Journal of Multivariate Analysis 130 (2014) 141–154

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Semiparametric estimation of a class of generalized linear models without smoothing

Alessio Sancetta *

Department of Economics, Royal Holloway, Egham TW20 0EX, United Kingdom

ARTICLE INFO

Article history:
Received 6 March 2013
Available online 13 May 2014

AMS subject classifications:
62G05
62G08
62G20

Keywords:
Empirical cumulant generating function
Exponential dispersion model
Generalized linear model
Single index model

ABSTRACT

In a generalized linear model, the mean of the response variable is a possibly non-linear function of a linear combination of explanatory variables. When the nonlinear function is unknown and is estimated nonparametrically from the data, these models are known as single index models. Using the relation of generalized linear models with the exponential family model, this paper shows how to use a modified version of the empirical cumulant generating function to estimate the linear function of the explanatory variables with no need of smoothing techniques. The resulting estimator is consistent and normally distributed. Extensive simulations, partially reported here, show that the method works in practice. The method can also be seen as complementary to existing fully nonparametric methods. In fact, it can provide an initial value that can be used to fine tune a nonparametric estimator of the link function in the first step of the estimation.

1. Introduction

Generalized linear models [24] allow the expectation of the response Y given the explanatory variables X to be nonlinear, through what is called the link function, e.g. $\mathbb{E}[Y|X] = G\left(X'\beta\right)$, for some univariate function G, whose inverse is called link function, i.e. $X'\beta = G^{-1}(\mathbb{E}[Y|X])$, as it links a linear function of the predictors to the conditional expectation. In some situations, it is not obvious what G should be. If G is not specified it can then be estimated from the data. Then, one calls this semiparametric model the single index model.

The generalized linear model makes direct reference to the exponential family model and the exponential dispersion family model [17,18]. On the other hand, the single index model makes reference to neither the specific functional form of G nor to the distribution of the errors, hence it is more general.

The literature on estimation of single index models abound. One approach is the average derivatives method, where one exploits the fact that

$$d\mathbb{E}[Y|X = x]/dx = dG\left(x'\beta\right)/dx \propto \beta,$$

and the prime $'$ stands for transposition. This requires a high dimensional kernel smoother and consequently is subject to the so called curse of dimensionality ([26,13], see [16], for an improved method and references therein). Another approach is to estimate G nonparametrically based on some initial estimate of β and then estimate β using the estimator for G, cycling through the procedure until convergence (e.g. [12,15,29,4,6], and references therein). One of such nonparametric models

* Corresponding author.
E-mail address: asancetta@gmail.com.
URL: http://sites.google.com/site/wwwsancetta/.

http://dx.doi.org/10.1016/j.jmva.2014.04.025
0047-259X/

is the Estimating Function Method (EFM) approach of Cui et al. [4]. This method achieves the same if not smaller variance than the estimator in [2]. For the EFM and other approaches, the first guess of β can be crucial for convergence to a global maximum. The problem is made even harder by the fact that the initial amount of smoothing used to estimate G strongly depends on the starting value of β. This initial problem could be avoided if one had a reasonably good estimate of the index parameter that does not require previous estimation of G based on some fully nonparametric approach. Recently, Fan et al. [6] have considered estimation of the quantile regression for the single index model in the presence of large number of regressors via penalization, essentially incorporating variable selection into the kernel smoothing estimation.

The goal of this paper is to impose the semiparametric restriction that the density of Y conditional on X belongs to the exponential dispersion family model with canonical link, and use this to estimate β. The estimation takes advantage of the fact that – under the aforementioned restrictions – the only infinite dimensional parameter is related to the conditional cumulant generating function of the response variables. Direct estimation of this would require nonparametric methods. However, this paper shows that it is possible to find a particular relation between the conditional mean and the unconditional expectation of some known function of the data. To the author knowledge this relation is new. Estimating unconditional expectations of known functions does not require any smoothing. Hence, in this context, the estimation of β can be turned into a nonlinear least square problem and estimated by Generalized Method of Moments (GMM). The resulting estimator is shown to be normally distributed. The method is applicable to continuous and binary dependent variables.

The next section presents the relation between the conditional mean and variance of the response and the unconditional expectation of some function of the data. This relation is the motivation for the estimator. Having defined the estimator, the asymptotic properties are derived under regularity conditions. Section 3 contains a discussion of the results and the conditions. The proofs are deferred to Section 4.

2. Statement of the problem

For some $\lambda > 0$, let P_λ be a probability measure with cumulant generating function $\lambda\psi(t) = \ln\left(\int e^{yt} dP_\lambda(y)\right)$ supposed to be finite for $t \in \mathcal{T}$ and $\mathcal{T}$ is some set containing the origin (called the effective domain of ψ, e.g. [18]). Then,

$$\frac{dP_\lambda(y|\eta)}{dP_\lambda(y)} = \exp\{\lambda(\eta y - \psi(\eta))\} \tag{1}$$

is a density in the exponential dispersion family with respect to (w.r.t.) the dominating measure P_λ. The family is very large as it is essentially defined through any probability measure P_λ having a finite moment generating function around the origin. Hence, the parameter space can be restricted to be the set of values $\eta \in \mathbb{R}$ and $\lambda > 0$ for which $\psi(\eta)$ is finite, and $\lambda\psi(\bullet)$ is the cumulant generating function of some P_λ. Throughout it is assumed that λ and η are inside the parameter space, assumed to be nonempty, so that $\lambda\psi(t)$ is always finite.

Here, interest is restricted to the canonical parameter $\eta := x'\beta$, for some explanatory variable $x \in \mathcal{X} \subseteq \mathbb{R}^K$ and a conformable vector β. This shall be a maintained condition throughout the paper. In its full generality, the exponential dispersion model assumes the canonical parameter to be a possibly nonlinear function of $x'\beta$. As discussed in [25,24], for $\eta = x'\beta$, (1) is a subset of the generalized linear model such that, given a sample $\{Y_i, X_i : i = 1, 2, \ldots, n\}$, a sufficient statistic for β is given by $\sum_{i=1}^n X_i Y_i$. Here, interest is restricted to this case only, where however λ (>0) is unrestricted. The effective domain of ψ implicitly define restrictions on x and β via η.

Example 1. Consider $\psi(\eta) = \eta^2/2$ and set $\lambda = \sigma^{-2}$ for some $\sigma^2 \in (0, \infty)$, so that the exponential dispersion model is the linear Gaussian model

$$\exp\left\{\frac{1}{\sigma^2}\left(yx'\beta - \frac{(x'\beta)^2}{2}\right)\right\} P_\lambda(y)$$

where $P_\lambda(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\}$. Then, $\{\eta \in \mathbb{R} : \psi(\eta) < \infty\} = \mathbb{R}$ so that the only restriction on x and β is that $x'\beta \in \mathbb{R}$.

In the above example, the parameters are essentially unrestricted. This is often not the case.

Example 2. Let $\psi(\eta) = -\ln(-\eta)$ and $\lambda > 0$ so that the exponential dispersion model is the gamma model

$$\exp\left\{\lambda\left(yx'b + \ln\left(-x'b\right)\right)\right\} P_\lambda(y)$$

where $P_\lambda(y) = \exp\{(\lambda - 1)\ln(\lambda y) + \ln\lambda - \ln\Gamma(\lambda)\}$, and $\Gamma(\lambda)$ is the gamma function. Hence, the model is defined for $\eta < 0$ only in order to make sure that $\psi(\eta) < \infty$. In this case, it is convenient to reparametrize in terms of $\tilde{b} = -b$ so that $\eta < 0$ is for example satisfied restricting x and $\tilde{b}$ to have only positive entries.

Another implication is that the restriction on η does restrict the distribution of the regressors when they are stochastic, or their range of values when deterministic. In the Gaussian example, X can take values in $\mathbb{R}^K$, but its distribution needs to be tight to avoid infinities.