Journal of Multivariate Analysis 160 (2017) 105–116

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Model specification test in a semiparametric regression model for longitudinal data

Hyunkeun Cho [a], Seonjin Kim [b,*]

[a] Department of Biostatistics, University of Iowa, 145 Riverside Drive, Iowa City, IA 52242, USA

[b] Department of Statistics, Miami University, 311 Upham Hall, Oxford, OH 45056, USA

ARTICLE INFO

Article history:
Received 14 April 2017
Available online 27 June 2017

AMS 2010 subject classifications:
62G10

Keywords:
Calibrated regression model
Kernel regression
Longitudinal data
Quadratic inference function
Semiparametric regression

ABSTRACT

We propose a model specification test for whether or not a postulated parametric model (null hypothesis) fits longitudinal data as well as a semiparametric model (alternative hypothesis) does. In the semiparametric model, we suppose that a baseline function of time is modeled nonparametrically, while the longitudinal covariate effect is assumed to be a parametric linear model. The existing kernel regression based likelihood ratio tests suffer from computing the likelihood function in the alternative hypothesis, because a specific parametric alternative is not desired. To circumvent this difficulty, we calibrate the semiparametric model to a regression model containing only the parametric parameters, and investigate the quadratic inference function in the calibrated model. The proposed approach yields an asymptotically unbiased parametric regression estimator without undersmoothing the baseline function. This provides us a simple and powerful test statistic that asymptotically follows a central chi-squared distribution with fixed degrees of freedom under the null hypothesis. Simulation studies show that the proposed test is able to identify the true parametric regression model consistently. We have also applied this test to real data and confirmed that the baseline function can be captured by a conjectured parametric form sufficiently well.

1. Introduction

A semiparametric regression model has attracted much attention for longitudinal data [18], where m_i measurements are repeatedly observed from subject $i \in \{1, \ldots, n\}$ over, i.e., for all $j \in \{1, \ldots, m_i\}$,

$$Y_i(t_{ij}) = \alpha(t_{ij}) + X_i(t_{ij})^\top \beta + \epsilon_i(t_{ij}), \tag{1}$$

where $Y_i(t_{ij})$ is the ith subject's response at time t_{ij}, $X_i(t_{ij})$ is a p-dimensional covariate vector at time t_{ij}, $\alpha(t)$ is an unspecified smooth baseline function of time t, β is a p-dimensional parameter vector of interest, and $\epsilon_i(t_{ij})$ satisfies $\mathrm{E}\{\epsilon_i(t_{ij})|X_i(t_{ij})\} = 0$. The model (1) enables us to estimate and draw inference about β parametrically, addressing the influence of time t on the response nonparametrically. Thus, most of literature on the semiparametric longitudinal model has focused on estimation and inference on β; see [3,5,10,12,15,17]. However, statistical inference about $\alpha(t)$ has received little attention.

In this paper, we endeavor to answer a fundamental question of whether a certain parametric regression model fits longitudinal data as well as the semiparametric regression model (1) does. While adopting a smooth function of $\alpha(t)$ in model (1) increases flexibility in estimation, it may also result in the loss of simple interpretation and inference about time

[*] Corresponding author.
E-mail addresses: hyunkeun-cho@uiowa.edu (H. Cho), kims20@miamioh.edu (S. Kim).

http://dx.doi.org/10.1016/j.jmva.2017.06.005

t on the response in the parametric model. This leads us to consider a semiparametric model specification test for the null hypothesis $\mathcal{H}_0 : \alpha(t) \in \{\alpha_\theta(t) : \theta \in \Theta\}$, where $\alpha_\theta(t)$ is a parametric function of time t with a q-dimensional parameter vector θ and Θ is the parametric space against the alternative hypothesis that $\alpha(t)$ is a smooth function.

Typical kernel regression based likelihood ratio tests confront challenges for this hypothesis testing because a specific parametric alternative is not desirable in the semiparametric model (1). Fan et al. [5] and Härdle et al. [6] provide kernel regression based likelihood ratio test statistics for independent data. However, the test statistic proposed by Härdle et al. [6] does not have an asymptotic chi-squared distribution, and thus the bootstrap is used to compute critical values. The generalized likelihood ratio statistic developed by Fan et al. [5] follows an asymptotic chi-squared distribution, yet additional complex computation is required to obtain the degrees of freedom. To the best of our knowledge, there is no semiparametric model specification test based on kernel regression in the literature for longitudinal models. In response to this gap in the literature, we have developed a simple and powerful inference for the semiparametric longitudinal model specification test.

The major difficulty in developing the desired specification test is caused by computing the likelihood function in the alternative. The asymptotic distribution of minus twice the loglikelihood function in this case is not clear because the nonparametric estimation of $\alpha(t)$ is involved. To attenuate this difficulty, following Xue and Zhu [17], we transform the semiparametric model (1) to a regression model including only β by calibrating the model (1) with $\mathrm{E}\{X_i(t_{ij})\}$ and $\mathrm{E}\{Y_i(t_{ij})\}$. This leads to the asymptotically unbiased estimator of β without undersmoothing the estimators of $\mathrm{E}\{X_i(t_{ij})\}$ and $\mathrm{E}\{Y_i(t_{ij})\}$. This bias correction is essential to construct the proposed test statistic because it allows us to compute the quadratic inference function (QIF) under the alternative hypothesis. Moreover, this QIF provides an inference function for model diagnostic and goodness-of-fit tests since it plays a similar role to minus twice the loglikelihood function.

The semiparametric model specification test is developed based on the difference in the QIFs under the abovementioned null and alternative hypotheses. The key idea is to compute the QIF in the calibrated model, which is equivalent to the alternative model (1). We show that the QIF evaluated at the estimator of β in the calibrated model follows an asymptotic chi-squared distribution with $p(d-1)$ degrees of freedom, where d is the number of basis matrices used in the QIF. Similarly, the QIF evaluated at the estimator of θ and β under the null hypothesis is asymptotically chi-squared with $(p+q)(d-1)$ degrees of freedom, where q is the dimension of θ. Consequently, it is shown that the proposed test statistic follows an asymptotically chi-squared distribution and its degrees of freedom is simply the same as $q(d-1)$, a multiplier of the dimension of θ.

The proposed inference procedure is readily implemented in that it does not specify a likelihood function. Another important advantage of the proposed procedure is that standard data-driven methods such as the plug-in method and cross-validation method can be used for bandwidth selection, since undersmoothing the estimators of $\mathrm{E}\{X_i(t_{ij})\}$ and $\mathrm{E}\{Y_i(t_{ij})\}$ is not needed to compute the QIF in the calibrated model. Simulation studies demonstrate that the proposed test statistic delivers an empirical size close to the nominal level, and detects the discrepancy between the null model and the true model successfully. We also apply the proposed model specification test to the real-life longitudinal data and confirm that a postulated parametric regression model is sufficient to fit the data. Moreover, both theoretical and simulation results confirm that the proposed method yields a more efficient estimator of β rather than ignoring the correlation; this is due to accommodating the within-subject correlation commonly existing in the longitudinal data. This efficiency gain is still achieved even though the assumed working correlation structure is specified incorrectly.

The remainder of this paper proceeds as follows. In Section 2.1, we provide the QIF in the calibrated model and investigate the asymptotic properties of the estimated parametric regression coefficients. In the rest of Section 2, we propose the semiparametric model specification test and introduce estimation and inference on the baseline function. We illustrate the methodology through simulation studies in Section 3, and apply the model to a real life data analysis of a HIV study in Section 4. We conclude with a discussion in Section 5. The proofs are placed in Appendix.

2. Methodology

2.1. Quadratic inference function in the calibrated model

In this section, we first provide a way to calibrate the semiparametric model (1) and investigate the quadratic inference function (QIF) in the calibrated model. The calibrating process is necessary in obtaining an asymptotically unbiased estimator of β without undersmoothing nonparametric estimators. In addition, the asymptotic properties of the QIF in the calibrated model plays a critical role in constructing the semiparametric model specification test statistic in Section 2.2.

Since $\mathrm{E}\{Y_i(t_{ij})\} = \alpha(t_{ij}) + \mathrm{E}\{X_i(t_{ij})\}^\top \beta$, the model (1) can be rewritten as

$$Y_i(t_{ij}) - \mathrm{E}\{Y_i(t_{ij})\} = [X_i(t_{ij}) - \mathrm{E}\{X_i(t_{ij})\}]^\top \beta + \epsilon_i(t_{ij}). \tag{2}$$

This calibrated model does not contain the unknown smooth baseline function $\alpha(t_{ij})$, yet it involves two unknown functions $\mathrm{E}\{Y_i(t_{ij})\}$ and $\mathrm{E}\{X_i(t_{ij})\}$. To estimate these functions, two nonparametric regressions are modeled as

$$Y_i(t_{ij}) = m_Y(t_{ij}) + \varsigma_i(t_{ij}), \qquad X_i(t_{ij}) = m_X(t_{ij}) + \zeta_i(t_{ij}),$$

where $m_Y(t_{ij}) = \mathrm{E}\{Y_i(t_{ij})\}$, $m_X(t_{ij}) = \mathrm{E}\{X_i(t_{ij})\}$, and $\varsigma_i(t_{ij})$ and $\zeta_i(t_{ij})$ are errors with mean zero. We estimate $m_Y(t_{ij})$ and $m_X(t_{ij})$ using a kernel smoothing

$$\hat{m}_Y(t_{ij}) = \frac{\sum_{k=1}^n \sum_{\ell=1}^{m_k} Y_k(t_{k\ell}) K_{k\ell}(t_{ij})}{\sum_{k=1}^n \sum_{\ell=1}^{m_k} K_{k\ell}(t_{ij})}, \qquad \hat{m}_X(t_{ij}) = \frac{\sum_{k=1}^n \sum_{\ell=1}^{m_k} X_k(t_{k\ell}) K_{k\ell}(t_{ij})}{\sum_{k=1}^n \sum_{\ell=1}^{m_k} K_{k\ell}(t_{ij})},$$