Computational Statistics and Data Analysis 56 (2012) 2598–2608

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric model for the dichotomized functional outcome after stroke: The Northern Manhattan Study

Huaihou Chen [a,*], Myunghee Cho Paik [a], Mandip S. Dhamoon [b], Yeseon Park Moon [b], Joshua Willey [b], Ralph L. Sacco [c], Mitchell S.V. Elkind [b]

[a] Department of Biostatistics, Mailman School of Public Health, Columbia University, 10032 New York, NY, USA
[b] Department of Neurology, College of Physician and Surgeons, Columbia University, USA
[c] Department of Neurology and Epidemiology and Human Genetics, Miller School of Medicine, Miami University, USA

ARTICLE INFO

Article history:
Received 5 April 2011
Received in revised form 1 February 2012
Accepted 2 February 2012
Available online 9 February 2012

Keywords:
Generalized estimating equation
Kernel method
Regression splines
Semiparametric longitudinal data analysis

ABSTRACT

The Northern Manhattan Study (NOMAS) is a prospective, population-based study. One of the goals of NOMAS is to characterize the functional status of stroke survivors over time after stroke. Based on generalized estimating equation models, previous parametric analysis showed that functional status declines over time and the trajectories of decline are different depending on insurance status. The two trends of functional status may not be linear, which motivates our semiparametric modeling. In this paper, we model the time trend nonparametrically, the associated covariates parametrically and an interaction term between the nonparametric time trend and a covariate. We consider both kernel weighted local polynomial-based and regression spline-based approaches for solving the semiparametric model, and propose a statistic to test for the interaction term. To evaluate the performance of the parametric model in the case of model misspecification, we study the bias and efficiency of the estimators from misspecified parametric models. We find that when the adjusted covariates are independent of the time, and the link function is identity, the estimators for those covariates are asymptotically unbiased, even if the time trend is misspecified. In general, however, under other conditions and nonidentity link, the misspecified parametric estimators are biased and less efficient even when they are unbiased. We compute the ARE and also conduct simulation studies and compare power for testing the adjusted covariate when the time trend is modeled parametrically versus nonparametrically. In the simulation studies, we observe significant gain in power of those semiparametric model-based estimators compared to the parametric model-based estimators in the cases when the time trend is nonlinear.

Published by Elsevier B.V.

1. Introduction

Stroke is the leading cause of serious long-term disability in the United States. In 2005, approximately 2.6% of US adults had suffered from stroke. Stroke causes substantial health-care expenditures. The estimated lifetime cost from an ischemic stroke is about $140,000 per patient. Because of the high prevalence of stroke, the burden of stroke disability is of primary public health importance. The Northern Manhattan Study (NOMAS) includes a population-based, prospective incident ischemic stroke follow-up study, which is designed to determine stroke incidence, outcomes, and risk factors in a multiethnic urban population (Sacco et al., 1998). One of the goals of NOMAS is to characterize the functional status of stroke

* Corresponding author.
E-mail addresses: hc2356@columbia.edu (H. Chen), mp9@columbia.edu (M.C. Paik).

0167-9473/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.csda.2012.02.001

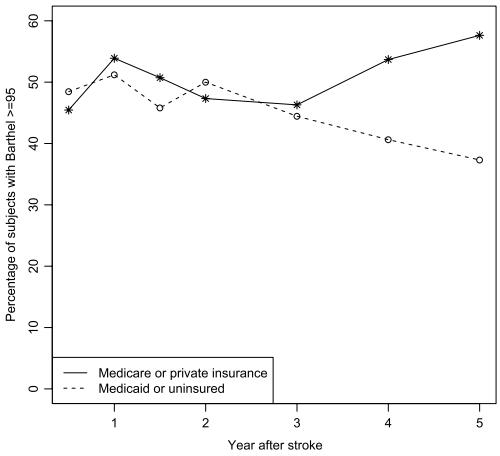

Fig. 1. Percentage of subjects with Barthel index $\geq$95 by insurance status.

survivors over time after stroke. All patients enrolled in the study were scheduled to have semi-annual follow up visits for the first two years after stroke, then annually until 5 years of follow-up. Dhamoon et al. (2009)analyzed functional outcomes among 379 subjects after stroke in the NOMAS cohort. The outcome was a binary indicator of functional independence, defined by Barthel Index greater than or equal to 95. Based on Generalized Estimating Equation (GEE) models, Dhamoon et al. (2009) reported that functional status declines over time, and the trajectories are different depending on insurance status, adjusting for demographics and known risk factors. In addition, the authors noted that the time trends may not be linear. Fig. 1 shows that time trends for the two groups defined by insurance status are different.

In practice, parametric models are parsimonious and useful. However, they may not be sufficient in describing the relationship between the outcome variable and the covariates. Also misspecification of the time trend could result in bias and inefficiency of coefficient estimates for the other covariates in a model. Semiparametric models relate the outcome variable with some covariates parametrically and other covariates nonparametrically, hence semiparametric models keep the flexibility of the nonparametric models for the time trend, and retain the parsimonious property of parametric models as well. Lin and Carroll (2001) proposed a semiparametric model along with the kernel weighted local polynomial method for solving it. They showed that the estimator using working independence is quite efficient. However, in practice, when the within-subject correlation is strong, it will be less efficient if one uses working independence (Wang, 2003; Chen and Jin, 2005). Three more efficient approaches have been proposed: Wang et al. (2005), Chen and Jin (2006) and Huang et al. (2007). The method in Wang et al. (2005) involved residuals from an initial model, for example, the working independence model in Lin and Carroll (2001), therefore, it is labor-intensive to be implemented. Chen and Jin (2006), using piecewise local polynomial to approximate the unknown function, first obtained consistent estimator for the parametric part, then used backfitting for the unknown function. Alternatively, Huang et al. (2007) applied regression splines to approximate the unknown function, and obtained both a smooth fit for the function and consistent estimators for the parametric part simultaneously. All the three aforementioned methods are more efficient than the one in Lin and Carroll (2001), when the within-subject correlation is strong. Among the three, Huang et al. (2007) is easier to be implemented than the other two.

One of our goals is to apply a semiparametric model to reanalyze functional outcome data after stroke to better understand the trajectory after stroke and to efficiently estimate the covariates. Since Fig. 1. shows that the trajectories of the two insurance groups are different, we need to distinguish the two functions in the semiparametric model. First, we consider an extension of the method in Lin and Carroll (2001) by allowing an interaction term between the nonparametric time trend and a variable in parametric part of the model. We show the asymptotic properties of the estimators and propose a statistic to test the interaction term between the time trend and a covariate. We also consider the regression spline-based method in Huang et al. (2007) and extend the model to the time-varying coefficient model.

Another goal is to evaluate bias and efficiency of parametric model in case of model misspecification. Specifically, we investigate the bias of the estimators from the possibly misspecified parametric model. We identify conditions that misspecified time trend still results in unbiased estimator of the covariates and compare their efficiency to the regression splines-based estimators. We find that when the adjusted covariates are independent of the time, and the link function is identity, the estimators for the covariates are asymptotically unbiased, even if the unknown function is misspecified. In general, however, under other conditions and non-identity link, the parametric estimators are biased and less efficient even when they are unbiased. This gives an important message in designing and executing studies in that if measurements are taken in a balanced way between the treatment or exposure groups, the resulting modeling will be robust against misspecification.

The structure of the paper is as follows: Sections 2 and 3 respectively review the parametric and semiparametric methods for longitudinal data analysis. Section 4 presents the semiparametric model with an interaction term, i.e., the time-varying coefficient model, the algorithm for solving it, and a statistic for the hypothesis of no interaction. We also show the