Robust estimators in Generalized Partially Linear Models

G. Boente[1] and Daniela Rodriguez [1]

[1] Universidad de Buenos Aires and CONICET Argentina

Abstract

Semiparametric models contain both a parametric and a nonparametric component. Sometimes the nonparametric component plays the role of a nuisance parameter. The aim of this talk is to consider semiparametric versions of the generalized linear models where the response y is to be predicted by covariates $(\mathbf{x}, t)$, where $\mathbf{x} \in I\!R^p$ and $t \in I\!R$. It will be assumed that the conditional distribution of $y|(\mathbf{x}, t)$ belongs to the canonical exponential family $\exp\left[y\theta(\mathbf{x}, t) - B\left(\theta(\mathbf{x}, t)\right) + C(y)\right]$, for known functions B and C. The generalized linear model (McCullagh and Nelder, 1989), which is a popular technique for modelling a wide variety of data, assumes that the mean is modelled linearly through a known link function, g, i.e.,

$$g(\mu\left(\mathbf{x}, t\right)) = \theta(\mathbf{x}, t) = \beta_0 + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \alpha t \ .$$

In many situations, the linear model is insufficient to explain the relationship between the response variable and its associated covariates. A natural generalization, which suffers from the *curse of dimensionality*, is to model the mean nonparametrically in the covariates. An alternative strategy is to allow most predictors to be modeled linearly while one or a small number of predictors enter the model nonparametrically. This is the approach we will follow, so that the relationship will be given by the semiparametric generalized partially linear model

$$\mu\left(\mathbf{x}, t\right) = \mathrm{E}\left(y|(\mathbf{x}, t)\right) = H\left(\eta(t) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right) \tag{1}$$

where $H = g^{-1}$ is a known link function, $\boldsymbol{\beta} \in I\!R^p$ is an unknown parameter and η is an unknown continuous function.

Severini and Wong (1992) introduced the concept of generalized profile likelihood, which was later applied to this model by Severini and Staniswalis (1994). In this method, the nonparametric component is viewed as a function of the parametric component, and $\sqrt{n}-$consistent estimates for the parametric component can be obtained when the usual optimal rate for the smoothing parameter is used. Such estimates fail to deal with outlying observations. In a semiparametric setting, outliers can have a devastating effect, since the extreme points can easily affect the scale and the shape of the function estimate of η, leading to possibly wrong conclusions on β.

Robust procedures for generalized linear models have been considered among others by Stephanski, Carroll and Ruppert (1986), Künsch, Stefanski and Carroll (1989), Bianco and Yohai (1995), Cantoni and Ronchetti (2001), Croux and Haesbroeck (2002) and Bianco, García Ben and Yohai (2005). The basic ideas from robust smoothing and from robust regression estimation have been adapted to deal with the case of independent observations following a partly linear regression model with $g(t) = t$; we refer to Gao and Shi (1997) and Bianco and Boente (2004), and He, Zhu and Fung (2002).

In this talk, we will first remind the classical approach to generalized partly linear models. The sensitivity to outliers of the classical estimates for these models is good evidence that robust

methods are needed. The problem of obtaining a family of robust estimates was first considered by Boente, He and Zhou (2006). However, their procedure is computationally expensive. We will introduce a general three–step robust procedure to estimate the parameter $\boldsymbol{\beta}$ and the function η, under a generalized partly linear model (1), that is easier to compute than the one introduce by Boente, He and Zhou (2006). It is shown that the estimates of $\boldsymbol{\beta}$ are root–n consistent and asymptotically normal. Through a Monte Carlo study, we compare the performance of these estimators with that of the classical ones. Besides, through their empirical influence function we study the sensitivity of the estimators. A robust procedure to choose the smoothing parameter is also discussed.

We will briefly discuss the generalized partially linear single index model which generalizes the previous one since the independent observations are such that $y_i | (\mathbf{x}_i, t_i) \sim F(\cdot, \mu_i)$ with $\mu_i = H\left(\eta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{t}_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$, where now $\mathbf{t}_i \in I\!R^q$, $\mathbf{x}_i \in I\!R^p$ and $\eta : I\!R \to I\!R$, $\boldsymbol{\beta} \in I\!R^p$ and $\boldsymbol{\alpha} \in I\!R^q$ ($\|\boldsymbol{\alpha}\| = 1$) are the unknown parameters to be estimated. Two families of robust estimators are introduced which turn out to be consistent and asymptotically normally distributed. Their empirical influence function is also computed. The robust proposals improve the behavior of the classical ones when outliers are present.

References

Bianco, A. and Boente, G. (2004). Robust estimators in semiparametric partly linear regression models. *Journal of Statistical Planning and Inference*, **122**, 229-252.

Bianco, A., García Ben, M. and Yohai, V. (2005) . Robust estimation for linear regression with asymmetric errors. *Canadian Journal of Statistics*, **33** 1-18.

Bianco, A. and Yohai, V. (1995). Robust estimation in the logistic regression model. *Lecture Notes in Statistics,* **109**, 17-34. Springer–Verlag, New York.

Boente, G., He, X. and Zhou, J. (2006). Robust Estimates in Generalized Partially Linear Models". *Annals of Statistics*, **34**, 2856-2878.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**, 1022-1030.

Croux, C. and Haesbroeck, G. (2002). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis*,**44**, 273-295.

Gao, J. and Shi, P. (1997). M–type smoothing splines in nonparametric and semiparametric regression models. *Statistica Sinica*, **7**, 1155-1169.

He, X., Zhu, Z. and Fung, W. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, **89**, 579-590.

Künsch, H., Stefanski, L. and Carroll, R. (1989). Conditionally unbiased bounded-influence estimation in general regression models with applications to generalized linear models. *Journal of the American Statistical Association*, **84**, 460-466.

Mc Cullagh, P. and Nelder, J.A. (1983). *Generalized linear models*. London: Chapman and Hall.

Severini, T. and Staniswalis, J. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, **89** , 501-511.

Severini, T. and Wong, W. (1992). Generalized profile likelihood and conditionally parametric models. *Annals of Statistics*, **20**, 1768- 1802.

Stefanski, L., Carroll, R. and Ruppert, D. (1986). Bounded score functions for generalized linear models. *Biometrika*, **73**, 413-424.