PARTIALLY LINEAR CENSORED REGRESSION MODELS USING HEAVY-TAILED DISTRIBUTIONS: A BAYESIAN APPROACH

LUIS M. CASTRO[1]*,
VICTOR H. LACHOS[2],
GUILLERMO P. FERREIRA[3],
REINALDO B. ARELLANO-VALLE[4]

[1] *Department of Statistics, Universidad de Concepción, Chile* (luiscastroc@udec.cl).

[2] *Department of Statistics, Universidade Estadual de Campinas, Brazil* (hlachos@ime.unicamp.br)

[3] *Department of Statistics, Universidad de Concepción, Chile* (gferreir@udec.cl)

[4] *Department of Statistics, Pontificia Universidad Católica de Chile, Chile* (reivalle@mat.puc.cl)

Abstract

Partial linear models are usually considered semiparametric models since they contain both a parametric linear term and a nonparametric component. In fact, in these type of models it is assumed that the mean response of interest is linearly dependent on some covariates, whereas its relation to other additional variables are characterized by non-parametric functions. Clearly, these models are 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, *e.g.*

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + g(t_i) + \epsilon_i, \tag{1}$$

$i = 1, \ldots, n$, where $\mathbf{x}_i$ is a vector of explanatory variables, $\boldsymbol{\beta}$ are the regression parameters, t_i is a known scalar, g is an unknown smooth function and ϵ_i is the error term.

These models have become an important tool in modeling econometric and biometric data and many authors have been proposed the estimation of parameters $\boldsymbol{\beta}$ and g in various context including kernel smoothing, smoothing splines and penalized splines from the classical and Bayesian points of view.

*This work is partially supported by Grant FONDECYT N°1110076 by the Chilean Government.

On the other hand, the problem of estimation for a regression model where the dependent variable is censored, has been studied in different fields, namely, econometric analysis, clinical essays, among many others. Most of the results on the normal regression model with a censored response variable are based on the Tobit model (Tobin, 1958), where the variable of interest Y_i, for $i = 1 \ldots, n$, is censored. Instead of this variable, we may observe a dependent variable Y_i^{o} given by $Y_i^{\mathrm{o}} = Y_i \mathbb{I}(Y_i > a)$, for some constant a, where $\mathbb{I}(\cdot)$ is the indicator function. The Tobit model corresponds to the censored linear regression model defined by

$$Y_i^{\mathrm{o}} = d_i Y_i \quad \text{and} \quad Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{2}$$

$i = 1, \ldots, n$, where $d_i = \mathbb{I}(Y_i > 0)$, $\boldsymbol{\beta}$ and $\mathbf{x}_i$ are defined in (1) and the error terms ϵ_i, $i = 1, \ldots, n$, are assumed to be independent and normally distributed, with zero mean and a common variance parameter σ_ϵ^2. Tobin (1958) focused on the estimation of the parameters $\boldsymbol{\beta}$ and σ_ϵ^2, on the basis of $n = n_0 + n_1$ observations $(d_1 y_1, \mathbf{x}_1^\top), \ldots, (d_n y_n, \mathbf{x}_n^\top)$, where n_0 and n_1 are the number of observations on the sets $N_0 = \{i : d_i = 0\} = \{i : y_i = 0\}$ and $N_1 = \{i : d_i = 1\} = \{i : y_i > 0\}$, respectively. From the relations mentioned above, the likelihood function for a random sample under the Tobit model is

$$L_N(\boldsymbol{\beta}, \sigma_\epsilon^2) = \prod_{i=1}^{n} \left[1 - \Phi\left(\frac{1}{\sigma_\epsilon}\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right]^{1-d_i} \left[\frac{1}{\sigma_\epsilon}\phi\left(\frac{1}{\sigma_\epsilon}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\right)\right]^{d_i}, \tag{3}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function (*pdf*) and cumulative density function (*cdf*) of a $\mathrm{N}(0, 1)$ distribution, respectively.

There are several extensions of Tobit model in the literature. For example, Blundell and Meghir (1987) discussed some generalisations of the Tobit model that allow for distinct processes determining the censoring rule and the continuous observations. Alternatively, semi-parametric censored models such as the binary response model, the ordered response model, the grouped dependent model, the multinomial response model among many others can be found in Powell (1994). Recently, Hutton and Stanghellini (2011) proposed a censored regression model assuming a skew-normal distribution in order to study health care interventions.

However, natural extensions for the Tobit model can be obtained by assuming *e.g.* that the distribution of the perturbations belong to the scale mixture of normal distributions

family (see Andrews and Mallows, 1974), from which the normal model can be obtained as a special case. From the classical point of view, Zhou and Tan (2009) proposed the Tobit factor analysis with multivariate Student-t distribution and Arellano-Valle et al. (2012) proposed an extension of the Tobit model considering that the error term follows a Student-t distribution. In these papers, the authors provided a useful extension of the Tobit model for statistical modeling of data sets involving observed variables with heavier tails than the normal distribution. Thus, in our paper, an extension of the normal censored regression model (2) is proposed, by considering both a parametric linear term and a nonparametric component and assuming that the error term belongs to the class of scale mixture of normal distributions. This class of distributions constitutes a class of thick-tailed distributions, some of which are the Student-t, slash and the contaminated normal distributions. The estimation and the study of the associated properties of the model is conducted under a Bayesian paradigm.

After fitting the model, it is important to check the model assumptions and conduct a sensitivity analysis in order to detect possible influential or extreme observations that can cause distortions on the results of the analysis. Following the pioneering work by Cook (1986), case-deletion and local influence diagnostics have been widely applied to many regression models in order to assess the effect of perturbations in the model and/or the data on the parameter estimates. Barros et al. (2010) have applied these methods to the normal Tobit model. However, to the best of our knowledge, there are neither studies on Bayesian inference for the Tobit model in the scale mixture of normal family and nor on Bayesian influence diagnostics related to this topic. Thus, in this paper, we discuss influence diagnostic analysis from a Bayesian perspective where the objective is to develop diagnostic measures based on the q-divergence as proposed by Peng and Dey (1995).

The plan of the paper is organized as follows. First, we present the scale mixture of normal distributions. Then, we present the semiparametric extension of the Tobit model and some of its properties. Next, we present the Bayesian implementation of the model, specifying priors distributions for the parameters of interest and the steps of the proposed MCMC algorithm. After that, we present the model selection and influence diagnostics. An application with a real data set of housewives wages is presented in the Section Application. Next, we present a simulation study in order to illustrate the performance of the proposed methodology. Finally, the paper closes with some conclusions.

Acknowledgments

L.M. Castro acknowledges funding support by Grant FONDECYT 11100076 from the Chilean government. The research of V.H.Lachos was supported by Grant 308109/2008-2 from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brazil) and Grant 2011/17400-6 from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP-Brazil). The research of R.B. Arellano-Valle was supported by Grant FONDECYT 1085241 from the Chilean government.

***K*eywords:** Bayesian modeling; Limited dependent variable; Non-linear regression model; Scale mixtures of normal distributions; Tobit model.

References

Andrews, D. and C. Mallows (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B 36*, 99–102.

Arellano-Valle, R., L. Castro, González-Farías, and K. Muñoz Gajardo (2012). Student-t censored regression model: properties and inference. *Statistical Methods and Applications DOI: 10.1007/s10260-012-0199-y.*

Barros, M., M. Galea, M. González, and V. Leiva (2010). Influence diagnostics in the tobit censored response model. *Statistical Methods and Applications 19*, 379–397.

Blundell, R. and C. Meghir (1987). Bivariate alternatives to the tobit model. *Journal of Econometrics 34*, 170–200.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B, 48*, 133–169.

Hutton, J. and E. Stanghellini (2011). Modelling bounded health scores with censored skew-normal distributions. *Statistics in Medicine 30*(4), 368–376.

Peng, F. and D. K. Dey (1995). Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics 23*, 199–213.

Powell, J. (1994). *Estimation of Semiparametric Models.*, pp. 5307–5368. Handbook of Econometrics, Amsterdam: Elsevier 6B.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*, 24–36.

Zhou, X. and C. Tan (2009). Maximum likelihood estimation of tobit factor analysis for multivariate t-distribution. *Communications in Statistics-Simulation and Computation 39*(1), 1–16.