Journal of Multivariate Analysis 100 (2009) 1182–1197

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Inference on a regression model with noised variables and serially correlated errors

Jinhong You [a], Xian Zhou [b,*], Li-Xing Zhu [c]

[a] Shanghai University of Finance and Economics, China
[b] Macquarie University, Australia
[c] Hong Kong Baptist University, Hong Kong

ARTICLE INFO

Article history:
Received 3 May 2006
Available online 28 October 2008

AMS 2000 subject classifications:
62G05
62G20
62M10

Keywords:
Regression with noised variables
De-noising
Serially correlated errors
ARMA model
Asymptotic normality
Consistency

ABSTRACT

Motivated by a practical problem, [Z.W. Cai, P.A. Naik, C.L. Tsai, De-noised least squares estimators: An application to estimating advertising effectiveness, Statist. Sinica 10 (2000) 1231–1243] proposed a new regression model with noised variables due to measurement errors. In this model, the means of some covariates are nonparametric functions of an auxiliary variable. They also proposed a de-noised estimator for the parameters of interest, and showed that it is root-n consistent and asymptotically normal when undersmoothing is applied. The undersmoothing, however, causes difficulty in selecting the bandwidth. In this paper, we propose an alternative corrected de-noised estimator, which is asymptotically normal without the need for undersmoothing. The asymptotic normality holds over a fairly wide range of bandwidth. A consistent estimator of the asymptotic covariance matrix under a general stationary error process is also proposed. In addition, we discuss the fitting of the error structure, which is important for modeling diagnostics and statistical inference, and extend the existing error structure fitting method to this new regression model. A simulation study is made to evaluate the proposed estimators, and an application to a set of advertising data is also illustrated.

1. Introduction

Motivated by studying the relationship between awareness and television rating points of TV commercials for Cadbury Dairy Milk chocolate brand, one of the major chocolate brands in England, [1] proposed a new regression model where covariables and response are observed with errors over time (considered as an auxiliary variable). This model is proved to be also useful in marketing science, finance and security analysis. For example, [2] used this model to analyze the risk of stocks. The details of this model are as follows. Let $(\xi_1, \ldots, \xi_p; \eta)$ be variables of interest that satisfy a linear relationship

$$\eta = \xi_1\beta_1 + \cdots + \xi_p\beta_p + z_1\alpha_1 + \cdots + z_q\alpha_q \tag{1.1}$$

with some additional covariates $z_1, \ldots, z_q$, where $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and $\alpha = (\alpha_1, \ldots, \alpha_q)^{\mathrm{T}}$ are unknown parameters to be estimated and the superscript (T) denotes the transpose of a vector or matrix. Measurements of $(\xi_1, \ldots, \xi_p; \eta; z_1, \ldots, z_q)$ are collected over time to yield a data set of $\{(x_{1i}, \ldots, x_{pi}; y_i, z_{1i}, \ldots, z_{qi}), i = 1, \ldots, n\}$ with

$$x_{si} = \xi_s(t_i) + u_{si} \quad \text{for } s = 1, \ldots, p \quad \text{and} \quad y_i = \eta_i + \varepsilon_i, \tag{1.2}$$

* Corresponding address: Department of Actuarial Studies, Macquarie University, Sydney, NSW 2109, Australia.
E-mail address: xzhou@efs.mq.edu.au (X. Zhou).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2008.10.011

where t_i is the *scaled* time for the ith measurement such that $0 \le t_i \le 1$, u_{si} and ε_i are measurement errors. The covariates $(z_{1i}, \ldots, z_{qi})$ are assumed to be observed without error and may include the constant component 1 to represent the intercept in the model. A key ingredient of model (1.1) is that the ξ_s are time dependent. In line with usual regression models, (1.1) and (1.2) may be rewritten as

$$y_i = \xi_1(t_i)\beta_1 + \cdots + \xi_p(t_i)\beta_p + z_{1i}\alpha_1 + \cdots + z_{qi}\alpha_q + \varepsilon_i \quad \text{and} \quad x_{si} = \xi_s(t_i) + u_{si}, \tag{1.3}$$

where $i = 1, \ldots, n$ and $s = 1, \ldots, p$. Model (1.3) can be viewed as a semiparametric measurement error model with an auxiliary variable t.

Using a wavelet de-noising process, [1] proposed a de-noised least squares estimator (DLSE) for the regression coefficients β and α when the errors are independent and identically distributed (i.i.d.). The independence assumption on the errors ε_i, however, is not always appropriate in applications since η is time dependent. Instead, the errors ε_i are often serially correlated and thus should be modeled as a time series. To accommodate the serial correlation in the errors, they further investigated the estimating problem of model (1.3) when the errors are from a moving average process of infinite order, denoted by MA(∞), which has the following form:

$$\varepsilon_i = \sum_{j=0}^{\infty} \phi_j e_{i-j}, \quad \text{with} \sum_{j=0}^{\infty} |\phi_j| < \infty, \tag{1.4}$$

where $\{e_j\}$ are i.i.d. random variables with $E(e_j) = 0$ and $\text{Var}(e_j) = \sigma_e^2 < \infty$. They derived the asymptotic normality of the DLSE by using nonparametric undersmoothing. Undersmoothing is not uncommon in nonparametric estimation. See, for example, [3–6]. However, it introduces difficulty in choosing the bandwidth, as it does not allow a data-driven choice. In order to overcome this problem, we propose in this paper a corrected de-noised estimator which does not need undersmoothing and can achieve an optimal strong convergence rate. Moreover, [2] presented a consistent estimator of the asymptotic covariance matrix only for an AR(1) error process, which is obviously rather restrictive. In this paper, we construct a consistent estimator of the covariance matrix under a general stationary error process. Furthermore, we discuss the estimation of the error structure in the model together with other parameters of interest, which is important to performing model diagnostics and making better statistical inference.

The rest of the paper is organized as follows. In Section 2 we propose a corrected DLS estimator, and derive its asymptotic normality and strong convergence rate. In Section 3, we construct a consistent estimator of the asymptotic covariance matrix under a general stationary error process. Section 4 investigates the fitting of the error structure. As an ARMA model is often a good approximation of the underlying process of the errors, we restrict Section 5 to the ARMA process and suggest a method for determining its order and then estimating the coefficients. Section 6 presents some simulation results and an example of application on a set of advertising data; this is followed by concluding remarks in Section 7. All proofs of theoretical results are postponed to Appendix A.

2. DLS estimation and corrected DLS estimator

In this section we first present the DLSE. We assume that the sequence of designs $\{t_i\}_{i=1}^n$ forms an asymptotically regular sequence [7] in the sense that

$$\int_0^{t_i} p(t)\mathrm{d}t = \frac{i-1}{n-1},$$

where $p(\cdot)$ denotes a positive density function on the interval [0, 1]. Obviously, this implies $\max_{1\le i\le n}(t_i - t_{i-1}) = O(n^{-1})$ with $t_0 = 0$. Like [2] we only de-noise the x variable as the de-noising of y does not enhance the performance of the estimator. It is worthwhile to notice that the results of this paper can be readily extended to cover random as well as multivariate covariate t.

According to (1.2) we can apply the usual nonparametric method to de-noise $\xi(t)$. Although there are several methods available in the literature which can be used to screen out noise, for ease of exposition, we use the Nadaraya–Watson kernel smoother. The de-noised variable $\hat{\xi}_s(t)$ is given by

$$\hat{\xi}_s(t) = \sum_{i=1}^{n} \omega_{ni}(t) x_{si} \quad \text{for } s = 1, \ldots, p \tag{2.1}$$

where $\omega_{ni}(t) = \kappa((t_i - t)/h)/(nh)$, $\kappa(\cdot)$ is a symmetric kernel function and h is the bandwidth.

For notational convenience, let $X = (X_1, \ldots, X_n)^{\mathrm{T}} \in \mathcal{R}^{n\times p}$, $X_i = (x_{1i}, \ldots, x_{pi})^{\mathrm{T}}$, $\hat{\Xi} = (\hat{\Xi}_1, \ldots, \hat{\Xi}_n)^{\mathrm{T}} \in \mathcal{R}^{n\times p}$, $\hat{\Xi}_i = (\hat{\xi}_1(t_i), \ldots, \hat{\xi}_p(t_i))^{\mathrm{T}}$, $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}} \in \mathcal{R}^{n\times q}$, $Z_i = (z_{1i}, \ldots, z_{pi})^{\mathrm{T}}$, $Y = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathcal{R}^n$, $\Xi = (\Xi_1, \ldots, \Xi_n)^{\mathrm{T}} \in \mathcal{R}^{n\times p}$, $\Xi_i = (\xi_1(t_i), \ldots, \xi_p(t_i))^{\mathrm{T}}$, $U = (U_1, \ldots, U_n)^{\mathrm{T}} \in \mathcal{R}^{n\times p}$, $U_i = (u_{1i}, \ldots, u_{pi})^{\mathrm{T}}$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}} \in \mathcal{R}^n$. Also define the following two matrices:

$$\Omega_n = \frac{1}{n}\begin{pmatrix} \Xi^{\mathrm{T}}\Xi & \Xi^{\mathrm{T}}Z \\ Z^{\mathrm{T}}\Xi & Z^{\mathrm{T}}Z \end{pmatrix}, \qquad \hat{\Omega}_n = \frac{1}{n}\begin{pmatrix} \hat{\Xi}^{\mathrm{T}}\hat{\Xi} & \hat{\Xi}^{\mathrm{T}}Z \\ Z^{\mathrm{T}}\hat{\Xi} & Z^{\mathrm{T}}Z \end{pmatrix}.$$