Journal of Multivariate Analysis 128 (2014) 51–61

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Testing for additivity in partially linear regression with possibly missing responses

Ursula U. Müller [a,*], Anton Schick [b], Wolfgang Wefelmeyer [c]

[a] Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA
[b] Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, USA
[c] Mathematical Institute, University of Cologne, Weyertal 86–90, 50931 Cologne, Germany

ARTICLE INFO

Article history:
Received 10 August 2012
Available online 17 March 2014

AMS subject classifications:
62G10
62G08
62G30

Keywords:
Partially linear regression
Additive regression
Local polynomial smoother
Marginal integration estimator
Uniform stochastic expansion
Responses missing at random

ABSTRACT

We consider a partially linear regression model with multivariate covariates and with responses that are allowed to be missing at random. This covers the usual settings with fully observed data and the nonparametric regression model as special cases. We first develop a test for additivity of the nonparametric part in the complete data model. The test statistic is based on the difference between two empirical estimators that estimate the errors in two ways: the first uses a local polynomial smoother for the nonparametric part; the second estimates the additive components by a marginal integration estimator derived from the local polynomial smoother. We present a uniform stochastic expansion of the empirical estimator based on the marginal integration estimator, and we derive the asymptotic distribution of the test statistic. The transfer principle of Koul et al. (2012) then allows a direct adaptation of the results to the case when responses are missing at random. We examine the performance of the tests in a small simulation study.

1. Introduction

Data sets with a large number of covariates are commonly observed in applications, in particular in biological studies. It is well known that many nonparametric methods do not perform well in this situation, which is often referred to as the 'curse of dimensionality'. A popular semiparametric model which is used to cope with this difficulty is the partially linear model. It combines the flexible nonparametric regression model with the basic linear regression model. In this article we consider a partially linear regression model of the form

$$Y = \vartheta^\top U + \varrho(X) + \varepsilon,$$

where ϑ is an unknown vector in $\mathbb{R}^p$ and ϱ is an unknown smooth function. The error ε has mean zero and is assumed to be independent of the pair (U, X), where U and X are (random) covariate vectors. In the ideal situation one observes the triplet (U, X, Y). However, in almost all real life data sets there are missing values. This is an important problem which needs to be handled with care, since the presence of missing data can easily distort statistical inferences if the wrong method is used. In this article we are specifically interested in the case when some responses Y are missing. Then one observes $(\delta, U, X, \delta Y)$ with δ an indicator random variable, with the interpretation that for $\delta = 1$ one observes the full triplet (U, X, Y), while for

* Corresponding author.
E-mail addresses: uschi@stat.tamu.edu (U.U. Müller), anton@math.binghamton.edu (A. Schick), wefelm@math.uni-koeln.de (W. Wefelmeyer).

http://dx.doi.org/10.1016/j.jmva.2014.03.003
0047-259X/

$\delta = 0$ one observes only the covariates (U, X). We make the common assumption that the responses are *missing at random*, which means that the conditional distribution of δ given (U, X, Y) depends only on the covariates (U, X),

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

Monographs on missing data are [19,36].

The partially linear regression model considered here has by definition a partially *additive* structure. We want to go one step further and test the hypothesis that the regression function is completely additive, i.e. even the smooth function ϱ is actually additive,

$$\varrho(x) = \varrho_1(x_1) + \cdots + \varrho_q(x_q), \quad x = (x_1, \ldots, x_q) \in \mathbb{R}^q.$$

It is important to have a diagnostic tool to assess additivity. As shown by Stone [32], additive models avoid the curse of dimensionality and are easy to interpret.

We will first develop a test procedure for the model with fully observed data, which we describe next. Then we will apply a method by Koul et al. [15], which they call the *transfer principle*, to derive a corresponding procedure for the model with missing responses. The *transfer principle* is a novel approach that makes it easy to derive procedures for certain missing data problems from those with fully observed data.

Assume that we observe n independent copies $(U_1, X_1, Y_1), \ldots, (U_n, X_n, Y_n)$ of (U, X, Y). Our test statistic for additivity will be of the form

$$T = n^{1/2}\|\hat{\mathbb{F}} - \tilde{\mathbb{F}}\| = n^{1/2} \sup_{t\in\mathbb{R}} |\hat{\mathbb{F}}(t) - \tilde{\mathbb{F}}(t)|$$

with two different residual-based empirical distribution functions $\hat{\mathbb{F}}$ and $\tilde{\mathbb{F}}$.

The first uses residuals of the form $\hat{\varepsilon}_j = Y_j - \hat{\vartheta}^\top U_j - \hat{\varrho}(X_j)$ with $\hat{\varrho}$ a local polynomial smoother based on the covariates X_j and the "observations" $Y_j - \hat{\vartheta}^\top U_j$. The second exploits the additivity assumption and works with residuals of the form $\tilde{\varepsilon}_j = Y_j - \hat{\vartheta}^\top U_j - \tilde{\varrho}(X_j)$ with $\tilde{\varrho}$ the marginal integration estimator derived from $\hat{\varrho}$. In both cases, $\hat{\vartheta}$ is some $\sqrt{n}$-consistent estimator of ϑ. Efficient estimators of ϑ for additive ϱ are constructed in [28]. Our test statistic T is a variant of the test statistic in [24], who test for additivity in a nonparametric regression model with heteroscedastic errors. Those authors study a bootstrap test based on their test statistic. Here we use the asymptotic distribution to develop our test. We show that, under additivity, T converges in distribution to $\kappa|Z|$, where Z is standard normal and κ is a constant depending on the underlying distribution. This leads us to the test $\mathbf{1}[T > \hat{\kappa} z_{\alpha/2}]$ which rejects the null hypothesis if T exceeds $\hat{\kappa} z_{\alpha/2}$ with $z_{\alpha/2}$ the $(1 - \alpha/2)$-quantile of the standard normal distribution and $\hat{\kappa}$ a consistent estimator of κ.

Our test for missing data uses the *complete case* version of the above test, which is constructed using only the observations with observed responses. More precisely, we reject the null hypothesis if T_c exceeds $\hat{\kappa}_c z_{\alpha/2}$, where T_c and $\hat{\kappa}_c$ are the complete case versions of T and $\hat{\kappa}$. The complete case version of a statistic $S_n = s_n((U_1, X_1, Y_1), \ldots, (U_n, X_n, Y_n))$ is of the form $S_c = s_N((U_{i_1}, X_{i_1}, Y_{i_1}), \ldots, (U_{i_N}, X_{i_N}, Y_{i_N}))$, where $(U_{i_1}, X_{i_1}, Y_{i_1}), \ldots, (U_{i_N}, X_{i_N}, Y_{i_N})$ are the $N = \sum_{j=1}^n \delta_j$ observations with observed responses. An implementation of the test is straightforward since it suffices to write a program for the model with fully observed data. This program then can be used for applications with responses missing at random: just delete all cases where only the covariates are available and work with the remaining N cases that are complete. Since we assume that the covariates and the errors are independent it is clear that the covariates *alone* do not carry information about the error distribution: the complete cases are *sufficient* for inference about functionals of the error distribution function F; see also the discussion in [15].

The reason for using the marginal integration estimator in $\tilde{\mathbb{F}}$ is that the stochastic expansion of $\tilde{\mathbb{F}}$ is then *different* from that of $\hat{\mathbb{F}}$ even under the hypothesis of additivity of ϱ, as will be shown in Section 2. This is necessary for the test based on T to have power under *contiguous* alternatives of the form $\varrho(x) = \varrho_1(x_1) + \cdots + \varrho_q(x_q) + n^{-1/2}s(x)$. The two stochastic expansions of $\hat{\mathbb{F}}$ and $\tilde{\mathbb{F}}$ imply in particular an expansion of our test statistic T under the hypothesis of additivity. From this we obtain the asymptotic distribution of T and hence an asymptotic critical value for the test.

We note that the marginal integration estimator is not particularly well suited for estimating the error distribution function. A better estimator would be the series estimator studied in Section 4 of [23]. The empirical distribution function of this estimator would however be stochastically equivalent to $\hat{\mathbb{F}}$ and therefore lead to a test with local asymptotic power equal to the significance level. The estimator $\hat{\mathbb{F}}$ was studied in [23], generalizing results by Müller et al. [21] who estimate the error distribution function in the *partially linear regression model* but only for one-dimensional X. The case $\vartheta = 0$ was studied by Müller et al. [22], and by Neumeyer and Van Keilegom [24], who assume heteroscedastic errors.

The components of the regression function in *additive* regression models can be estimated in several ways. Stone [32] uses an additive spline estimator. The backfitting method of Breiman and Friedman [2], and Buja et al. [3], estimates the additive components one by one and iterates this procedure. Orthogonal series estimators for semiparametric regression models are studied by Eubank et al. [11], Andrews [1], Donald and Newey [8], Eubank [10], Li [16] and Delecroix and Protopopescu [6]; for partially linear additive regression models see [23]. Here we use the marginal integration method of Newey [25], Tjøstheim and Auestad [35] and Linton and Nielsen [18]. The method starts with an estimator $\hat{\varrho}$ for a multivariate nonparametric regression function and obtains estimators for the additive components by integrating out all but one of the variables, usually with empirical estimators based on the remaining components of the covariates. Linton [17] uses marginal integration to