Computational Statistics and Data Analysis 128 (2018) 87–103

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimation and hypothesis test for partial linear multiplicative models

Jun Zhang [a], Zhenghui Feng [b,*], Heng Peng [c]

[a] College of Mathematics and Statistics, Institute of Statistical Sciences, Shenzhen-Hong Kong Joint Research Center for Applied Statistical Sciences, Shenzhen University, 518060, Shenzhen, China
[b] School of Economics and the Wang Yanan Institute for Studies in Economics, Xiamen University, 361005, Xiamen, China
[c] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

ARTICLE INFO

Article history:
Received 9 January 2018
Received in revised form 1 May 2018
Accepted 28 June 2018
Available online 10 July 2018

Keywords:
Kernel smoothing
Local linear smoothing
Model checking
Partial linear models
Variable selection

ABSTRACT

Estimation and hypothesis tests for partial linear multiplicative models are considered in this paper. A profile least product relative error estimation method is proposed to estimate unknown parameters. We employ the smoothly clipped absolute deviation penalty to do variable selection. A Wald-type test statistic is proposed to test a hypothesis on parametric components. The asymptotic properties of the estimators and test statistics are established. We also suggest a score-type test statistic for checking the validity of partial linear multiplicative models. The quadratic form of the scaled test statistic has an asymptotic chi-squared distribution under the null hypothesis and follows a non-central chi-squared distribution under local alternatives, converging to the null hypothesis at a parametric convergence rate. We conduct simulation studies to demonstrate the performance of the proposed procedure and a real data is analyzed to illustrate its practical usage.

1. Introduction

Let $(\boldsymbol{X}, Z, Y)$ be a random vector, and assume that $(\boldsymbol{X}, Z)$ and Y satisfy the following partial linear multiplicative model (PLMM):

$$Y = \exp\left(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X} + g(Z)\right)\epsilon, \tag{1.1}$$

where Y is the response variable, $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$, $Z \in \mathbb{R}^1$, $g(\cdot)$ is an unknown smooth function. Both Y and ϵ considered in model (1.1) are positive variables. The model error ϵ satisfies $E(\ln(\epsilon)) = 0$ and also $E(\epsilon - \epsilon^{-1}|\boldsymbol{X}) = 0$. Model (1.1) is equivalent to $\ln(Y) = \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X} + g(Z) + \ln(\epsilon)$. To make $g(\cdot)$ unique, condition $E(\ln(\epsilon)) = 0$ is used to identify $g(\cdot)$. The latter condition $E(\epsilon - \epsilon^{-1}|\boldsymbol{X}) = 0$ is used for the least relative error estimation (Chen et al., 2010, 2016) of $\boldsymbol{\beta}_0$. $\boldsymbol{\beta}_0$ is unknown and need to be estimated. In this paper, we focus on univariate Z only, although the proposed procedure is directly applicable for multivariate $\boldsymbol{Z}$. However, the extension might be practically less useful due to the curse of dimensionality.

Multiplicative linear regression model is studied by Chen et al. (2010, 2016) based on the fundamental assumptions that the response variable and model error are positive and that the logarithmic transform of the response variable is linear. These models are useful in analyzing financial and biomedical data with positive responses, such as the price of financial assets and body fat indexes. In recent years, various efforts have been made to balance the interpretation of linear models with the flexibility of nonparametric models, because an incorrect model of the regression function can lead to

* Corresponding author.
E-mail address: zhfengwise@xmu.edu.cn (Z. Feng).

https://doi.org/10.1016/j.csda.2018.06.017
0167-9473/

excessive modeling biases and erroneous conclusions. PLMM is a realistic, parsimonious candidate when one believes that the relationship between the response variable and some of the covariates has a parametric form, while the relationship between the response variable and the remaining covariate may not be linear. PLMM enjoys the simplicity properties of multiplicative linear regression model and the flexibility of nonparametric models because it combines both parametric and nonparametric components.

Model (1.1) is very flexible. First of all, by taking a logarithmic transformation, model (1.1) becomes to classical partial linear model (Härdle et al., 2000; Heckman, 1986; Speckman, 1988; Liang and Li, 2009; Liang et al., 1999; Xie and Huang, 2009; Yang et al., 2016; Müller and van de Geer, 2015; Zhu and Ng, 2003; Wang and Jing, 2003; Liang et al., 2012). Such a logarithmic transformation is reasonable because of the theoretical and computational simplicity. Second, it is a generalisation of multiplicative linear regression models or accelerated failure models. When $g(\cdot) \equiv 0$, model (1.1) degenerates to the multiplicative regression models (Chen et al., 2010, 2016). To estimate the parameters in multiplicative regression models, Chen et al. (2010) proposed the least absolute relative error (LARE) estimation by minimising $\sum_{i=1}^{n}(|\epsilon_i^{-1}-1|+|\epsilon_i-1|)$. As noted in Chen et al. (2016), the LARE estimation enjoys the robustness and scale-free property, however, this criterion is unsmooth and its computation is very complicated. Besides, confidence intervals for parameters are not very accurate due to the complexity of asymptotic covariance matrix, which involves the density of the error ϵ. As a remedy, Chen et al. (2016) proposed the least product relative error (LPRE) criterion by minimising $\sum_{i=1}^{n}(|\epsilon_i^{-1}-1|\times|\epsilon_i-1|)$, which is equivalent to minimise $\sum_{i=1}^{n}(\epsilon_i^{-1}+\epsilon_i)$. Because the LPRE criteria is strictly convex and infinitely differentiable, the optimisation procedure is much easier.

In this study, we extend the parametric multiplicative models (Chen et al., 2010, 2016) to a partial linear multiplicative model (PLMM) (1.1) and consider estimation and hypothesis testing for this model. Firstly, we propose the profile least product relative error (PLPRE) estimation for $\boldsymbol{\beta}_0$ and $g(\cdot)$. This is expected to be more efficient than the logarithmic transformation method. The convexity of the LPRE criteria leads to a penalised LPRE method for variable selection based on the recently developed smoothly clipped absolute deviation (SCAD) method (Fan and Li, 2001) to achieve variable selection. Secondly, we consider statistical inference for $\boldsymbol{\beta}_0$ to test whether $\boldsymbol{\beta}_0$ satisfies some linear combinations or not. A Wald-type statistic is proposed and is shown that the limiting distribution under the null hypothesis is a centered chi-squared distribution, and consider a local alternative hypothesis. Moreover, a restricted estimator of $\boldsymbol{\beta}_0$ is proposed under the null hypothesis in terms of its asymptotic properties. Finally, we aim to develop a lack-of-fit test for checking the adequacy of PLMM. A score-type statistic is proposed and is shown to be asymptotically centered normal distributed under null hypothesis. Once the asymptotic variance of this test statistic is estimated, a scaled statistic can be constructed in a quadratic form based on our previous statistic. We show that this scaled statistic has an asymptotic centered chi-squared distribution under the null hypothesis and has a non-central chi-squared distribution under local alternatives, converging to the null hypothesis at a parametric convergence rate. Monte Carlo simulation experiments are conducted to examine the performance of the proposed estimation and test procedure.

This paper is organised as follows. In Section 2, we propose the estimation procedure for the parameters, introduce the algorithms, and present the asymptotic results. We also introduce a penalised estimation to achieve variable selection. In Section 3, we derive a Wald-type test statistic for the testing problem, provide a restricted estimator under the null hypothesis, and obtain its theoretical properties. In Section 4, we develop a score-type test statistic for checking the adequacy of partial linear multiplicative models, and study the theoretical properties of this test statistic. Section 5 presents the results of simulation studies, and Section 6 reports the statistical analysis of real data. All proofs of theorems are given in the online "Supplementary Material".

2. Estimation

2.1. PLPRE for $\boldsymbol{\beta}_0$ and $g(z)$

Suppose $\{\boldsymbol{X}_i, Z_i, Y_i\}_{i=1}^{n}$ is an i.i.d. sample from model (1.1), where $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{pi})^{\mathrm{T}}$. The estimation procedure is summarised as follows. We transform model (1.1) into $\ln(Y) = \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X} + g(Z) + \ln(\epsilon)$. To estimate $g(\cdot)$, we use the local linear smoothing technique and approximate $g(z)$ by $g(z_*) + g'(z_*)(z - z_*)$ in a neighborhood of z. For given $\boldsymbol{\beta}$, the local linear estimator of $(g(z), g'(z))$ is obtained by minimising (2.1) with respect to (b_0, d_0),

$$
\begin{aligned}
&(\hat{g}(z,\boldsymbol{\beta}), \hat{g}'(z,\boldsymbol{\beta})) \\
&= \arg\min_{b_0,d_0} \sum_{i=1}^{n} \left\{\ln(Y_i) - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i - b_0 - d_0(Z_i - z)\right\}^2 K_h(Z_i - z),
\end{aligned}
\tag{2.1}
$$

where $K_h(Z_i - z) = h^{-1}K((Z_i - z)/h)$, $K(\cdot)$ is a kernel function and h is a bandwidth. A direct calculation from (2.1) entails that

$$
\hat{g}(z,\boldsymbol{\beta}) = \frac{T_{n,20}(z,\boldsymbol{\beta})T_{n,01}(z,\boldsymbol{\beta}) - T_{n,10}(z,\boldsymbol{\beta})T_{n,11}(z,\boldsymbol{\beta})}{T_{n,00}(z,\boldsymbol{\beta})T_{n,20}(z,\boldsymbol{\beta}) - T_{n,10}^2(z,\boldsymbol{\beta})},
\tag{2.2}
$$

where $T_{n,l_1l_2}(z,\boldsymbol{\beta}) = \sum_{i=1}^{n} K_h(Z_i - z)(Z_i - z)^{l_1}[\ln(Y_i) - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i]^{l_2}$ for $l_1 = 0, 1, 2$, $l_2 = 0,1$.