gamma)$ with $s = d + \gamma > 3m/2$. Suppose further that the error variable has mean zero, a finite moment of order $\zeta > 4s/(2s - m)$ and a density f that is Hölder with exponent $\xi > m/(2s-m)$. Consider the estimator $\hat{r}$ from above with densities $w_1, \dots, w_m$ that are $(m+2)$-times continuously differentiable and have compact support $[-1, 1]$. Let the bandwidth satisfy $c_n \sim (n \log n)^{-1/(2s)}$. Then, with $\hat{\varepsilon}_j = Y_j - \hat{r}(X_j)$,

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n} \sum_{j=1}^{n} \big\{ \mathbf{1}(\hat{\varepsilon}_j \le t) - \mathbf{1}(\varepsilon_j \le t) - \varepsilon_j f(t) \big\} \Big| = o_p(n^{-1/2}).$$

We now apply the transfer principle for asymptotically linear statistics given by Koul et al. (2012) to adapt the results from Theorem 1.1 for the MAR model. The complete case estimator for $F(t)$ is given by

$$\hat{\mathbb{F}}_c(t) = \frac{1}{N}\sum_{j=1}^{n} \delta_j \mathbf{1}(\hat{\varepsilon}_{j,c} \le t) = \frac{1}{N}\sum_{j=1}^{n} \delta_j \mathbf{1}\{Y_j - \hat{r}_c(X_j) \le t\}, \tag{1.2}$$

where $\hat{r}_c$ is the complete case version of $\hat{r}$, i.e. the component $\hat{\beta}_{c0}$ of a minimizer

$$\hat{\beta}_c = \operatorname*{arg\,min}_{\beta=(\beta_i)_{i\in I(d)}} \sum_{j=1}^{n} \delta_j \Big\{ Y_j - \sum_{i\in I(d)} \beta_i \psi_i\Big(\frac{X_j - x}{c_n}\Big)\Big\}^2 w\Big(\frac{X_j - x}{c_n}\Big). \tag{1.3}$$

Using the transfer principle requires the conditional distribution of (X, Y) given $\delta = 1$ to meet the assumptions on the (unconditional) joint distribution of (X, Y) from Theorem 1.1. In our case it is easy to see that this affects only the covariate distribution G: the MAR assumption combined with the independence of X and ε yield that ε and (X, δ) are independent. Hence the parameters f and r stay the same when switching from the unconditional to the conditional distribution. In particular, the complete case statistic $\hat{\mathbb{F}}_c(t)$ is a consistent estimator for $F(t)$ in the MAR model (since F remains unchanged). Hence we can keep all but one of our assumptions: only assumption (G) must be restated.

Assumption ($\mathbf{G}_1$) The conditional distribution of the covariate vector X given $\delta = 1$ is quasi-uniform on the cube $[0, 1]^m$, i.e. it has a density which is bounded and bounded away from zero on $[0, 1]^m$.

The transfer principle says that the complete case version of the estimator from Theorem 1.1 has the corresponding expansion (1.1). This expansion is equivalent to

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j}{E\delta} \big\{ \mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t) - \varepsilon_j f(t) \big\} \Big| = o_p(n^{-1/2}).$$

Hence we have, uniformly in $t \in \mathbb{R}$,

$$\hat{\mathbb{F}}_c(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j}{E\delta} \mathbf{1}(\hat{\varepsilon}_{j,c} \le t) + o_p(n^{-1/2}) = F(t) + \frac{1}{n}\sum_{j=1}^{n} b(\delta_j, \varepsilon_j, t) + o_p(n^{-1/2}),$$

with influence function $b(\delta, \varepsilon, t) = \delta / E\delta \, \{\mathbf{1}(\varepsilon \le t) - F(t) + f(t)\varepsilon\}$. This is indeed the *efficient* influence function for estimating $F(t)$: see Corollary 2.3 in Section 2. This brings us to the main result of this paper.

Theorem 1.2 : *Consider the nonparametric regression model with responses missing at random. Suppose the assumptions of Theorem 1.1 are satisfied, now with (G_1) in place of (G). Then the* complete case *estimator $\hat{\mathbb{F}}_c$ of the error distribution satisfies the stochastic expansion (1.1), i.e.*

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{N}\sum_{j=1}^{n} \delta_j \big\{ \mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t) - \varepsilon_j f(t) \big\} \Big| = o_p(n^{-1/2}).$$

If the error density furthermore fulfills assumption (F), stated in section 2, then $\hat{\mathbb{F}}_c(t)$ *is asymptotically efficient in the sense of Hájek and Le Cam for estimating* $F(t)$, $t \in \mathbb{R}$, *with influence function*

$$b(\delta, \varepsilon, t) = \frac{\delta}{E\delta}\big\{\mathbf{1}(\varepsilon \le t) - F(t) + f(t)\varepsilon\big\}.$$

Remark 1: If the transfer principle were not available, the expansion in Theorem 1.2 could be derived by mimicking the (rather elaborate) proofs of Lemma 1 in Müller et al. (2009) and of Theorem 2.2 in Müller, Schick and Wefelmeyer (2007), who estimate the error distribution in a general semiparametric regression model. The arguments are essentially the same – what is new now is the presence of indicators. The approach is as follows. Analogously to Müller et al. (2009; see equation (1.4) in that paper), one derives an approximation $\hat{a}_c(x)$ of the difference $\hat{r}_c(x) - r(x)$,

$$\sup_{x\in\mathbb{R}} |\hat{r}_c(x) - r(x) - \hat{a}_c(x)| = o_p(n^{-1/2}). \tag{1.4}$$

Note that the statements $\hat{\varepsilon}_{j,c} \le t$ and $\varepsilon_j \le t + \hat{r}_c(x) - r(x)$ are equivalent. Now use this and (1.4) and replace the two empirical distribution functions $\hat{\mathbb{F}}_c$ and $N^{-1}\sum_{j=1}^n \delta_j \mathbf{1}(\varepsilon_j \le t)$ in the formula by expectations (cf. Müller et al., 2007, proof of Theorem 2.2). This gives

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n}\sum_{j=1}^n \frac{\delta_j}{E\delta}\{\mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t)\} - \{F_{\hat{a}_c}(t) - F(t)\}\Big| = o_p(n^{-1/2}),$$

where

$$F_a(t) = E\Big[\frac{\delta_j}{E\delta}\mathbf{1}\{\varepsilon \le t + a(X)\}\Big] = E[\mathbf{1}\{\varepsilon \le t + a(X)\}|\delta = 1] = \int F\{t + a(x)\}\, G_1(dx)$$

with G_1 denoting the conditional distribution of X given $\delta = 1$; $F(t)$ is the expectation of the second term of the sum, i.e. $F(t) = F_a(t)$ for $a = 0$. A Taylor expansion applied to $F_{\hat{a}_c}(t) - F(t)$ in the above expansion yields

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n}\sum_{j=1}^n \frac{\delta_j}{E\delta}\big\{\mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t)\big\} - f(t)\int \hat{a}_c(x) G_1(dx)\Big| = o_p(n^{-1/2}).$$

The desired expansion now follows from this combined with

$$\int \hat{a}_c(x) G_1(dx) = \frac{1}{n}\sum_{j=1}^n \frac{\delta_j}{E\delta}\varepsilon_j + o_p(n^{-1/2}).$$

The last approximation is the complete case version of equation (1.3) in Müller et al. (2009). It can be verified by inspecting the proof of Lemma 1 in that paper, where properties of locally polynomial smoothers are derived. Keep in mind that our estimators are constructed from the complete cases (equation (1.3) above), which explains the indicators in the above formula.

Note that the uniform expansion implies a functional central limit theorem. Also note that the efficiency property of our proposed simple estimator $\hat{\mathbb{F}}_c$ yields that competing imputation type estimators will not be able to outperform it in large samples. In Section 3 we illustrate this result with simulations for two examples. The first example demonstrates the efficiency of the complete case estimator $\hat{\mathbb{F}}_c$ by comparing it with a 'tuned' estimator using an imputation technique similar to one studied by González-Manteiga and Pérez-González (2006). For our second example, we perform simulations similar to those in Müller, Schick and Wefelmeyer (2012), who use a martingale transform approach to test for normal errors in the full model. The test statistics involve the estimators from the first example.

2. Efficiency

We now calculate the *efficient influence function* for estimating the functional $E\{h(\varepsilon)\}$ using observations $(X_i, \delta_i Y_i, \delta_i)$, $i = 1, \ldots, n$. We first follow the arguments of Müller, Schick and Wefelmeyer (2006), who study efficient estimation of general differentiable functionals with data of the above form. We summarize their main arguments and refer to that paper for more details. We then focus on the functional $E\{h(\varepsilon)\}$, which Müller, Schick and Wefelmeyer (2004) study in the full model. This allows us to adapt parts of their proofs to the MAR model considered here.

We do not assume a parametric model for the regression function or for the distribution of the observations. The parameter set Θ of the statistical model therefore includes a family of covariate distributions $\mathscr{G}$, a family of error distributions $\mathscr{F}$, a set of regression functions $\mathscr{R}$, and a family of response probability distributions $\mathscr{B}$. i.e. $\Theta = \mathscr{G} \times \mathscr{F} \times \mathscr{R} \times \mathscr{B}$. We impose the following assumptions:

Assumption (F) The error density f is absolutely continuous with almost everywhere derivative f' and finite Fisher information $J = \int \ell^2(z) f(z) dz$, where $\ell = -f'/f$ denotes the score function.

Since the construction of the efficient influence function utilizes the directional information in Θ, we will now identify the set $\dot{\Theta}$ of all perturbations related to the statistical model, which may be thought of as directions. The joint distribution $P(dx, dy, dz)$ depends on the marginal distribution $G(dx)$ of X, the conditional probability $\pi(x)$ that δ equals one given $X = x$, and the conditional distribution $Q(x, dy)$ of Y given $X = x$. Formally we have

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

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. Now consider perturbations G_{nu}, π_{nw} and Q_{nv} of G, π and Q,

respectively, that are *Hellinger differentiable* in the following sense:

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

$$\int\int \Big[n^{1/2}\big\{ dB^{1/2}_{\pi_{nw}(x)} - dB^{1/2}_{\pi(x)} \big\} - \frac{1}{2}\{\cdot - \pi(x)\} w(x) dB^{1/2}_{\pi(x)} \Big]^2 G(dx) \to 0,$$

$$\int\int \Big[(n^{1/2}\big\{ dQ^{1/2}_{nv}(x,\cdot) - dQ^{1/2}(x,\cdot) \big\} - \frac{1}{2} v(x,\cdot) dQ^{1/2}(x,\cdot) \Big]^2 G_1(dx) \to 0,$$

with G_1 as the conditional distribution of X given that $\delta = 1$. This requires that u belongs to $\mathcal{L}_{2,0}(G)$, i.e. $u \in \mathcal{L}_2(G)$ and $\int u\, dG = 0$, that w belongs to

$$\mathcal{L}_2(G_\pi) = \Big\{ w \in \mathcal{L}_2(G) : \int w^2(x)\pi(x)\{1-\pi(x)\} dG(x) < \infty \Big\}$$

with $G_\pi(dx) = \pi(x)\{1-\pi(x)\}G(dx)$, and that v belongs to

$$\mathcal{V}_0 = \Big\{ v \in \mathcal{L}_2(Q \otimes G_1) : \int v(x,y) dQ(x,dy) = 0 \Big\}.$$

Note that models for G_1, π and Q will imply further restrictions on the perturbations in order to satisfy those model assumptions. So u, w and v must be restricted to subspaces of $\mathcal{L}_{2,0}(G)$, $\mathcal{L}_2(G_\pi)$ and $\mathcal{V}_0$, respectively. In this paper no model assumptions on G and π have been made, so we only have to identify the appropriate subspace $\mathcal{V}$ of $\mathcal{V}_0$. Since the covariates and the errors are assumed to be independent, we may write $Q(x,dy) = f\{y - r(x)\}dy$. With this notation the constraint on $v \in \mathcal{V}_0$ states $\int v(x,y) f\{y - r(x)\} dy = 0$. In order to derive the explicit form of $\mathcal{V}$ we introduce perturbations s and t of the unknown functions f and r and write

$$Q_{nv}(x,dy) = Q_{nst}(x,dy) = f_{ns}(y - r_{nt})dy,$$

where $f_{ns}(z) = f(z)\{1 + n^{-1/2}s(z)\}$, $r_{nt}(x) = r(x) + n^{-1/2}t(x)$ for $s \in \mathcal{S}$ and $t \in \mathcal{T}$. Here

$$\mathcal{S} = \Big\{ s \in \mathcal{L}_2(F) : \int s(z) f(z) dz = 0, \int z s(z) f(z) dz = 0 \Big\},$$

which comes from the requirement that the perturbed density f_{ns} integrates to one and has mean zero. We can take $\mathcal{T} = \mathcal{L}_2(G_1)$ since we do not assume a parametric form for r. In the following we write "$\doteq$" to denote asymptotic equivalence, i.e. equality up to an additive term of order $o_p(n^{-1/2})$. As in Müller (2009), who considers a *parametric* (nonlinear) regression function, we have

$$\begin{aligned} f_{ns}(y - r_{nt}(x)) &= f\{y - r_{nt}(x)\}\big[1 + n^{-1/2} s\{y - r_{nt}(x)\}\big] \\ &= f\{y - r(x) - n^{-1/2}t(x)\}\big[1 + n^{-1/2} s\{y - r(x) - n^{-1/2}t(x)\}\big] \\ &\doteq f\{y - r(x)\}\Big(1 + n^{-1/2}\big[s\{y - r(x)\} + \ell\{y - r(x)\}t(x)\big]\Big). \end{aligned}$$

Hence $Q_{nst}(x, dy) \doteq f\{y - r(x)\}\Big(1 + n^{-1/2}\big[s\{y - r(x)\} + \ell\{y - r(x)\}t(x)\big]\Big)$ and $\mathcal{V}$ has the form

$$\mathcal{V} = \Big\{v(x, y) = s\{y - r(x)\} + \ell\{y - r(x)\}t(x) : s \in \mathcal{S}, t \in \mathcal{T}\Big\}.$$

Thus we construct $\dot{\Theta}$ as the set containing all possible Hellinger perturbations of the statistical model parameters, or just $\dot{\Theta} = \mathcal{L}_{2,0}(G) \times \mathcal{S} \times \mathcal{L}_2(G_1) \times \mathcal{L}_2(G_\pi)$. The perturbed distribution $P_{n\gamma}$, with $\gamma = (u, s, t, w)$ in $\dot{\Theta}$, of the observation $(X, \delta Y, \delta)$ is then

$$P_{n\gamma}(dx, dy, dz) \doteq G_{nu}(dx)B_{\pi_{nw(x)}}(dz)\left\{zQ_{nst}(x, dy) + (1 - z)\delta_0(dy)\right\}.$$

It follows that $P_{n\gamma}$ is Hellinger differentiable with tangent

$$d_\gamma(X, \delta Y, \delta) = u(X) + \delta\{s(\varepsilon) + \ell(\varepsilon)t(X)\} + \{\delta - \pi(X)\}w(X).$$

The efficient influence function of a differentiable functional is characterized by its canonical gradient, which is defined as an orthogonal projection of a gradient onto the tangent space. We take the tangent space T as the closure of the linear subspace formed by d_γ. Since d_γ is a sum of orthogonal elements we can write

$$T = \{u(X) : u \in \mathcal{L}_{2,0}(G)\} \oplus \mathcal{V} \oplus \{(\delta - \pi(X))w(X) : w \in \mathcal{L}_2(G_\pi)\}.$$

We are interested in the linear functional $E\{h(\varepsilon)\}$. In order to specify a gradient of $E\{h(\varepsilon)\}$ we need the directional derivative $\gamma_h \in \dot{\Theta}$ of $E\{h(\varepsilon)\}$, which is characterized by a limit as follows. As in Müller et al. (2004) we have, for every $s \in \mathcal{S}$,

$$\lim_{n\to\infty} n^{1/2}\Big[\int h(z)f_{ns}(z)\,dz - E\{h(\varepsilon)\}\Big] = E\{h(\varepsilon)s(\varepsilon)\} = E\{h_0(\varepsilon)s(\varepsilon)\},$$

with h_0 given as the projection of h onto $\mathcal{S}$,

$$h_0(z) \;=\; h(z) - \int h\,dF - \frac{z}{\sigma^2}\int xh(x)\,dF(x),$$

where σ^2 denotes the error variance. Hence $E\{h(\varepsilon)\}$ is differentiable with directional derivative $\gamma_h = (0, h_0, 0, 0)$ and gradient $h_0(\varepsilon)$. By the convolution theorem (see, for example, Schick (1993), Section 2), the unique canonical gradient $g^*(X, \delta Y, \delta)$ is obtained as the orthogonal projection of $h_0(\varepsilon)$ onto the tangent space T. Hence it must be of the form

$$g^*(X, \delta Y, \delta) = u^*(X) + \delta\{s^*(\varepsilon) + \ell(\varepsilon)t^*(X)\} + \{\delta - \pi(X)\}w^*(X) \tag{2.1}$$

and is characterized by

$$E\{h_0(\varepsilon)s(\varepsilon)\} = E\{g^*(X, \delta Y, \delta)d_\gamma(X, \delta Y, \delta)\} \tag{2.2}$$

for every $\gamma \in \dot{\Theta}$. A straightforward calculation yields for the right-hand side of (2.2):

$$\begin{aligned}
&E\{g^*(X,\delta Y,\delta)d_\gamma(X,\delta Y,\delta)\}\\
&= E\big\{u^*(X)u(X)\big\} + E\delta E\big\{s^*(\varepsilon)s(\varepsilon)\big\} + E\big\{\ell_0(\varepsilon)s^*(\varepsilon)\big\}E\big\{\pi(X)t(X)\big\}\\
&+ E\big\{\ell_0(\varepsilon)s(\varepsilon)\big\}E\big\{\pi(X)t^*(X)\big\} + JE\big\{t^*(X)t(X)\big\} + E\big[\pi(X)\{1-\pi(X)\}w^*(X)w(X)\big],
\end{aligned}$$

where $\ell_0(\varepsilon)$ is the projection of $\ell(\varepsilon)$ onto $\mathcal{V}$, that is $\ell_0(\varepsilon) = \ell(\varepsilon) - \varepsilon/\sigma^2$. For convenience, we introduce the quantity J_0 which is calculated analogously to J as

$$J_0 = \int \ell_0^2\, dF = \int \Big\{\ell(z) - \frac{z}{\sigma^2}\Big\}^2 dF(z) = J - \frac{1}{\sigma^2}.$$

From (2.2) it is easy to see that $u^* = w^* = 0$. Setting $u = t = w = 0$ in (2.2) we obtain

$$\int h_0 s\, dF = E\delta \int s s^*\, dF + \int \ell_0 s\, dF \int \pi t^*\, dG$$

for all s. This gives $s^*(z) = (E\delta)^{-1}\big\{h_0(z) - \ell_0(z)\int \pi t^*\, dG\big\}$. Now set $u = s = w = 0$ in (2.2) and insert s^* to get

$$\begin{aligned}
0 &= \int \ell_0 s^*\, dF \int \pi t\, dG + J\int \pi t^* t\, dG = \int \ell_0 s^*\, dF E\delta \int t\, dG_1 + JE\delta \int t^* t\, dG_1\\
&= \int h_0\ell_0\, dF \int t\, dG_1 - J_0 E\delta \int t\, dG_1 \int t^*\, dG_1 + JE\delta \int t^* t\, dG_1
\end{aligned}$$

for all $t \in \mathcal{L}_2(G_1)$. Now consider $\mathcal{L}_2(G_1)$ written (as in Müller et al. (2004)) as an orthogonal sum of functions with mean 0 and of constants, i.e. $\mathcal{L}_2(G_1) = \mathcal{L}_{2,0}(G_1) \oplus [1]$, which means that we can write $t = (t - \int t\, dG_1) + \int t\, dG_1$. The above equation now becomes

$$0 = JE\delta \int \big(t - \int t\, dG_1\big)\big(t^* - \int t^*\, dG_1\big)\, dG_1 + \int h_0\ell_0\, dF \int t\, dG_1 + \frac{E\delta}{\sigma^2}\int t\, dG_1 \int t^*\, dG_1$$

for all $t \in \mathcal{L}_2(G_1)$. This yields

$$t^* - \int t^*\, dG_1 = 0, \quad \int t^*\, dG_1 = -\sigma^2(E\delta)^{-1}\int h_0\ell_0\, dF,$$

and thus $t^* = -\sigma^2(E\delta)^{-1}\int h_0\ell_0\, dF$. Combining the above we obtain the following result:

Lemma 2.1: *The canonical gradient of $E\{h(\varepsilon)\}$ is $g^*(X,\delta Y,\delta)$ and characterized by $(0, s^*, t^*, 0)$, where*

$$s^*(z) = \frac{1}{E\delta}\big[h_0(z) + \sigma^2 E\{h_0(\varepsilon)\ell_0(\varepsilon)\}\ell_0(z)\big] \quad \textit{and} \quad t^* = -\frac{\sigma^2}{E\delta}E\{h_0(\varepsilon)\ell_0(\varepsilon)\},$$

with $\sigma^2 = E(\varepsilon^2)$, $h_0(\varepsilon) = h(\varepsilon) - \int h\, dF - \varepsilon\sigma^{-2}\int z h(z)\, dF(z)$ and $\ell_0(\varepsilon) = \ell(\varepsilon) - \varepsilon/\sigma^2$.

An estimator $\hat{\mu}$ of $E\{h(\varepsilon)\}$ is *efficient* in the sense of Hájek and Le Cam if it is asymptotically linear with influence function equal to the canonical gradient $g^*(X, \delta Y, \delta)$ that characterizes $E\{h(\varepsilon)\}$, i.e. if

$$n^{1/2}\big\{\hat{\mu} - E\{h(\varepsilon)\}\big\} = n^{-1/2}\sum_{i=1}^{n} g^*(X_i, \delta_i Y_i, \delta_i) + o_p(1).$$

A straightforward calculation using this combined with Lemma 2.1 and formula (2.1) yields:

Corollary 2.2: *Consider the nonparametric regression model with responses missing at random. An efficient estimator $\hat{\mu}$ of $E\{h(\varepsilon)\}$ must satisfy the expansion*

$$n^{1/2}\big[\hat{\mu} - E\{h(\varepsilon)\}\big] = n^{-1/2}\sum_{i=1}^{n} \frac{\delta_i}{E\delta}\big[h(\varepsilon_i) - E\{h(\varepsilon)\} - \varepsilon_i E\{\ell(\varepsilon)h(\varepsilon)\}\big] + o_p(1).$$

Remark 2: Müller et al. (2004) construct residual-based estimators $n^{-1}\sum_{i=1}^{n} h(\hat{\varepsilon}_i)$ for estimating $E\{h(\varepsilon)\}$ in the full model. In their Section 2 they give conditions for the i.i.d. representation

$$n^{-1/2}\sum_{i=1}^{n} h(\hat{\varepsilon}_i) = n^{-1/2}\sum_{i=1}^{n}[h(\varepsilon_i) - E\{h'(\varepsilon)\}\varepsilon_i] + o_p(1),$$

which characterizes an efficient estimator. (For simplicity we assume in this remark that h is differentiable.) Note that $E\{h'(\varepsilon)\} = E\{\ell(\varepsilon)h(\varepsilon)\}$. Hence, using the transfer principle, we see that the complete case versions of their estimators have the expansion from the previous corollary. Hence they are efficient in the MAR model.

The function $h(\varepsilon) = \mathbf{1}(\varepsilon \le t)$ is of particular interest since many statistical methods are residual-based and require estimation of the error distribution function. Using Corollary 2.2 with this particular $h(\varepsilon)$, we obtain an expansion for the residual-based empirical distribution function:

Corollary 2.3: *Consider the nonparametric regression model with responses missing at random. An estimator $\hat{F}$ of the error distribution function F is efficient if it satisfies the expansion*

$$n^{1/2}\big\{\hat{F}(t) - F(t)\big\} = n^{-1/2}\sum_{i=1}^{n} \frac{\delta_i}{E\delta}\big\{\mathbf{1}(\varepsilon_i \le t) - F(t) + \varepsilon_i f(t)\big\} + o_p(1).$$

This is the expansion of the complete case estimator $\hat{\mathbb{F}}_c$ from the previous section, which completes the proof of Theorem 1.2.

3. Simulation results

To conclude we present a brief simulation study of the previous results. We also apply a goodness-of-fit test for normal errors to the residuals. For both examples we assume

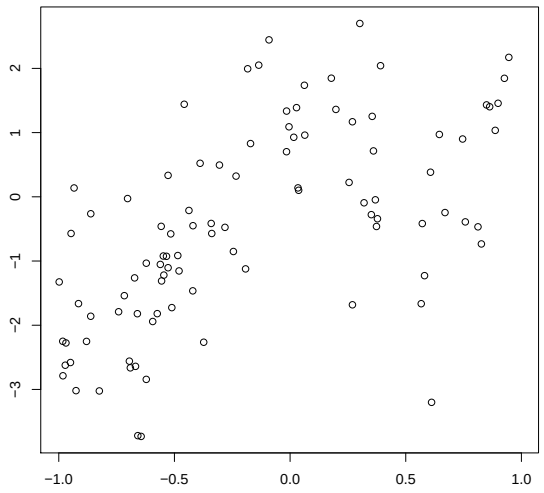

Figure 1. $r(x) = x^3 - x^2 + x + \cos\left(\frac{3\pi}{2}x\right)$, $-1 \leq x \leq 1$, with $N(0,1)$ errors.

a nonparametric regression model as before, $Y = r(X) + \varepsilon$. In order to depict the nonparametric nature of the regression function r, we choose for the simulations

$$r(x) = x^3 - x^2 + x + \cos\left(\frac{3\pi}{2}x\right).$$

The covariates were generated from a uniform distribution and the errors from a normal distribution, $X_i \sim U(-1,1)$ and $\varepsilon_i \sim N(0,1)$ for $i = 1, \ldots, n$; see Figure 1 which shows a scatterplot of a simulated dataset. Finally, the indicators δ_i have a Bernoulli($\pi(x)$) distribution, with $\pi(x) = P(\delta = 1|X = x)$. For the simulations we use the logistic distribution function for $\pi(x)$, with a mean of zero and scale parameter 1,

$$\pi(x) = \frac{1}{1 + e^{-x}}.$$

Therefore, the mean amount of missing data is around 50% and ranges between 27% and 73%. For the above choices the assumptions of Theorem 1.2 are satisfied. We work with $d = 1$, the local linear smoother, with bandwidth $c_n = 1.25\{n \log(n)\}^{-1/4}$.

3.1. *Example 1: Simulation of asymptotic mean squared error*

We consider two estimators of the error distribution function. The first estimator is the proposed complete case estimator $\hat{\mathbb{F}}_c$ and the second is a 'tuned' version of $\hat{\mathbb{F}}_c$ that utilizes an imputation technique. Similar to González-Manteiga and Pérez-González (2006), we take the initial local polynomial complete case estimator $\hat{r}_c$ (see equation (1.3)) to produce the completed sample $(X_i, \hat{Y}_i)$. We chose $\hat{Y}_i = \hat{r}_c(X_i)$ for each $i = 1, \ldots, n$. This

is a variation of the approach of González-Manteiga and Pérez-González who work with $\hat{Y}_i = \delta_i Y_i + (1-\delta_i)\hat{r}_c(X_i)$, i.e. with a "partial imputation" technique. A new local polynomial fit, $\hat{r}^*(\cdot)$, is then constructed from the completed sample. If Y is observed we can compute adjusted residuals of the form $\hat{\varepsilon}^* = Y - \hat{r}^*(X)$. Using these residuals we obtain the new tuned estimator

$$\hat{\mathbb{F}}_\iota(t) = N^{-1}\sum_{j=1}^{n}\delta_j \mathbf{1}\big(\hat{\varepsilon}_j^* \le t\big).$$

From the previous sections we know that the complete case estimator $\hat{\mathbb{F}}_c$ is an (asymptotically) efficient estimator of the error distribution function. The discussion in Remark 1 suggests that the tuned estimator $\hat{\mathbb{F}}_\iota$ is also efficient, i.e. both estimators are asymptotically equivalent: we expect that $\hat{\mathbb{F}}_\iota$ can be expanded in the same way as $\hat{\mathbb{F}}_c$,

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}\frac{\delta_j}{E\delta}\big\{\mathbf{1}(\hat{\varepsilon}_j^* \le t) - \mathbf{1}(\varepsilon_j \le t)\big\} - f(t)\int \hat{a}^*(x)G_1(dx)\Big| = o_p(n^{-1/2}),$$

where $\hat{a}^*(x)$ is now an approximation of the difference $\hat{r}^*(x) - r(x)$ (cf. equation (1.4) in Remark 1). The term involving the integral can be written as

$$f(t)\int \hat{a}^*(x)G_1(dx) = f(t)\int \hat{a}_c(c)G_1(dx) + f(t)\int\{\hat{a}^*(x) - \hat{a}_c(c)\}G_1(dx),$$

with the last term being asymptotically negligible since $\hat{a}^*(x) - \hat{a}_c(c)$ approximates the difference $\hat{r}^*(x) - \hat{r}_c(x)$ of two consistent estimators of $r(x)$. The arguments from Remark 1 then yield

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{N}\sum_{j=1}^{n}\delta_j\big\{\mathbf{1}(\hat{\varepsilon}_j^* \le t) - \mathbf{1}(\varepsilon_j \le t) - \varepsilon_j f(t)\big\}\Big| = o_p(n^{-1/2}),$$

i.e. both $\hat{\mathbb{F}}_c$ and $\hat{\mathbb{F}}_\iota$ have the same asymptotic expansion.

In order to further check the conjecture that both estimators are asymptotically equivalent, we conducted a simulation study using 1000 trials. We considered four sample sizes and five different values of t at which the error distribution function was evaluated. The findings are summarized in Table 1. Note that we also implemented another estimator, which uses partial imputation to complete the sample as suggested by González-Manteiga and Pérez-González. Since our approach performed slightly better, we report only the results for our version of $\hat{\mathbb{F}}_\iota$ which is based on $\hat{Y}_i = \hat{r}_c(X_i)$, i.e. *all* responses are imputed, and not just the missing ones. For the second smoothing step we chose the same bandwidth as in the first step, $c_n = 1.25\{n\log(n)\}^{-1/4}$.

These results show that the simulated MSE (multiplied by n) of our efficient estimator is close to the true asymptotic MSE (which equals the asymptotic variance and can be calculated using Corollary 2.3). We also see that the asymptotic MSE estimates of $\hat{\mathbb{F}}_\iota$ behave in a similar way to those of $\hat{\mathbb{F}}_c$, in particular for large sample sizes. This provides further evidence that the two approaches are indeed asymptotically equivalent. The simulated MSE's of $\hat{\mathbb{F}}_\iota$, however, more closely match the true asymptotic MSE across values of t at low sample sizes. This could be a second order effect and we believe the

Asymptotic mean squared error (MSE)										
	$t=-1.5$		$t=-1$		$t=0$		$t=1$		$t=1.5$	
n	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$
50	0.1141	0.0987	0.2705	0.2087	0.1702	0.1884	0.2865	0.2220	0.1179	0.1009
250	0.1018	0.0930	0.1800	0.1634	0.2021	0.2071	0.2022	0.1972	0.1201	0.1165
1000	0.0991	0.0945	0.1668	0.1625	0.1865	0.1997	0.1706	0.1780	0.1000	0.1008
10000	0.0925	0.0920	0.1567	0.1537	0.2068	0.2274	0.1690	0.1752	0.0953	0.0975
true	0.0911	–	0.1498	–	0.1816	–	0.1498	–	0.0911	–

Table 1. Simulated and true asymptotic MSE

most likely explanation is that $\hat{\mathbb{F}}_\iota$ can be regarded as an enhanced version of $\hat{\mathbb{F}}_c$. However, when $t=0$ both estimators $\hat{\mathbb{F}}_\iota$ and $\hat{\mathbb{F}}_c$ perform very similarly for all sample sizes. Since this value of t is also the mode of the distribution, we believe that the tuning technique using imputation is least helpful in this case.

3.2. *Example 2: Simulating a goodness-of-fit test for normal errors*

We now consider a test proposed by Müller et al. (2012) for the full model with multivariate covariates. This test was also examined by Koul et al. (2012) in the MAR model with a one-dimensional covariate, but without simulations. Both articles study versions of a martingale transform test developed by Khmaladze and Koul (2009). Under the null hypothesis, these tests tend in distribution to $\sup_{0\le t\le 1}|B(t)|$, with $B(t)$ the standard Brownian motion, i.e. they are asymptotically distribution free. This is very useful since the corresponding complete case statistics have the same limiting distributions in this case, which is a consequence of the transfer principle. This means that the decision rule remains unchanged in the MAR model. For example, setting the level of the test to 0.05, we reject H_0 if the test statistic exceeds 2.2414, the upper 5% quantile of the distribution of $\sup_{0\le t\le 1}|B(t)|$.

Writing $\phi(x)$ for the density of the $N(0,1)$ distribution and σ^2 for the error variance, the null hypothesis of normal errors is

$$H_0:\ \exists\sigma>0 \quad f(x)\ =\ \frac{1}{\sigma}\phi\Big(\frac{x}{\sigma}\Big), \qquad x\in\mathbb{R}.$$

In order to introduce the test statistic T_c set $h(x)=(1,-\phi'(x)/\phi(x),-(x\phi(x))'/\phi(x))^T$ and

$$H(t)\ =\ \int_{-\infty}^{t} h^T(x)\Gamma^{-1}(x)\phi(x)dx,$$

with $\Gamma(x)=\int_x^\infty h(z)h^T(z)\phi(z)dz$ (see Müller et al. (2012) and Koul et al. (2012) for an explicit form of $\Gamma(x)$ and for more details). Following Koul et al. (2012) we have the test statistic

$$T_c=\sup_{t\in\mathbb{R}}\Big|\frac{1}{\sqrt{N}}\sum_{j=1}^{n}\delta_j\big\{\mathbf{1}(\hat{Z}_{j,c}\le t)-H(t\wedge\hat{Z}_{j,c})h(\hat{Z}_{j,c})\big\}\Big|.$$

Test for normal errors								
	$N(0,2)$		$\chi^2_1 - 1$		t_4		Laplace$(0,2)$	
n	T_c	T_ι	T_c	T_ι	T_c	T_ι	T_c	T_ι
50	0.022	0.025	0.489	0.535	0.099	0.108	0.095	0.119
200	0.030	0.028	1.000	1.000	0.457	0.463	0.459	0.483

Table 2. Simulated level, given by N(0,2) figures, and power for T_c and T_ι.

Note that this statistic is based on our proposed estimator $\hat{\mathbb{F}}_c$ but with *scaled* residuals $\hat{Z}_{j,c} = \hat{\varepsilon}_{j,c}/\hat{\sigma}_c$, where $\hat{\sigma}_c$ is the complete case version of the residual-based empirical estimator, i.e. $\hat{\sigma}_c = \sqrt{\hat{\sigma}_c^2}$ with

$$\hat{\sigma}_c^2 = \frac{1}{N}\sum_{j=1}^{n} \delta_j \hat{\varepsilon}_{j,c}^2 = \frac{1}{N}\sum_{j=1}^{n} \delta_j \{Y_j - \hat{r}_c(X_j)\}^2.$$

Under the MAR assumption ε and δ are independent. Hence $\hat{\sigma}_c^2$ is a consistent estimator of $\mathrm{Var}(\varepsilon|\delta = 1) = \mathrm{Var}(\varepsilon) = \sigma^2$.

We are interested in studying the performance of T_c in the MAR model, and also wish to compare it with the corresponding statistic T_ι that is based on the tuned estimator $\hat{\mathbb{F}}_\iota$, i.e. T_ι has exactly the same form as T_c but with all $\hat{\varepsilon}_{j,c}$ replaced by the adjusted residuals $\hat{\varepsilon}_j^* = Y_j - \hat{r}^*(X_j)$.

For the simulations we consider the same scenario as in the previous example, but now also admit some other models for the error distribution. First we look at the $N(0,2)$ distribution to allow verification of the (5%) level of the test. For the power considerations we generated errors from a mean shifted $\chi^2(1)$ distribution, a $t(4)$ distribution and a Laplace distribution with mean 0 and variance 2. The simulation study is based on 1000 runs and samples of size 50 and 200.

Table 2 shows that when the errors are normally distributed (and the null hypothesis is true), the test using T_c rejects the null hypothesis 2.2% of the time for samples of size 50, and 3% of the time for samples of size 200. This indicates that the test using T_c is slightly conservative. Turning to T_ι we see similar conservative behavior: here the hypothesis of normality is rejected 2.5% and 2.8% of the time for sample sizes 50 and 200, respectively. When the null hypothesis is not true, the power figures are fairly close for both tests. The test using T_ι seems to be more powerful for low sample sizes. The differences are less pronounced for the larger sample size of 200, suggesting that the two tests are asymptotically equivalent – which is what we would expect given the discussion and the simulation results in the previous example. Summing up, both test procedures have similar performance. The test based on T_c appears to be the better choice for moderately large (or large) samples, as it is easier to implement.

Acknowledgements

Ursula U. Müller was supported by NSF Grant DMS 0907014. The authors thank the referees for a number of suggestions that improved the manuscript. We also thank Susan Davis for her helpful comments on an earlier draft.

References

[1] Efromovich, S. (2011a), 'Nonparametric Regression with Responses Missing at Random', *Journal of Statistical Planning and Inference*, 141, 3744-3752.

[2] Efromovich, S. (2011b), 'Nonparametric Regression with Predictors Missing at Random', *Journal of the American Statististical Association*, 106, 306-319.

[3] Khamaldze, E.V. and Koul, H.L. (2004), 'Martingale transforms goodness-of-fit tests in regression models', *Annals of Statistics*, 32, 995-1034.

[4] Khamaldze, E.V. and Koul, H.L. (2009), 'Goodness-of-fit problem for errors in nonparametric regression: distribution free approach', *Annals of Statistics*, 37, 3165-3485.

[5] Koul, H.L., Müller, U.U. and Schick, A. (2012), 'The transfer principle: a tool for complete case analysis', *Annals of Statistics*, 40, 3031-3049.

[6] González-Manteiga, W. and Pérez-González A. (2006), 'Goodness-of-fit tests for linear regression models with missing response data', *Canadian Journal of Statistics* 34, 149-170.

[7] Liang, H., Wang, S. and Carroll, R. (2007), 'Partially linear models with missing response variables and error-prone covariates', *Biometrika*, 94, 185-198.

[8] Little, R.J.A. and Rubin, D.B. (2002), *Statistical analysis with missing data*, Second edition, Wiley-Interscience.

[9] Molenberghs, G. and Kenward, M. (2007), *Missing Data in Clinical Studies*, Wiley.

[10] Müller, U.U. (2009), 'Estimating linear functionals in nonlinear regression with responses missing at random', *Annals of Statistics*, 37, 2245-2277.

[11] Müller, U.U., Schick, A. and Wefelmeyer, W. (2004), 'Estimating linear functionals of the error distribution in nonparametric regression', *Journal of Statistical Planning and Inference*, 119, 75-93.

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

[13] Müller, U.U., Schick A. and Wefelmeyer W. (2007), 'Estimating the error distribution in semiparametric regression', *Statistics and Decisions*, 25, 1-18.

[14] Müller, U.U., Schick, A. and Wefelmeyer, W. (2009), 'Estimating the error distribution function in nonparametric regression with multivariate covariates', *Statistics and Probability Letters*, 79, 957-964.

[15] Müller, U.U., Schick, A. and Wefelmeyer, W. (2012), 'Estimating the error distribution function in semiparameteric additive regression models', *Journal of Statistical Planning and Inference*, 142, 552-566.

[16] Neumeyer, N. and Van Keilegom, I. (2010), 'Estimating the error distribution in nonparametric multiple regression with applications to model testing', *Journal of Multivariate Analysis*, 101, 1067-1078.

[17] Schick, A. (1993), 'On efficient estimation in regression models', *Annals of Statistics*, 21, 1486-1521.

[18] Stute, W. (1997), 'Nonparametric model checks for regression', *Annals of Statistics*, 25, 613-641.

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