Statistics and Probability Letters 117 (2016) 31–39

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Efficient estimation of the error distribution function in heteroskedastic nonparametric regression with missing data

Justin Chown

Ruhr-Universität Bochum, Fakultät für Mathematik, Lehrstuhl für Stochastik, 44780 Bochum, Germany

ARTICLE INFO

Article history:
Received 27 May 2015
Received in revised form 13 April 2016
Accepted 13 April 2016
Available online 3 May 2016

MSC:
primary 62G05
secondary 62G08
62G20

Keywords:
Heteroskedastic
Nonparametric regression
Local polynomial smoother
Transfer principle
Efficient estimator
Empirical distribution function

ABSTRACT

We propose a residual-based empirical distribution function to estimate the distribution function of the errors of a heteroskedastic nonparametric regression with responses missing at random based on completely observed data, and we show this estimator is asymptotically most precise.

1. Introduction

An important problem encountered in practice occurs when variation in the data is found to be dynamic. A typical example is when responses Y are regressed onto a vector of m covariates X and the errors of that regression have variation changing in X. Under this condition, many statistical procedures no longer provide consistent inference. For example, consider a study of crop yields under different application amounts of a fertilizer. Should the variation in yields depend on the amount of fertilizer applied, then the classical F-test will no longer provide a consistent method of inference for a regression of the yields toward the amount of fertilizer applied because it assumes the model errors have constant variation. Examples of heteroskedastic data may be found in Greene (2000), Vinod (2008), Sheather (2009) and Asteriou and Hall (2011).

We are interested in the case where the responses are missing and observe a random sample $(X_1, \delta_1 Y_1, \delta_1), \ldots, (X_n, \delta_n Y_n, \delta_n)$ of data that is composed of independent and identically distributed copies of a base observation $(X, \delta Y, \delta)$. Here δ is an indicator variable taking values one, when Y is observed, and zero, otherwise. Throughout this article, we will interpret a datum $(X, 0, 0)$ as corresponding to a categorically missing response, i.e. when $\delta = 0$, the first zero in the datum only describes the product $0 \times Y = 0$, almost surely, because we make the common assumption that $P(|Y| < \infty) = 1$. For this work, we make the following assumption concerning the covariates X:

Assumption 1. The covariate vector X has a distribution that is quasi-uniform on the cube $[0, 1]^m$; i.e. X has a density that is both bounded and bounded away from zero on $[0, 1]^m$.

E-mail address: justin.chown@ruhr-uni-bochum.de.

http://dx.doi.org/10.1016/j.spl.2016.04.009
0167-7152/© 2016 Elsevier B.V. All rights reserved.

We assume the responses are *missing at random* (MAR), and, paraphrasing Chown and Müller (2013), we will refer to the probability model with responses missing at random as the MAR model. This means the distribution of δ given both the covariates X and the response Y depends only on the covariates X, i.e.

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = \pi(X). \tag{1.1}$$

Since we do observe some responses Y, we will assume that π is almost everywhere bounded away from zero on $[0,\ 1]^m$. It is then clear that $E\delta = E[\pi(X)]$ is positive. The MAR assumption is commonly used and it is very reasonable in many missing data situations (see Chapter 1 of Little and Rubin, 2002).

In this article we study the heteroskedastic nonparametric regression model

$$Y = r(X) + \sigma(X)e,$$

with the error e independent of the covariate vector X. In order to identify the functions r and σ, we will additionally assume the error e has mean zero and unit variance. For this work, we are interested in the case of smooth functions r and σ (see below for an explicit definition), and we will assume that σ is a positive-valued function so that it is a well-defined scale function. Hence, the model above is a well-defined heteroskedastic nonparametric regression model with identifiable components. This model is closely related to that studied in Chown and Müller (2013), who study the case of $\sigma(\cdot) \equiv \sigma_0$, a positive constant, i.e. $\sigma(x) = \sigma_0$ for almost every x. As a consequence, many results will be familiar. Here we will estimate the two functions r and σ with nonparametric function estimators that are constructed from the assumed smoothness properties of these functions. We will then use these estimates in our proposed estimator of the distribution function of the errors F.

To begin, we first consider (1.1) and observe that

$$E[\delta h(e)] = E\delta E[h(e)] \quad \text{and} \quad E[\delta h(e)|X] = \pi(X)E[h(e)]$$

for suitable measurable functions h. The relations above naturally lead to complete case estimators for each of F, r and σ. We investigate the residual-based empirical distribution function, $\hat{\mathbb{F}}_c$, given as

$$\hat{\mathbb{F}}_c(t) = \frac{1}{N}\sum_{j=1}^{n} \delta_j \mathbf{1}\big[\hat{\varepsilon}_{j,c} \le t\big] = \frac{1}{N}\sum_{j=1}^{n} \delta_j \mathbf{1}\left[\frac{Y_j - \hat{r}_c(X_j)}{\hat{\sigma}_c(X_j)} \le t\right], \quad t \in \mathbb{R}. \tag{1.2}$$

Here $N = \sum_{j=1}^{n} \delta_j$ is the number of completely observed pairs (X, Y) and the subscript "c" indicates the estimator is based on the subsample of complete cases described below, which is, in general, different from the original sample of data. Similar to the estimator of Chown and Müller (2013), this is a complete case estimator. To explain the idea, we first take our sample $(X_1, \delta_1 Y_1, \delta_1), \ldots, (X_n, \delta_n Y_n, \delta_n)$ and reorder it according to whether or not $\delta_j = 1$, $j = 1, \ldots, n$. This means we rewrite it as $(X_1, Y_1, 1), \ldots, (X_N, Y_N, 1), (X_{N+1}, 0, 0), \ldots, (X_n, 0, 0)$. Due to the i.i.d. nature of the original sample, ordering the data in this way both changes nothing and highlights the existence of two subsamples. We can write the first subsample as $(X_1, Y_1), \ldots, (X_N, Y_N)$, where $N \le n$ is the random size of this subsample, which we call the *complete cases*. Hence, our estimator uses only the part of the original sample where responses Y are actually observed. This means we use only the available residuals $\hat{\varepsilon}_{j,c} = \{Y_j - \hat{r}_c(X_j)\}/\hat{\sigma}_c(X_j)$, $j = 1, \ldots, N$, where $\hat{r}_c$ is a suitable complete case estimator for the regression function r and $\hat{\sigma}_c$ is a suitable complete case estimator of the scale function σ. Since we are only using a part of the original data based on the auxiliary information that $\delta = 1$, which now has different stochastic properties than the original data, we will, nevertheless, argue below that $\hat{\mathbb{F}}_c$ is both a consistent and an efficient estimator for F.

In this work, we use local polynomial estimators of the first and second conditional moments $r(x) = E(Y|X = x)$ and $r_2(x) = E(Y^2|X = x)$, respectively, which we will use later to construct our estimators $\hat{r}_c$ and $\hat{\sigma}_c$. Local polynomial estimation follows naturally by a Taylor expansion argument, and, therefore, follows from both of the functions r and σ satisfying certain smoothness conditions; i.e. we assume both r and σ lie on the Hölder space of functions $H(d, \varphi)$ with domain $[0,\ 1]^m$. Paraphrasing Müller et al. (2009), a function from $[0,\ 1]^m$ to $\mathbb{R}$ belongs to $H(d, \varphi)$, if it has continuous partial derivatives up to order d and the partial derivatives of order d are Hölder with exponent $0 < \varphi \le 1$. We will write $H_1(d, \varphi)$ for the unit ball of $H(d, \varphi)$ (see Müller et al., 2009, for an explicit definition).

To define the local polynomial estimators of degree d, we first introduce some notation. Let $I(d)$ be the set of multi-indices $i = (i_1, \ldots, i_m)$ such that $i_1 + \cdots + i_m \le d$. These multi-indices correspond with the partial derivatives of r and r_2 (and hence σ) whose order is at most d. The local polynomial estimators of r and r_2 are respectively given by $\hat{\gamma}_{a,0}$, for $a = 1, 2$, where $\hat{\gamma}_{a,0}$ denotes the $0 = (0, \ldots, 0)$ entry of the vector

$$\hat{\gamma}_a = \operatorname*{arg\,min}_{\gamma=(\gamma_i)_{i\in I(d)}} \sum_{j=1}^{n} \delta_j \left\{Y_j^a - \sum_{i\in I(d)} \gamma_i \psi_i\left(\frac{X_j - x}{\lambda_n}\right)\right\}^2 w\left(\frac{X_j - x}{\lambda_n}\right), \quad a = 1, 2.$$

Here

$$\psi_i(x) = \frac{x_1^{i_1}}{i_1!} \cdots \frac{x_m^{i_m}}{i_m!}, \quad x = (x_1, \ldots, x_m) \in [0,\ 1]^m,$$

$w(x) = w_1(x_1) \cdots w_m(x_m)$ is a product of densities, and $\{\lambda_n\}_{n\ge1}$ is a sequence of positive numbers satisfying $\lambda_n \to 0$, as $n \to \infty$, which we call a bandwidth. Hence, we introduce our respective function estimators of r and σ pointwise at each

x as $\hat{r}_c(x) = \hat{\gamma}_{1,0}$ and $\hat{\sigma}_c(x) = \{\hat{\gamma}_{2,0} - \hat{\gamma}_{1,0}^2\}^{1/2}$. Note that $\delta_1, \dots, \delta_n$ appear in the formula above because we require only the complete cases to estimate both r and r_2, and the minimization procedure above is unaffected by the proportion π of missing data.

Neumeyer and Van Keilegom (2010) construct an estimator related to $\hat{\mathbb{F}}_c$ for the full model using local polynomial estimators of the first and second conditional moments, i.e. the simpler case where $\delta_j = 1,\ j = 1, \dots, n$. However, these authors require the density function of the covariates g to be differentiable. For our model, we work with the conditional density function g_1 of the covariates X given $\delta = 1$. Using the identity for the conditional distribution function G_1 of the covariates X given $\delta = 1$, we have $G_1(dx) = \{\pi(x)/E\delta\}G(dx)$. Hence, any differentiability requirements imposed on the density function g would also apply to the missingness proportion π through g_1.

To alleviate this differentiability requirement, we turn our attention to the results of Müller et al. (2007, 2009), who impose no such requirement. Investigating the proof of Lemma 1 of Müller et al. (2009), reveals straightforward modifications of those results for local polynomial function estimation in a homoskedastic model to the heteroskedastic model considered here. Since this approach requires Assumption 1, using the relation between G_1 and G above and the bounding assumption on π, we observe that G_1 is quasi-uniform whenever G is quasi-uniform. We arrive at the following crucial technical corollary to Lemma 1 of Müller et al. (2009) for the estimators $\hat{r}_c$ and $\hat{\sigma}_c$.

Corollary 1. *Let Assumption 1 hold. Suppose the regression function r and the scale function σ both belong to $H(d, \varphi)$ with domain $[0,\ 1]^m$. Further suppose the error variable has mean equal to zero, variance equal to one and a finite moment of order $q > 4s/(2s - m)$, with $s = d + \varphi > 3m/2$. Assume the missingness proportion π is almost everywhere bounded away from zero on $[0,\ 1]^m$, and the densities $w_1, \dots, w_m$ are $(m+2)$-times continuously differentiable and have compact support $[-1,\ 1]$. Let $\lambda_n \sim (n\log(n))^{-1/(2s)}$. Then there is a random function $\hat{a}_{1,c}$, associated to the complete case local polynomial estimate $\hat{r}_c$ of r, such that*

$$P\big(\hat{a}_{1,c} \in H_1(m, \alpha)\big) \to 1,$$

for some $\alpha > 0$,

$$\int_{[0,\ 1]^m} \big|\hat{a}_{1,c}(x)\big|^{1+b} g_1(x)\, dx = o_p(n^{-1/2}),$$

for $b > m/(2s - m)$, and

$$\sup_{x\in[0,\ 1]^m} \big|\hat{r}_c(x) - r(x) - \hat{a}_{1,c}(x)\big| = o_p(n^{-1/2}).$$

If, additionally, the error variable has a finite moment of order $2q$, then there is a random function $\hat{a}_{2,c}$, associated to the complete case local polynomial estimate $\hat{r}_{2,c}$ of r_2, such that

$$P\big(\hat{a}_{2,c} \in H_1(m, \alpha)\big) \to 1,$$

$$\int_{[0,\ 1]^m} \big|\hat{a}_{2,c}(x)\big|^{1+b} g_1(x)\, dx = o_p(n^{-1/2})$$

and

$$\sup_{x\in[0,\ 1]^m} \big|\hat{r}_{2,c}(x) - r_2(x) - \hat{a}_{2,c}(x)\big| = o_p(n^{-1/2}).$$

Paraphrasing Remark 5 of Müller et al. (2009), there is a trade-off between the required smoothness of the regression and scale functions (indicated by the variable s) and the existence of higher order moments for the error variable e (indicated by q). This means that higher degree polynomials, used to approximate r and σ, require higher order moments of e to exist. Further, we can see that a larger bandwidth may be used to estimate these functions when they are smooth but a smaller bandwidth will be required when these functions are rough. In light of the above results, we are able to obtain analogous conclusions to those of Lemma A.2 of Neumeyer and Van Keilegom (2010).

Proposition 1. *Suppose the first set of assumptions of Corollary 1 are satisfied. Then we have*

$$\int_{[0,\ 1]^m} \frac{\hat{r}_c(x) - r(x)}{\sigma(x)} g_1(x)\, dx = \frac{1}{N}\sum_{j=1}^n \delta_j e_j + o_p(n^{-1/2}).$$

Now suppose the additional assumptions of Corollary 1 are satisfied. Then we have

$$\int_{[0,\ 1]^m} \frac{\hat{\sigma}_c(x) - \sigma(x)}{\sigma(x)} g_1(x)\, dx = \frac{1}{N}\sum_{j=1}^n \delta_j \frac{e_j^2 - 1}{2} + o_p(n^{-1/2}).$$

Remark 1. Analogous results to Corollary 1 hold for the full model where $\delta_j = 1,\ j = 1, \dots, n$, and $N = n$, almost surely, in both cases where the covariate distribution is G and G_1. Here the local polynomial estimators $\hat{r}$ for r (with associated $\hat{a}_1$)

and $\hat{r}_2$ for r_2 (with associated $\hat{a}_2$) are respectively defined exactly as $\hat{r}_c$ and $\hat{r}_{2,c}$ are defined above, but now the indicators $\delta_1, \ldots, \delta_n$ are all equal to one. Hence, we obtain estimators $\hat{r}$ for r and $\hat{\sigma}$ for σ for which analogous results of Proposition 1 hold in the full model in both cases where the covariate distribution is G and G_1. The case of covariates having distribution G confirms the findings of Lemma A.2 of Neumeyer and Van Keilegom (2010), which are required to prove their main result.

As noted in Remark 1 of Chown and Müller (2013), one proves the above statements analogously to how Müller et al. (2009) prove their results (inspect the proof of Lemma 1 of that paper). The only changes occur by introducing σ and the indicators $\delta_1, \ldots, \delta_n$. Since we also estimate r_2, this requires strengthening the moment conditions on the error variable e from q to $2q$ because $Y^2 = r^2(X) + \sigma^2(X) + 2r(X)\sigma(X)e + \sigma^2(X)(e^2 - 1)$, which follows from the model equation above. An additional requirement needed by Neumeyer and Van Keilegom (2010) for their results to hold is that $\sup_{t\in\mathbb{R}} |t^2F''(t)| < \infty$. This assumption implies the curvature of the function space underlying the probability model is finite. However, we can measure this curvature using Fisher information. This means we can merely assume that F has finite Fisher information for both location and scale, written as Assumption 2, which is a lighter assumption than $\sup_{t\in\mathbb{R}} |t^2F''(t)| < \infty$. We now arrive at our third auxiliary result, which confirms the results of Neumeyer and Van Keilegom (2010). The proof of this result is held to the supplemental material (see Appendix A).

Assumption 2. The error density f is absolutely continuous with almost everywhere derivative f' and finite Fisher information for both location and scale; i.e.

$$\int (1+z^2)\left(\frac{f'(z)}{f(z)}\right)^2 F(dz) < \infty.$$

Theorem 1 (*Expansion for the Full Model*)**.** *Assume the covariates X are distributed according to G. Let the required assumptions of Proposition 1 be satisfied concerning the local polynomial estimators $\hat{r}$ and $\hat{r}_2$ (see Remark 1). Further, let Assumption 2 hold with the error density f additionally satisfying $\sup_{t\in\mathbb{R}} f(t) < \infty$ and $\sup_{t\in\mathbb{R}} |tf(t)| < \infty$. Then, for $\hat{\varepsilon}_1 = \{Y_1 - \hat{r}(X_1)\}/\hat{\sigma}(X_1), \ldots, \hat{\varepsilon}_n = \{Y_n - \hat{r}(X_n)\}/\hat{\sigma}(X_n)$, we have*

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

We now adapt the results of Theorem 1 to the MAR model using the *transfer principle for complete case statistics* given in Koul et al. (2012). Expanding on the observations of Chown and Müller (2013), it follows that we can factor the joint distribution of (X, Y) into two components: the distribution G of the covariates X and the conditional distribution of the responses Y given X, i.e. the distribution F of the errors e. Now, using the MAR assumption, we observe that Y and δ are independent given X. This implies only the distribution G changes to G_1 when moving from full model to the MAR model, e.g. complete case statistics are based on observations (X, Y) with a joint conditional distribution given $\delta = 1$, which can now be factored into G_1 and F. Hence, the functionals F, r and σ remain the same in the MAR model. This implies the complete case statistic $\hat{\mathbb{F}}_c$ is a consistent estimator for F in the MAR model. However, in order to apply the transfer principle, we need to restate the result of Theorem 1 for covariates that have distribution G_1, which corresponds to the data used in our complete case estimator $\hat{\mathbb{F}}_c$. The proof of this result follows immediately from the proof of Theorem 1 (see the supplemental material, Appendix A) with the discussion in Remark 1.

Corollary 2 (*Expansion for the Full Model Using G_1*)**.** *Assume the covariates X are distributed according to G_1, and π is almost everywhere bounded away from zero on $[0, 1]^m$. Let the required assumptions of Proposition 1 be satisfied concerning the local polynomial estimators $\hat{r}$ and $\hat{r}_2$ (see Remark 1). Further, let Assumption 2 hold with the error density f additionally satisfying $\sup_{t\in\mathbb{R}} f(t) < \infty$ and $\sup_{t\in\mathbb{R}} |tf(t)| < \infty$. Then, for $\hat{\varepsilon}_1 = \{Y_1 - \hat{r}(X_1)\}/\hat{\sigma}(X_1), \ldots, \hat{\varepsilon}_n = \{Y_n - \hat{r}(X_n)\}/\hat{\sigma}(X_n)$, we have*

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

Combining the results above with the transfer principle for complete case statistics, we can immediately derive the expansion of our complete case estimator $\hat{\mathbb{F}}_c$. We investigate the efficiency bound for regular estimators of F in the MAR model in Section 2, i.e. estimators whose limit distributions do not depend on any direction of approach. In Corollary 3 (see Section 2), we provide the efficient influence function characterizing the class of efficient estimators of F in the MAR model. Since the influence function of our complete case estimator $\hat{\mathbb{F}}_c$ matches the efficient influence function, this characterizes $\hat{\mathbb{F}}_c$ as an efficient estimator for F in the MAR model, which implies that $\hat{\mathbb{F}}_c$ is an asymptotically most precise (least dispersed) estimator. We now arrive at the main result of this section:

Theorem 2 (*Expansion for the MAR Model*)**.** *Consider the heteroskedastic nonparametric regression model with responses missing at random. Let Assumption 2 hold with the error density f additionally satisfying $\sup_{t\in\mathbb{R}} f(t) < \infty$ and $\sup_{t\in\mathbb{R}} |tf(t)| < \infty$,* *and let the assumptions of Corollary 1 hold. Then the complete case estimator $\hat{\mathbb{F}}_c$ of the error distribution function F satisfies the uniform stochastic expansion*

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

Furthermore, $\hat{\mathbb{F}}_c$ is asymptotically linear, uniformly in $t \in \mathbb{R}$, with influence function

$$\phi(\delta, e, t) = \frac{\delta}{E\delta}\left[\mathbf{1}\big[e \le t\big] - F(t) + f(t)\left\{e + \frac{t}{2}\big(e^2 - 1\big)\right\}\right],$$

and $\hat{\mathbb{F}}_c$ is asymptotically efficient, in the sense of Hájek and Le Cam, for estimating F.

Proof. The assumptions of Theorem 1 and Corollary 2 are satisfied. Hence, for the full model, we have

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

when the covariates X are distributed under either G or G_1. Since $\hat{\mathbb{F}}_c$ is the complete case version of the estimator in the display above, it follows from Remark 2.5 of Koul et al. (2012) for the first assertion to hold, i.e.

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

This expansion is equivalent to

$$\sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{j=1}^n \frac{\delta_j}{E\delta}\left[\mathbf{1}\big[\hat{\varepsilon}_{j,c} \le t\big] - \mathbf{1}\big[e_j \le t\big] - f(t)\left\{e_j + \frac{t}{2}\big(e_j^2 - 1\big)\right\}\right]\right| = o_p(n^{-1/2}),$$

and we find, uniformly in $t \in \mathbb{R}$,

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

where the function $\phi(\delta, e, t) = (\delta/E\delta)[\mathbf{1}[e \le t] - F(t) + f(t)\{e + t/2(e^2 - 1)\}]$ is the influence function for $\hat{\mathbb{F}}_c$. Since the assumptions of Corollary 3 in Section 2 are satisfied, it follows for the influence function ϕ to be the efficient influence function for estimating F, which concludes the proof.

We note the uniform expansion above implies the existence of a functional central limit theorem. In addition, the property that $\hat{\mathbb{F}}_c$ is efficient means that competing estimators will not achieve higher precision for large samples. This includes estimators that employ imputation approaches to estimate the missing responses. A consequence of this conclusion is that imputation procedures employed to estimate F may only be effective in small samples. Therefore, we recommend the use of the complete case estimator $\hat{\mathbb{F}}_c$ for conducting various hypothesis tests concerning the heteroskedastic MAR model. Section 2 details the remaining results necessary for proving Theorem 2. Section 3 concludes the article with a numerical study of the previous results.

2. Efficiency

In this section we will construct the efficient influence function for estimating a linear functional $E[h(e)]$ based on observations of the form $(X, \delta Y, \delta)$, and later specialize this result to $F(t) = E[\mathbf{1}[e \le t]]$, $t \in \mathbb{R}$. We will first follow the arguments of Chown and Müller (2013), who study this problem for the special case of a constant variance function. In addition, we follow the arguments of Müller et al. (2006), who consider linear functionals of the joint distribution of X and Y with data of the above form. Finally, we use insight from the arguments of Schick (1994), who studies estimation of functionals from various heteroskedastic regression models. We only summarize their main arguments and refer the reader to these papers for further details. This allows us to adapt parts of those proofs to the model considered here. Consequently, we only sketch the proofs of the results in this section. To continue, we require Assumption 2 to hold.

In the following, no assumption of a parametric model (finite dimensional) is imposed on any of the regression function, the scale function or the joint distribution of the observations. This means the parameter set Θ consists of the unknown functions of the statistical model: a family of covariate distributions $\mathscr{G}$ satisfying Assumption 1, a family of error distributions $\mathscr{F}$ that have mean zero, unit variance, finite fourth moment and satisfy Assumption 2, a space of regression functions $\mathscr{R}$ that belong to $H(d, \varphi)$, a space of scale functions $\mathscr{S}$ that is a subspace of $\mathscr{R}$ composed of positive-valued functions and a family of response probability distributions $\mathscr{B}$ that are characterized by the functions from $[0,\ 1]^m$ to $(0, 1]$. More precisely, $\Theta = \mathscr{G} \times \mathscr{F} \times \mathscr{R} \times \mathscr{S} \times \mathscr{B}$.

We now proceed as in Section 2 of Chown and Müller (2013). Since the construction of the efficient influence function utilizes directional information in Θ, we now identify the set of perturbations $\dot{\Theta}$, which may be thought of as directions. Observe the joint distribution $P(dx, dy, dz)$ takes the form

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

where $B_p = p\delta_1 + (1-p)\delta_0$ denotes the Bernoulli distribution with parameter p and δ_t as the Dirac measure at t. The model considered here deviates from that considered in Chown and Müller (2013) only in the conditional distribution Q of Y given X. This means we first need to consider the spaces $\mathcal{L}_{2,0}(G)$, $\mathcal{L}_2(G_\pi)$ and $\mathcal{V}_0$. Here $\mathcal{L}_{2,0}(G)$ is the space of functions that are square integrable and have mean zero with respect to G, $\mathcal{L}_2(G_\pi)$ is a subspace of $\mathcal{L}_2(G)$, where the functions w now satisfy $E[w^2(X)\pi(X)\{1-\pi(X)\}] < \infty$, and $\mathcal{V}_0$ is the space of functions satisfying $\int v(x, y)Q(dy|x) = 0$. It then follows for perturbations G_{nu} of G, π_{nw} of π and Q_{nv} of Q that are Hellinger differentiable requires the functions u, w and v to be further restricted to appropriate subspaces. Since we have only assumed a model for Q, this only requires us to resolve the subspace $\mathcal{V}$ of $\mathcal{V}_0$.

Using the independence of the covariates X and errors e, we may write

$$\frac{d}{dy}Q(y|x) = f\left(\frac{y - r(x)}{\sigma(x)}\right)\frac{1}{\sigma(x)}.$$

Hence, in order to derive the explicit form of $\mathcal{V}$, we introduce further perturbations s, t and m of the unknown functions f, r and σ, respectively, and write

$$\frac{d}{dy}Q_{nv}(y|x) = \frac{d}{dy}Q_{nstm}(y|x) = f_{ns}\left(\frac{y - r_{nt}(x)}{\sigma_{nm}(x)}\right)\frac{1}{\sigma_{nm}(x)},$$

where $f_{ns}(z) = f(z)\{1 + n^{-1/2}s(z)\}$, $r_{nt}(x) = r(x) + n^{-1/2}t(x)$ and $\sigma_{nm}(x) = \sigma(x) + n^{-1/2}m(x)$ for $s \in \mathcal{S}$, $t \in \mathcal{L}_2(G_1)$ and $m \in \mathcal{L}_2(G_1)$. Here

$$\mathcal{S} = \left\{s \in \mathcal{L}_2(F) : \int s(z)f(z)dz = 0, \int zs(z)f(z)dz = 0 \text{ and } \int z^2 s(z)f(z)dz = 0\right\},$$

which is derived by the constraints that f_{ns} must integrate to one, have mean zero and have unit variance. In the following we will write "$\doteq$" to denote asymptotic equivalence; i.e. equality up to an additive term of order $o_p(n^{-1/2})$. In addition, we introduce the notation $\mathrm{l}(z) = (\ell_1(z), \ell_2(z))^T$, for $\ell_1(z) = -f'(z)/f(z)$ and $\ell_2(z) = -1 - zf'(z)/f(z)$, $\mathbf{k}(x) = (t(x)/\sigma(x), m(x)/\sigma(x))^T$, $\mathbf{e}_1 = (1, 0)^T$ and $\mathbf{e}_2 = (0, 1)^T$. Similar to the calculations of Chown and Müller (2013) and Schick (1994), who consider, more generally, directionally differentiable regression and scale functions, we have, by a brief sketch,

$$f_{ns}\left(\frac{y - r_{nt}(x)}{\sigma_{nm}(x)}\right)\frac{1}{\sigma_{nm}(x)} \doteq f\left(\frac{y - r(x)}{\sigma(x)}\right)\frac{1}{\sigma(x)} \times \left\{1 + n^{-1/2}\left[\mathbf{k}^T(x)\mathrm{l}\left(\frac{y - r(x)}{\sigma(x)}\right) + s\left(\frac{y - r(x)}{\sigma(x)}\right)\right]\right\}.$$

Hence,

$$\frac{d}{dy}Q_{ns\mathbf{k}}(y|x) \doteq f\left(\frac{y - r(x)}{\sigma(x)}\right)\frac{1}{\sigma(x)}\left\{1 + n^{-1/2}\left[\mathbf{k}^T(x)\mathrm{l}\left(\frac{y - r(x)}{\sigma(x)}\right) + s\left(\frac{y - r(x)}{\sigma(x)}\right)\right]\right\}$$

and $\mathcal{V}$ takes the form

$$\mathcal{V} = \left\{v(x, y) = \mathbf{k}^T(x)\mathrm{l}\left(\frac{y - r(x)}{\sigma(x)}\right) + s\left(\frac{y - r(x)}{\sigma(x)}\right) : \mathbf{k} \in \mathcal{L}_2(G_1) \times \mathcal{L}_2(G_1) \text{ and } s \in \mathcal{S}\right\}.$$

Thus we have perturbations $\dot{\Theta} = \mathcal{L}_{2,0}(G) \times \mathcal{S} \times \{\mathcal{L}_2(G_1) \times \mathcal{L}_2(G_1)\} \times \mathcal{L}_2(G_\pi)$. Observe, for any $\gamma = (u, s, \mathbf{k}, w)$ in $\dot{\Theta}$, the perturbed distribution $P_{n\gamma}(dx, dy, dz)$ of an observation $(X, \delta Y, \delta)$ is then

$$P_{n\gamma}(dx, dy, dz) = G_{nu}(dx)B_{\pi_{nw}(x)}(dz)\Big\{zQ_{ns\mathbf{k}}(dy|x) + (1-z)\delta_0(dy)\Big\}.$$

It follows that P is Hellinger differentiable with tangent

$$d_\gamma\big(X, \delta Y, \delta\big) = u(X) + \big\{\delta - \pi(X)\big\}w(X) + \delta\big\{\mathbf{k}^T(X)\mathrm{l}(e) + s(e)\big\},$$

and we arrive at the form of the tangent space as

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

Consequently, we have local asymptotic normality. This means the following expansion holds:

$$\sum_{j=1}^{n}\log\left(\frac{dP_{n\gamma}}{dP}\big(X_j, \delta_j Y_j, \delta_j\big)\right) = n^{-1/2}\sum_{j=1}^{n} d_\gamma\big(X_j, \delta_j Y_j, \delta_j\big) - \frac{1}{2}E\Big[d_\gamma^2\big(X, \delta Y, \delta\big)\Big] + o_p(1).$$

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

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

with h_0 given as a projection of h onto $\mathcal{S}$:

$$h_0(z) = h(z) - E[h(e)] - zE[eh(e)] - \frac{z^2 - E[e^3]z - 1}{E[e^4] - E^2[e^3] - 1}\Big\{E\big[e^2h(e)\big] - E\big[e^3\big]E[eh(e)] - E[h(e)]\Big\}.$$

Thus, $E[h(e)]$ is directionally differentiable with directional derivative $(0, h_0, \mathbf{0}, 0)$ and gradient $h_0(e)$. By the convolution theorem (see, for example, Section 2 of Schick, 1993) the unique canonical gradient $g^*(X, \delta Y, \delta)$ is found by orthogonally projecting the gradient $h_0(e)$ onto the tangent space T. Thus, $g^*(X, \delta Y, \delta)$ must take the form

$$g^*\big(X, \delta Y, \delta\big) = u^*(X) + \big\{\delta - \pi(X)\big\}w^*(X) + \delta\big\{\mathbf{k}^{*T}(X)\mathrm{l}(e) + s^*(e)\big\}. \qquad (2.1)$$

Now proceeding as in Section 2 of Chown and Müller (2013), we obtain the following result:

Lemma 1. *The canonical gradient of* $E[h(e)]$ *is* $g^*(X, \delta Y, \delta)$ *and is characterized by* $(0, s^*, \mathbf{k}^*, 0)$, *where*

$$s^*(z) = \frac{1}{E\delta}h_0(z) - E_1\big[\mathbf{k}^{*T}(X)\big]\mathrm{l}_0(z) \quad \textit{and} \quad \mathbf{k}^* \equiv -\frac{1}{E\delta}J_d^{-1}E\big[\mathrm{l}_0(e)h_0(e)\big],$$

with h_0 *given above and the quantities*

$$\mathrm{l}_0(z) = \mathrm{l}(z) - z\mathbf{e}_1 - \frac{z^2 - E[e^3]z - 1}{E[e^4] - E^2[e^3] - 1}\Big\{2\mathbf{e}_2 - E\big[e^3\big]\mathbf{e}_1\Big\} \quad \textit{and}$$

$$J_d^{-1} = \frac{1}{E[e^4] - E^2[e^3] - 1}\begin{bmatrix} E[e^4] - 1 & -2E[e^3] \\ -2E[e^3] & 4 \end{bmatrix}.$$

We will call an estimator $\hat{\mu}$ for $E[h(e)]$ efficient, in the sense of Hájek and Le Cam, if it is asymptotically linear with corresponding influence function equal to the canonical gradient $g^*(X, \delta Y, \delta)$ that characterizes $E[h(e)]$. This means $\hat{\mu}$ satisfies the expansion

$$n^{1/2}\big\{\hat{\mu} - E[h(e)]\big\} = n^{-1/2}\sum_{j=1}^{n} g^*\big(X_j, \delta_j Y_j, \delta_j\big) + o_p(1).$$

We combine this fact with Lemma 1 and (2.1) to obtain the following result:

Theorem 3. *Consider the heteroskedastic nonparametric regression model with responses missing at random. An estimator* $\hat{\mu}$ *of* $E[h(e)]$ *is efficient, if it satisfies the expansion*

$$n^{1/2}\big\{\hat{\mu} - E[h(e)]\big\} = n^{-1/2}\sum_{j=1}^{n}\frac{\delta}{E\delta}\Big[h_0(e_j) - E^T\big[h_0(e_j)\mathrm{l}_0(e_j)\big]J_d^{-1}\mathrm{l}_d(e_j)\Big] + o_p(1),$$

where h_0 *is given above,* l_0 *and* J_d^{-1} *are given in Lemma 1 and*

$$\mathrm{l}_d(z) = z\mathbf{e}_1 + \frac{z^2 - zE[e^3] - 1}{E[e^4] - E^2[e^3] - 1}\Big\{2\mathbf{e}_2 - E\big[e^3\big]\mathbf{e}_1\Big\}.$$

In this article, we are interested in the function $h(z) = \mathbf{1}[z \le t]$ because we estimate $F(t) = E[\mathbf{1}[e \le t]]$ using $\hat{\mathbb{F}}_c$. We now obtain, using Theorem 3 with this h, the expansion for an efficient estimator of the error distribution function F.

Corollary 3. *Consider the heteroskedastic nonparametric regression model with responses missing at random. An estimator* $\hat{F}$ *of* F *is efficient, in the sense of Hájek and Le Cam, if it satisfies the expansion*

$$n^{1/2}\big\{\hat{F}(t) - F(t)\big\} = n^{-1/2}\sum_{j=1}^{n}\frac{\delta}{E\delta}\bigg[\mathbf{1}\big[e_j \le t\big] - F(t) + f(t)\Big\{e_j + \frac{t}{2}\big(e_j^2 - 1\big)\Big\}\bigg] + o_p(1).$$

Table 1
Simulated asymptotic bias and variance (in parentheses) of $\hat{\mathbb{F}}_c$.

Asymptotic bias and variance of $\hat{\mathbb{F}}_c$				
n	$t = 0$	$t = -1$	$t = -2$	$t = -3$
100	−0.0318 (0.1957)	0.0030 (0.1516)	0.0563 (0.0588)	0.0977 (0.0231)
200	−0.0818 (0.2124)	0.1646 (0.1555)	0.0777 (0.0722)	0.0965 (0.0241)
500	−0.0746 (0.2022)	0.1806 (0.1285)	0.0008 (0.0496)	0.0301 (0.0089)
1000	−0.0826 (0.1848)	0.1389 (0.1033)	−0.0382 (0.0348)	0.0006 (0.0030)

Table 2
Simulated AMSE and AMISE of $\hat{\mathbb{F}}_c$.

AMSE and AMISE of $\hat{\mathbb{F}}_c$					
n	AMSE				AMISE
	$t = 0$	$t = -1$	$t = -2$	$t = -3$	
100	0.1967	0.1516	0.0619	0.0326	0.8248
200	0.2191	0.1826	0.0782	0.0334	0.9248
500	0.2077	0.1611	0.0496	0.0098	0.7184
1000	0.1916	0.1226	0.0362	0.0030	0.5812
∞	0.1817	0.0913	0.0270	0.0025	0.4231

3. Simulations

We conclude this article with a small numerical study of the previous results. In the following we work with

$$r\big(x_1, x_2\big) = 1 + x_1 - x_2 + 2e^{-\frac{1}{2}\sqrt{x_1^2+x_2^2}} \quad \text{and} \quad \sigma\big(x_1, x_2\big) = \sqrt{1 + 2x_1^2 + 2x_2^2}$$

to preserve the nonparametric nature of the study. The covariates X_1 and X_2 are each randomly generated from a $U(-1, 1)$ distribution, and the errors e are generated from a standard normal distribution. The indicators δ are randomly generated from a Bernoulli $(\pi(X_1, X_2))$ distribution, with $\pi(X_1, X_2) = P(\delta = 1|X_1, X_2)$. Here we use $\pi(x_1, x_2) = 1-1/(1+e^{-(x_1+x_2)/2})$. Consequently, the average amount of missing data is about 50% (ranging between 26% and 74%). We work with $d = 3$, the locally cubic smoother, to estimate both of the functions r and σ. For our choice of using a product of tricubic kernel functions and bandwidth $\lambda_n = 3(n\log(n))^{-1/7}$, the assumptions of Theorem 2 are satisfied.

To check the performance of our proposed estimator, we have conducted simulations of 1000 runs using samples of sizes 100, 200, 500 and 1000. The distribution function has been estimated at the t-values 0, −1, −2 and −3 (the results for t-values 1, 2, and 3 are very similar). Table 1 shows the results of the simulated asymptotic bias and variance of $\hat{\mathbb{F}}_c$, which is calculated by multiplying the simulated bias by the square-root of each sample size and multiplying the simulated variance by each sample size. These quantities are predicted to be constant across sample sizes by Theorem 2, and, therefore, will change only with the value of t. Table 2 shows the results of the simulated asymptotic mean squared error (AMSE) and the simulated asymptotic mean integrated squared error (AMISE), which are calculated similarly to the simulated asymptotic variance. In addition, we have calculated the AMSE and AMISE for an infinitely large sample using the results of Theorem 2, which are given by the figures labeled with sample size ∞.

Beginning with Table 1, we can see the asymptotic bias in $\hat{\mathbb{F}}_c$ is slightly negative near zero, increases to become positive when moving away from zero and, finally, decreases toward zero again when moving into the tails of the distribution. This is in contrast to the asymptotic variance, which appears to be largest near zero and only decreases toward zero when moving into the tails of the distribution. Nevertheless, we can see the values appear reasonably stable at the larger sample sizes 500 and 1000 as desired. Turning our attention now to Table 2, we can plainly see the estimator $\hat{\mathbb{F}}_c$ appears to have both AMSE and AMISE values decreasing toward the respective predicted limiting values (given by the ∞ figures). This indicates the predictions made by Theorem 2 are indeed adequate for describing the limiting behavior of $\hat{\mathbb{F}}_c$. In conclusion we find the complete case estimator $\hat{\mathbb{F}}_c$ useful and practical for estimating the distribution of the errors F in the heteroskedastic MAR model.

Acknowledgments

Justin Chown acknowledges financial support from the contract ‘Projet d’Actions de Recherche Concertées’ (ARC) 11/16-039 of the ‘Communauté française de Belgique’, granted by the ‘Académie universitaire Louvain’, the IAP research network P7/06 of the Belgian Government (Belgian Science Policy) and the Collaborative Research Center “Statistical modeling of nonlinear dynamic processes” (SFB 823, Teilprojekt C4) of the German Research Foundation (DFG). I would like to thank the two referees for their careful reading of the manuscript and the very helpful comments that greatly improved the quality of this work.

Appendix A. Supplementary material

This text contains the proof of Theorem 1. Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.spl.2016.04.009.

References

Asteriou, D., Hall, S.G., 2011. Applied Econometrics. Palgrave MacMillan, New York, New York, ISBN: 9780230271821.

Chown, J., Müller, U.U., 2013. Efficiently estimating the error distribution in nonparametric regression with responses missing at random. J. Nonparametr. Stat. 25, 665–677.

Greene, W.H., 2000. Econometric Analysis. Prentice Hall, Upper Saddle River, New Jersey, ISBN: 9780130132970.

Koul, H.L., Müller, U.U., Schick, A., 2012. The transfer principle: a tool for complete case analysis. Ann. Statist. 40, 3031–3049.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. In: Wiley Series in Probability and Mathematical Statistics, Wiley, Hoboken, New Jersey, ISBN: 9780471183860.

Müller, U.U., Schick, A., Wefelmeyer, W., 2004. Estimating linear functionals of the error distribution in nonparametric regression. J. Statist. Plann. Inference 119, 75–93.

Müller, U.U., Schick, A., Wefelmeyer, W., 2006. Imputing responses that are not missing. In: Nikulin, M., Commenges, D., Huber, C. (Eds.), Probability, Statistics and Modelling in Public Health. Springer, New York, New York, ISBN: 9780387260235, pp. 350–363.

Müller, U.U., Schick, A., Wefelmeyer, W., 2007. Estimating the error distribution in semiparametric regression. Statist. Decisions 25, 1–18.

Müller, U.U., Schick, A., Wefelmeyer, W., 2009. Estimating the error distribution function in nonparametric regression with multivariate covariates. Statist. Probab. Lett. 79, 957–964.

Neumeyer, N., Van Keilegom, I., 2010. Estimating the error distribution in nonparametric multiple regression with applications to model testing. J. Multivariate Anal. 101, 1067–1078.

Schick, A., 1993. On efficient estimation in regression models. Ann. Statist. 21, 1486–1521.

Schick, A., 1994. On efficient estimation in regression models with unknown scale functions. Math. Methods Statist. 3, 171–212.

Sheather, S.J., 2009. A Modern Approach to Regression with R. In: Springer Series in Language and Communication, Springer, New York, New York, ISBN: 9780387096087.

Vinod, H.D., 2008. Hands-on Intermediate Econometrics Using R: Templates for Extending Dozens of Practical Examples. World Scientific, Hackensack, New Jersey, ISBN: 9789812818850.