ISSN: 2641-6336
DOI: 10.33552/ABBA.2022.04.000598
Annals of
Biostatistics & Biometric Applications
Iris Publishers

Mini Review

Recent Studies on Regularization Methods in Semiparametric Models with Longitudinal Data

Omidali Aghababaei Jazi*

Department of Mathematical and Computational Sciences, University of Toronto Mississauga

***Corresponding author:** Omidali Aghababaei Jazi, Department of Mathematical and Computational Sciences, University of Toronto Mississauga.

Received Date: November 3, 2022

Published Date: November 18, 2022

Abstract

Longitudinal data arise when an outcome variable is measured on the same subjects on several occasions. Semiparametric regression models are commonly used for modelling of longitudinal data as they provide more flexibility in studying the association between regression factors and a longitudinal outcome. Regularization methods for these models have received much attention over the last two decades in order to reduce the complexity of the models and improve their predictability. In this short article, I will review recent studies on regularization methods in semiparametric models with longitudinal data.

Keywords: Longitudinal data; Regularization; Semiparametric models; Informative follow-up

Introduction

A longitudinal study is a research design that involves repeated observations of the same variables over time. Semiparametric regression models are commonly used for modelling of collected data in a longitudinal study as they provide edibility as well as a balance between parametric and nonparametric models. Statistical methods under semiparametric regression models have been proposed by researchers; see [1,2,3] among others.

A common problem in analyzing such data is how to select relevant variables and estimate their coefficients in semiparametric models for longitudinal data. This will reduce the complexity of models and the vari-ability of estimators and further improve the interpretability and predictability of models. For instance, in treating major depressive disorder, different covariates such as age, gender, socioeconomic status, etc can be considered at baseline. It is important to choose the most relevant covariates in the model because including irrelevant covariates increases variance of parameter estimators and as such reduces their efficiency. The classical variable selection methods such as best subset selection and stepwise deletion suffer from some drawbacks such as computational cost, instability of the selection process, and unknown asymptotic properties of their estimators. Moreover, they cause further challenges in semiparametric regression analysis such as the choice of smoothing parameter for each sub model [4]. Finally, significance tests based on stepwise procedures may lead to greatly inflated Type I error rates; see [5] for further discussion.

To overcome the aforementioned problems, regularization methods have been developed over the past two decades. [6,7] proposed the least absolute shrinkage and selection operator (LASSO) penalty function for linear regression model and the Cox model. [8,9] developed variable selection procedures via nonconcave penalized likelihood in generalized linear models and extended it to the semiparametric Cox proportional hazards

model and the frailty model. They proposed the smoothly clipped absolute deviation (SCAD) penalty function which results in penalized estimators that are un-biased, sparse, and continuous. The authors established asymptotic properties of the resulting penalized estimators and showed that with a proper choice of the tuning parameter, they possess the same asymptotic properties as if the true model were known a priori; this is known as the oracle property. Other major studies include [10-12] among others.

While there is a considerably rich literature on variable selection methods for cross-sectional data, studies for longitudinal data are rather limited due to the challenges posed by incorporating within subjects' correlation. [13] developed a quasi-likelihood information criterion (QIC) analogous to Akaike in-formation criterion (AIC) which relies on a likelihood function while the GEE method does not require a parametric distribution. [14] studied variable selection for nonparametric varying-coe cient models. [15] investigated variable selection for mixed-effects model with continuous responses. Other studies are [16-19].

For semiparametric regression models and cross-sectional data, [20] studied variable se-lection using nonconcave penalized likelihood. [4] proposed two novel estimation methods for longitudinal data with irregular observation times. They further proposed a penalized weighted least squares approach using a quadratic loss between the observed data and the model that involves only the unknown parameter. Their approach, however, does not account for informative observation times that depend on observed outcome values, and focuses on semiparametric linear models. The issue is known as longitudinal data with informative observation times (follow-up) [21]. Standard estimation methods such as generalized estimating equation (GEE) [22] in this situation may lead to biased estimators; see for example [23-28], and [21] proposed estimation methods via weighted estimating equations and joint modelling for semiparametric models, respectively. [29] presented and compared the existing semiparametric methods and proposed some extension of the methods.

An example of such data is the bladder cancer study [30] conducted by the Veterans Administration Cooperative Urological Research Group, in which there are 85 patients. Of which 47 were assigned to the control group and 38 were assigned to thiotepa treatment. The study was conducted over 53 months, and participants in the control group had a physician visit once every three months, while those in the treatment group required a visit with their physician once a month. Figure 1 depicts abacus plot for 30 randomly selected patients from the data. It is apparent that the data is highly irregular and outcome dependent. Each patient which is represented by a line does not have a consistent meeting schedule. Some patients drop out entirely, there are some patients that meet less frequently than required, and some patients that require additional follow-ups. [31] have recently proposed a penalized estimating function procedure that is suitable for estimation methods under semiparametric models for longitudinal data with informative follow-up, whether the estimation is devised by the weighting approach or joint modelling of longitudinal outcome and informative observation times. They demonstrated that the resulting penalized estimators are consistent and with proper choice of the tuning parameter and penalty functions possess the oracle property.

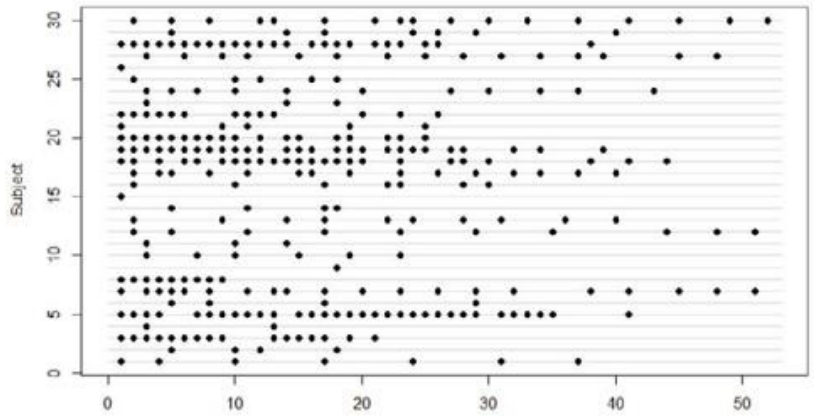

Figure 1: Abacus Plot of bladder cancer study.

Discussion

Longitudinal data that are subject to irregular and informative observation times with large numbers of co-variates are increasingly accessible thanks to the growing availability of administrative data. Regularization methods avoid the drawbacks of the stepwise and best subset selection such as innated Type I error rates and computational cost, and provides statistical practitioners with a flexible and computationally e client tool. Optimization algorithms such as the Minorize-Maximization (MM) algorithm can solve the problem that some penalty functions

may be undefined at zero and further if a variable is deleted in the iterative algorithm, it will be necessarily excluded from the nail model. The current regularization methods can be applied to generalized semiparametric regression models for such as the semiparametric log-linear regression model. Furthermore, the methods can be implemented under other commonly used penalty functions such as Adaptive LASSO and MCP. Moreover, standard methods such as inverse probability of censoring weighting (IPCW) may be adapted to address informative censoring. Finally, the current regularization methods for longitudinal data with irregular and informative follow-up consider axed number of variables in the outcome model. They can be extended to variable selection in both observation times and outcome models as well as high-dimensional case where the number of covariates increases with sample size.

Acknowledgement

None.

Conflict of Interest

No conflict of interest.

References

1. Zeger S L, Diggle P J (1994) Semiparametric models for longitudinal data with application to cd4 cell numbers in HIV seroconverters. Biometrics 50(3): 689-699.
2. Martinussen T, Scheike T H (1999) A semiparametric additive regression model for longitudinal data. Biometrika 86(3): 691-702.
3. Lin D, Ying Z (2001) Semiparametric and nonparametric regression analysis of longitudinal data. Journal of the American Statistical Association 96(453): 103-126.
4. Fan J, Li R (2004) New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. Journal of the American Statistical Association 99(467): 710-723.
5. Heinze G, Dunkler D (2017) Five myths about variable selection. Transplant International 30(1): 6-10.
6. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267-288.
7. Tibshirani R (1997) The lasso method for variable selection in the cox model. Statistics in Medicine, 16(4): 385-395.
8. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96(456): 1348-1360.
9. Fan J, Li R (2002) Variable selection for cox's proportional hazards model and frailty model. Annals of Statistics pp. 74-99.
10. Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American statistical association 101(476): 1418-1429.
11. Johnson B A, Lin D, Zeng D (2008) Penalized estimating functions and variable selection in semiparametric regression models. Journal of the American Statistical Association, 103(482):672-680.
12. Zhang C H (2010) Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics 38(2): 894-942.
13. Pan W (2001) Akaike's information criterion in generalized estimating equations. Biometrics 57(1): 120-125.
14. Wang L, Li H, Huang J Z (2008) Variable selection in nonparametric varying-coe cient models for analysis of repeated measurements. Journal of the American Statistical Association 103(484): 1556-1569.
15. Ni X, Zhang D, Zhang H H (2010) Variable selection for semiparametric mixed models in longitudinal studies. Biometrics 66(1): 79-88.
16. Fu W J (2003) Penalized estimating equations. Biometrics 59(1): 126-132.
17. Dziak J J (2006) Penalized quadratic inference functions for variable selection in longitudinal research.
18. Xu P, Wu P, Wang Y, Zhu L (2010) A GEE based shrinkage estimation for the generalized linear model in longitudinal data analysis. Technical report, technical report, Department of Mathematics, Hong Kong Baptist University
19. Xue L, Qu A, Zhou J (2010) Consistent model selection for marginal generalized additive model for correlated data. Journal of the American Statistical Association 105(492): 1518-1530.
20. Li R, Liang H (2008) Variable selection in semiparametric regression modeling. Annals of statistics 36(1): 261-286.
21. Liang Y, Lu W, Ying Z (2009) Joint modeling and analysis of longitudinal data with informative observation times. Biometrics 65(2): 377-384.
22. Liang K Y, Zeger S L (1986) Longitudinal data analysis using generalized linear models. Biometrika 73(1): 13-22.
23. Lipsitz S R, Fitzmaurice G M, Ibrahim J G, Gelber R, Lipshultz S, et al. (2002) Parameter estimation in longitudinal studies with outcome-dependent follow-up. Biometrics 58(3): 621-630.
24. Fitzmaurice G M, Lipsitz S R, Ibrahim J G, Gelber R, Lipshultz S, et al. (2006) Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. Biostatistics 7(3): 469-485.
25. Chen Y, Ning J, Cai C (2015) Regression analysis of longitudinal data with irregular and informative observation times. Biostatistics 16(4): 727-739.
26. Sun J, Park D H, Sun L, Zhao X (2005) Semiparametric regression analysis of longitudinal data with informative observation times. Journal of the American Statistical Association 100(471): 882-889.
27. Buzkova P, Lumley T (2008) Semiparametric log-linear regression for longitudinal measurements subject to outcome-dependent follow-up. Journal of Statistical Planning and Inference 138(8): 2450-2461.
28. Buzkova P, Lumley T (2009) Semiparametric modeling of repeated measurements under outcome-dependent follow-up. Statistics in medicine 28(6): 987-1003.
29. Tan K S, French B, Troxel A B (2014) Regression modeling of longitudinal data with outcome-dependent observation times: extensions and comparative evaluation. Statistics in medicine 33(27): 4770-4789.
30. Andrews D F, Herzberg A M (2012) Data: a collection of problems from many elds for the student and research worker. Springer Science & Business Media.
31. Aghababaei Jazi O, Pullenayegum E (2022) Variable selection in semiparametric regression models for longitudinal data with informative observation times. Statistics in Medicine 41(17): 3281-3298.