Journal of Econometrics 156 (2010) 322–336

Contents lists available at ScienceDirect

Journal of Econometrics

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

# Additive cubic spline regression with Dirichlet process mixture errors☆

Siddhartha Chib[a,*], Edward Greenberg[b]

[a] Olin Business School, Washington University in St. Louis, St. Louis MO 63130, United States

[b] Department of Economics, Washington University in St. Louis, St. Louis MO 63130, United States

ARTICLE INFO

*Article history:*
Received 5 May 2008
Received in revised form
23 October 2009
Accepted 7 November 2009
Available online 22 November 2009

*Keywords:*
Additive regression
Bayes factors
Cubic spline
Non-parametric regression
Dirichlet process
Dirichlet process mixture
Marginal likelihood
Markov chain Monte Carlo
Metropolis–Hastings
Model comparison
Ordinal data

ABSTRACT

The goal of this article is to develop a flexible Bayesian analysis of regression models for continuous and categorical outcomes. In the models we study, covariate (or regression) effects are modeled additively by cubic splines, and the error distribution (that of the latent outcomes in the case of categorical data) is modeled as a Dirichlet process mixture. We employ a relatively unexplored but attractive basis in which the spline coefficients are the unknown function ordinates at the knots. We exploit this feature to develop a proper prior distribution on the coefficients that involves the first and second differences of the ordinates, quantities about which one may have prior knowledge. We also discuss the problem of comparing models with different numbers of knots or different error distributions through marginal likelihoods and Bayes factors which are computed within the framework of Chib (1995) as extended to DPM models by Basu and Chib (2003). The techniques are illustrated with simulated and real data.



## 1. Introduction

Our objective in this article is to specify and estimate flexible Bayesian regression models for continuous and categorical outcomes. By flexible we mean models that are relatively free of assumptions about the functional forms through which covariates affect the response and of assumptions about the distribution of the unobserved error or, in the case of categorical outcomes, the distribution of the underlying latent data. Flexibility in the choice of such assumptions is especially desirable in fields such as biostatistics and the social sciences, where theory rarely provides guidance either about the form of the covariate effects, other than the presumption that the effects are smooth in the covariates, or about the error distribution.

The regression function has been modeled in several ways (see, for example, O'Hagan (1978); Angers and Delampady (1992); Müller et al. (1996); Chipman et al. (1997); Clyde et al. (1998); Vannucci and Corradi (1999)). In this article, we assume that the regression function is additive with each function of the covariates modeled as a cubic spline (for example, Härdle, 1990; Green and Silverman, 1994; Pagan and Ullah, 1999; Li and Racine, 2006). In this approach, it is necessary to specify a set of basis functions for the cubic spline. In the practice to date (for example, Congdon, 2007, chap. 4; Denison et al., 2002; Ruppert et al., 2003, chap. 16) attention has been restricted to the truncated power series basis and polynomial *B*-spline basis. In each case, the parameters of the basis functions have no easy interpretation. From the Bayesian viewpoint, the lack of interpretability of the coefficients is inconvenient because it hinders the construction of proper prior distributions that can be motivated by defensible a priori reasoning. The common strategy of specifying improper or default prior distributions is not satisfactory if the goal is to compare alternative non-parametric formulations through such formal means as marginal likelihoods and Bayes factors.

One innovation of this article is in the use of a relatively unexplored basis in which the spline coefficients have the attractive feature of being the unknown function ordinates at the knots. We exploit this feature to develop a proper prior distribution on the coefficients that involves the first and second differences of the ordinates, quantities about which one may be expected to have some prior knowledge. We also indicate how a simulation-based approach can be used to specify the hyperparameters of our

☆ We thank the editor and four referees for their many useful comments on the article.

* Corresponding author. Tel.: +1 314 935 4657.
*E-mail addresses:* chib@wustl.edu (S. Chib), edg@artsci.wustl.edu (E. Greenberg).

0304-4076/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2009.11.002

prior distribution. The basis we employ is described in Lancaster and Šalkauskas (1986, secs. 3.7 and 4.2). We, henceforth, refer to it as the LS basis. This basis is mentioned briefly by Wood (2006, Sec. 4.1.2) in relation to cubic regression splines, but it is not used in the computations. Other bases parameterized in terms of ordinate values appear in Poirier (1973) and Green and Silverman (1994), but without any connection to Bayesian problems or to the issues that concern us in this article.

In the spline literature, the error distribution is usually assumed to be parametric, especially when the outcome is not continuous. In ordinal models, which we use as the running example for categorical outcomes, the model is almost always specified with logit or probit links. One of our goals is to show that it is not difficult to move beyond such parametric assumptions. In our development, we assume that the distribution of the error is an unknown parameter that we model as a Dirichlet process mixture (DPM). The DPM is a general family of prior distributions on probability measures that was introduced by Ferguson (1973) and Antoniak (1974) and has found many applications in statistics and econometrics. Some of the early uses of this specification in economics are Tiwari et al. (1988), Hirano (2002), and Chib and Hamilton (2002). In modeling the error distribution for continuous outcomes, we assume that, conditioned on an unknown location and variance, the error distribution is normal. We then assume that the location and variance of this normal distribution have an unknown distribution that is modeled as a Dirichlet process, leading to an error distribution that is an arbitrary location–variance mixture of normal distributions. A key property of this DPM specification, established by Ferguson (1983), is that it is capable of approximating any unknown distribution. An alternative approach, which we do not pursue, could involve the weighted mixtures of Dirichlet processes (Dunson et al., 2007). For ordinal outcomes, where the ordered category probabilities are defined in terms of an increasing sequence of cut-points, we invoke a version of the DPM model in which mixing occurs only over the variance parameter because in that case the unknown location is confounded with the cut-points.

An important characteristic of our models is that they are easy to understand and estimate. The models can be expressed in the form of a linear regression for observed outcomes in the case of continuous outcomes and for latent outcomes in the case of ordinal outcomes. The predictors in these regressions are derived from the data on the underlying covariate and our basis functions for the cubic spline. The coefficients are the unknown function ordinates at the various knots. The derivative of each function with respect to its covariate is easily calculated. As far as estimation is concerned, the posterior distribution of the model parameters and other unknowns can be summarized by relatively straightforward Markov chain Monte Carlo (MCMC) methods. For continuous outcomes, conditioned on the parameters of the DPM process, the set-up is similar to a Gaussian heteroskedastic regression model, which simplifies several sampling steps. Similarly, conditioned on all the other unknowns and the data, the sampling of the DPM parameters is done according to the methods of Escobar and West (1995) and MacEachern and Müller (1998). The fitting is completed by steps in which the unknown smoothness parameters are sampled. The MCMC sampling algorithm for the ordinal model is similar in the latent variable framework of Albert and Chib (1993). It differs from the continuous model because the posterior distribution includes the cut-points and the latent variables that are introduced to model the ordinal outcomes.

A major focus of our work is on the comparison of different versions of our models (defined, for example, by alternative covariates, fewer or additional knots, or parametric assumptions about the error distribution). For this purpose, we discuss the computation of marginal likelihoods and Bayes factors. We provide algorithms for computing the marginal likelihood for both the continuous and ordinal models within the framework of Chib (1995) as extended to DPM models by Basu and Chib (2003). These algorithms are not complicated and require virtually the same code that is used in the fitting of the models.

Our article can be viewed as a contribution to an emerging literature on flexible Bayesian regression models. For instance, Leslie et al. (2007) pursue similar objectives but in the context of regression splines and with a different basis than ours. The article does not consider the question of the prior on the spline coefficients or the computation of the marginal likelihood. Griffin and Steel (2007) analyze a new Dirichlet process regression smoother in which the functional form for the covariate structure is centered over a class of regression models rather than taking the form of a spline. Finally, Geweke and Keane (2007) and Villani et al. (2007) consider a Bayesian regression model in which the error distribution is modeled by a discrete mixture of normal variables. The mean function in the former is modeled by general quadratic, cubic, and quartic polynomials in two covariates, while splines are used in the latter. These two articles primarily focus on time series problems and do not tackle the question of Bayes factors for model comparisons. None of these four contributions extend their methods to models of ordinal outcomes.

The rest of the article is organized as follows. In Section 2, we present the basic models. Section 3 contains the cubic spline basis for modeling the unknown covariate functions and introduces the identifying restrictions. We specify the prior distribution for the parameters in Section 4 and, in Section 5, develop the prior-posterior analyses of the models and show how the posterior distribution of the unknowns can be summarized by MCMC methods. The computation of the marginal likelihood is considered in Section 6. Section 7 deals with some special cases. Examples with simulated and real data are contained in Section 8. Section 9 has our conclusions.

## 2. Models

### 2.1. Continuous outcomes

Assume that $y_i$ is the $i$th observation in a sample of $n$ observations $y = (y_1, \ldots, y_n)$ and that the model generating $y_i$ depends on a $k_0$-vector of covariates $x_{i0}$, consisting of an intercept and nominal variables, and $q$ additional covariates $w_{i1}, \ldots, w_{iq}$. Now, let

$$y_i = x_{i0}'\beta_0 + g_1(w_{i1}) + \cdots + g_q(w_{iq}) + \varepsilon_i, \quad i \le n, \tag{2.1}$$

where the $g_j(\cdot)(j \le q)$ are unknown functions, and the error $\varepsilon_i$ is independent of the covariates. Thus, in this model, the covariates $x_{i0}$ are assumed to have a parametric effect on the expected value of the response, and the $w_{ij}$ are assumed to enter the model nonparametrically.

The distribution of the error is assumed to be a DPM. Although other non-parametric formulations of the error distributions are possible, the DPM specification has the strengths of being both parsimonious and tractable. Formally, conditioned on an unknown location $\mu_i$ and positive variance $\sigma_i^2$, we assume that the error distribution is normal $\mathrm{N}\left(\mu_i, \sigma_i^2\right)$. We then suppose that $\phi_i = (\mu_i, \sigma_i^2)$ has an unknown probability measure $G$ over $((-\infty, \infty) \times (0, \infty), \mathcal{B} \times \mathcal{B}_+)$, where the prior on $G$ is given by the Dirichlet process (Ferguson, 1973) with concentration parameter $\alpha$ and base distribution $G_0$. Marginalized over $\phi_i$ and $G$, it follows that the error distribution is an arbitrary location–variance mixture of normal distributions. In particular, we assume that

$$\begin{aligned}
&\varepsilon_i|\phi_i \sim \mathrm{N}\left(\mu_i, \sigma_i^2\right)\\
&\phi_i|G \sim G\\
&G \sim \mathrm{DP}(\alpha G_0),
\end{aligned}$$