Journal of Statistical Planning and Inference 142 (2012) 1198–1214

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Estimating the density of a possibly missing response variable in nonlinear regression

Ursula U. Müller [1]

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

ARTICLE INFO

Article history:
Received 30 January 2010
Received in revised form
20 September 2011
Accepted 29 November 2011
Available online 7 December 2011

Keywords:
Least dispersed estimator
Semiparametric regression
Empirical likelihood
Influence function
Gradient

ABSTRACT

This paper considers linear and nonlinear regression with a response variable that is allowed to be "missing at random". The only structural assumptions on the distribution of the variables are that the errors have mean zero and are independent of the covariates. The independence assumption is important. It enables us to construct an estimator for the response density that uses all the observed data, in contrast to the usual local smoothing techniques, and which therefore permits a faster rate of convergence. The idea is to write the response density as a convolution integral which can be estimated by an empirical version, with a weighted residual-based kernel estimator plugged in for the error density. For an appropriate class of regression functions, and a suitably chosen bandwidth, this estimator is consistent and converges with the optimal parametric rate $n^{1/2}$. Moreover, the estimator is proved to be efficient (in the sense of Hájek and Le Cam) if an efficient estimator is used for the regression parameter.

1. Introduction

We study regression models of the form $Y = r_\vartheta(X) + \varepsilon$, where r_ϑ is a linear or nonlinear regression function that depends smoothly on a finite-dimensional parameter vector ϑ. We assume that the covariate vector X and the error variable ε are independent, and that the errors have mean zero and finite variance. We will not make any further model assumptions on the distributions of X and ε, in other words our model is a *semiparametric* regression model. Note that this model with an *unknown* error distribution is particularly relevant in situations where it is not appropriate to assume that the errors are from a normal distribution, or from some other specific distribution, which would allow estimation of the regression function using likelihood techniques.

We are interested in situations where some of the responses Y are missing. More precisely, we will assume that Y is *missing at random* (MAR). This means that we always observe X, but only observe Y in those cases where some indicator δ equals one, and the indicator δ is conditionally independent of Y given X, i.e. $P(\delta = 1|X,Y) = P(\delta = 1|X) = \pi(X)$. MAR is a common assumption and is reasonable in many situations (see Little and Rubin, 2002, Chapter 1). One example would be the problem of non-responses in survey questions: assume, for example, that additional data about socioeconomic status are available. It is possible that the response probabilities are different for subjects with different socioeconomic

E-mail address: uschi@stat.tamu.edu
URL: http://www.stat.tamu.edu/~uschi/

[1] Ursula U. Müller was supported by NSF Grant DMS-0907014.

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.11.021

backgrounds. It is also possible that subjects from the same status group are equally likely to respond, regardless what the response would be.

Note that the more intuitive notion of randomness for the missing value mechanism is called *missing completely at random* (MCAR). Here the missing value mechanism does not depend on observed or unobserved measurements, i.e. $P(\delta=1|X,Y)$ is a constant, $P(\delta=1|X,Y)=\pi(X)=\pi$. (By assuming MAR responses we will also cover the MCAR situation, since $\pi(X)=\pi$ is simply a special case.) The situation when data are *not missing at random* (NMAR) will not be studied here: in this case $P(\delta=1|X,Y)=\pi(X,Y)$ is a function of X and Y. It therefore depends on data that are missing, which means that inference requires auxiliary information. With the MAR assumption, $P(\delta=1|X,Y)=\pi(X)$ depends only on observable data, i.e. the mechanism $\pi(X)$ is "ignorable": it need not be modeled since it can be estimated from the data.

In this paper we study density estimators, specifically consistent estimators of the density of the response variable Y which converge with the unusual (fast) rate $n^{1/2}$. The simplest (slowly converging) estimator of the density of a variable Y at some point y is a kernel estimator based on observed responses

$$\frac{1}{N}\sum_{i=1}^{n}\delta_i k_b(y-Y_i), \tag{1.1}$$

where N is the number of completely observed data points, $N=\sum_{j=1}^{n}\delta_j$, and where $k_b(y)=k(y/b)/b$ with kernel function k and bandwidth $b>0$. If there is additional information available in the form of a single covariate or a covariate vector X, then it is intuitively clear that an estimator which uses the additional information should be better than the kernel estimator above. This idea is, for example, used by Wang (2008) for a related regression model with MAR responses, but without assuming independence of covariates and errors. He introduces a probability weighted estimator and an imputed estimator, and proves local asymptotic normality—but with rates slower than $n^{1/2}$, which is typical for kernel estimators. Also related is Mojirsheibani (2007), who studies partial imputation for response density estimators in a nonparametric regression setting with MAR responses. He also obtains convergence rates that are slower than $n^{1/2}$.

Here we construct an estimator for the response density from a sample $(X_i,\delta_i Y_i,\delta_i)$. Under appropriate conditions on the regression function and the distribution, our estimator will converge with the desired rate $n^{1/2}$, and, beyond that, will be efficient. The case with missing responses is an important generalization of the case with fully observed data, which is covered as a special case with all indicators $\delta_i=1$. This is a research area where little work has been done, even if we include cases where all data are observed.

In order to introduce the estimator we write M for the covariate distribution and f for the error density. We also suppose that $r_\vartheta(X)$ has a density g. Then the response density, say $q(y)$, can be written as a convolution integral

$$q(y)=\int f\{y-r_\vartheta(x)\}M(dx)=\int f(y-u)g(u)\,du=\int f(u)g(y-u)\,du=E\{g(y-\varepsilon)\}=f*g(y).$$

This representation suggests two plug-in estimators of the integral: firstly, a convolution of kernel density estimators

$$\hat{q}(y)=\hat{f}*\hat{g}(y)\quad\text{with}\quad \hat{f}(z)=\frac{1}{N}\sum_{j=1}^{n}\delta_j k_b(z-\hat{\varepsilon}_j),$$

$$\hat{g}(z)=\frac{1}{n}\sum_{j=1}^{n}k_b\{z-r_{\hat{\vartheta}}(X_j)\}, \tag{1.2}$$

where $\hat{\varepsilon}_i=Y_i-r_{\hat{\vartheta}}(X_i)$ are the residuals based on a $n^{1/2}$-consistent estimator $\hat{\vartheta}$ of ϑ. Another obvious and perhaps even simpler estimator is

$$\int \hat{f}\{y-r_{\hat{\vartheta}}(x)\}\,d\hat{M}(x)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}\{y-r_{\hat{\vartheta}}(X_i)\} \tag{1.3}$$

with $\hat{f}$ from above. Here $\hat{M}$ is just the empirical covariate distribution function. For technical reasons we will work with estimator (1.2) in this paper. However, it is easy to see that this estimator can always be written in the form (1.3). The reverse does not hold in general. The two estimators are the same if, for example, the kernel k in $\hat{f}$ is the standard normal density. (See Section 5 for more details.)

Note that the two estimators (1.2) and (1.3) use all observations, whereas the usual kernel estimator (1.1) only uses a fraction of the data, namely the responses Y_i in a neighborhood of y, which explains the faster convergence rate. The convolution approach is therefore, in general, better than the usual approach (1.1), and even better than the usual estimator based on complete data pairs. A degenerate case is given if the regression function is a constant, $r_\vartheta(x)=\vartheta$. In this case we do not estimate an integral: $q(y)$ is just a shift of the error density, $q(y)=\int f(y-\vartheta)M(dx)=f(y-\vartheta)$—which is estimated with the usual slow rates.

Now suppose that the response density can be written as a non-degenerate convolution integral. The estimator (1.2) will, in general, be $n^{1/2}$-consistent but not efficient. In order to make it efficient we have to use an efficient estimator for $\hat{\vartheta}$. We will also have to incorporate the mean zero constraint on the error distribution, which we achieve by adding Owen's empirical likelihood weights. The weighted estimator of the error density, $\hat{f}_w$, is then a weighted version of $\hat{f}$, with weights $\hat{w}_j$ based on residuals $\hat{\varepsilon}_j=Y_j-r_{\hat{\vartheta}}(X_j)$ such that the weighted residuals are centered around zero, $\sum_{j=1}^{n}\hat{w}_j\delta_j\hat{\varepsilon}_j=0$. Our final