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 $(X_1, \delta_1 Y_1, \delta_1), \dots, (X_n, \delta_n Y_n, \delta_n)$ 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 $\{\lambda_n\}_{n\ge 1}$ 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