Statistical Methodology 18 (2014) 14–31

Contents lists available at ScienceDirect

Statistical Methodology

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

Partially linear censored regression models using heavy-tailed distributions: A Bayesian approach

Luis M. Castro [a], Victor H. Lachos [b,*], Guillermo P. Ferreira [a], Reinaldo B. Arellano-Valle [c]

[a] *Departamento de Estadística, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Concepción, Chile*

[b] *Departamento de Estatística, IMECC, Universidade Estadual de Campinas, Campinas, Sao Paulo, Brazil*

[c] *Departamento de Estadística, Facultad de Matemática, Pontificia Universidad Católica de Chile, Santiago 22, Chile*

ARTICLE INFO

Article history:
Received 4 June 2012
Received in revised form
4 October 2013
Accepted 11 October 2013

Keywords:
Bayesian modeling
Censored regression models
Nonlinear regression model
Scale mixtures of normal distributions

ABSTRACT

Linear regression models where the response variable is censored are often considered in statistical analysis. A parametric relationship between the response variable and covariates and normality of random errors are assumptions typically considered in modeling censored responses. In this context, the aim of this paper is to extend the normal censored regression model by considering on one hand that the response variable is linearly dependent on some covariates whereas its relation to other variables is characterized by nonparametric functions, and on the other hand that error terms of the regression model belong to a class of symmetric heavy-tailed distributions capable of accommodating outliers and/or influential observations in a better way than the normal distribution. We achieve a fully Bayesian inference using pth-degree spline smooth functions to approximate the nonparametric functions. The likelihood function is utilized to compute not only some Bayesian model selection measures but also to develop Bayesian case-deletion influence diagnostics based on the q-divergence measures. The newly developed procedures are illustrated with an application and simulated data.

* Corresponding author. Tel.: +55 1932018420.
E-mail addresses: luiscastroc@udec.cl (L.M. Castro), hlachos@gmail.com, hlachos@ime.unicamp.br (V.H. Lachos), gpferrei@uc.cl (G.P. Ferreira), reivalle@mat.puc.cl (R.B. Arellano-Valle).

1572-3127/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.stamet.2013.10.003

1. Introduction

The problem of estimating regression models where the dependent variable is censored has been studied in different fields, namely, econometric analysis, clinical assays and environmental analysis among others. For example, in econometrics the study of the labor force participation of married women is usually conducted under the normal censored regression model known as Tobit model (see [49,23]). In this case, the observed response is the wage rate. This variable is typically considered as censored below zero, *i.e.*, for working women positive values of the wage rates are registered whereas for the non-working women the observed wage rates are zero (see [13,36]).

In the statistical literature, the normal censored regression model corresponds to a linear regression model where the response variable of theoretical interest, Y, is censored. Instead, we observe a dependent variable Y^* defined by

$$Y_i^* = \begin{cases} c_i, & Y_i \le c_i, \\ \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, & Y_i > c_i, \end{cases} \tag{1}$$

$i = 1, \ldots, n$, where $\boldsymbol{\beta}$ is a vector of regression parameters, $\mathbf{x}_i$ is a vector of explanatory variables, the error terms $\epsilon_i,\ i = 1, \ldots, n$, are assumed to be independent and normally distributed with zero mean and a common variance parameter σ_ϵ^2, and c_i is the censoring level. For simplicity, in this paper we assume that the response variable is left-censored, however extensions to arbitrary censoring are immediate.

There are several proposals of censored models in the literature. For example, Blundell and Meghir [4] discussed some generalizations of the normal censored model (known as Tobit model) that allow, for distinct processes, determining the censoring rule and the continuous observations. Alternatively, semiparametric censored models such as the binary response model, the ordered response model, the grouped dependent model and the multinomial response model, among many others, can be found in [39]. Recently, Hutton and Stanghellini [25] proposed a censored regression model assuming a skew-normal distribution to study health care interventions and Yue and Hong [50] considered the use of the Tobit model in quantile nonlinear regression with asymmetric Laplace errors for the study of medical expenditures.

However, symmetric extensions of the censored normal model can be obtained by assuming *e.g.* that the distribution of the perturbations belongs to the scale mixture of normal (SMN) distributions family (see [1]), from which the normal model can be obtained as a special case. In fact, Zhou and Tan [52] proposed a Tobit factor analysis with multivariate Student-*t* distribution whereas Arellano-Valle et al. [2] proposed an extension of the Tobit model considering that the error term follows a Student-*t* distribution. In those papers, the authors provided useful extensions of the normal censored model for statistical modeling of datasets involving observed variables with heavier tails than the normal distribution.

Consequently, our proposal deals with an extension of the normal censored regression model (1) by considering both a parametric linear term and a nonparametric component as follows

$$Y_i^* = \begin{cases} c_i, & Y_i \le c_i, \\ \mathbf{x}_i^\top \boldsymbol{\beta} + g(t_i) + \epsilon_i, & Y_i > c_i, \end{cases} \tag{2}$$

$i = 1, \ldots, n$, where $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are given as in (1), t_i is a known scalar, g is an unknown smooth function and ϵ_i is the error term. Model (2) is known as partial linear model[1] since it includes as special cases the linear regression model (without the nonparametric component) and the usual nonparametric regression model (without the regression parameters). This type of model is useful in situations where the response variable is linearly related to some of the covariates and, at the same time, depends on other covariates in a nonlinear way.

For model (2) we also assume that the error term ϵ_i belongs to the class of SMN distributions. We consider the use of the SMN class of distribution because it is a class of thick-tailed distributions

[1] A complete review of these types of models can be found in [42].