Semiparametric inference for the scale-mixture of normal partial linear regression model with censored data

Mehrdad Naderi, Elham Mirfarah, Matthew Bernhardt and Ding-Geng Chen

ABSTRACT

In the censored data exploration, the classical linear regression model which assumes normally distributed random errors is perhaps one of the commonly used frameworks. However, practical studies have often criticized the classical linear regression model because of its sensitivity to departure from the normality and partial nonlinearity. This paper proposes to solve these potential issues simultaneously in the context of the partial linear regression model by assuming that the random errors follow a scale-mixture of normal (SMN) family of distributions. The postulated method allows us to model data with great flexibility, accommodating heavy tails and outliers. By implementing the B-spline approximation and using the convenient hierarchical representation of the SMN distributions, a computationally analytical EM-type algorithm is developed for obtaining maximum likelihood (ML) parameter estimates. Various simulation studies are conducted to investigate the finite sample properties, as well as the robustness of the model in dealing with the heavy tails distributed datasets. Real-world data examples are finally analyzed for illustrating the usefulness of the proposed methodology.

ARTICLE HISTORY

Received 20 October 2020
Accepted 16 May 2021

KEYWORDS

B-spline; EM-type algorithm; interval-censored data; scale-mixture of normal family of distributions; semiparametric modeling

1. Introduction

Regression models form the basis for a large number of statistical inference procedures. The main purpose of regression analysis is to explore the relationship between a response variable \( Y \) and a \( p \)-dimensional covariate vector \( x \in \mathbb{R}^p \). More precisely, the regression models aim to find the expected value of \( Y \) for a given level of covariate vector \( x \), say \( E(Y|x) = g(x) \). These models can be found in a broad array of scientific fields, including econometrics, engineering, and medical studies, and allow judgments to be made on data within these fields. The parametric regression models in which the function \( g(\cdot) \) can be specified in terms of a small number of parameters are widely used for data exploration since the parameters can be interpreted as the effects of covariates on the response variable. For instance, the classical linear regression model assumes that \( Y = \beta^T x + \epsilon \), where \( \epsilon \) represents the error term followed by a normal distribution and \( \beta^T = (\beta_0, \beta_1, \ldots, \beta_{p-1}) \).
However, the nonlinearity (or partial nonlinearity), as well as invalid distribution assumptions, might increase the model misspecification and misleading inference. In this regard, the semiparametric techniques can provide an alternative platform for data analysis. One of the most acknowledged semiparametric models in the regression framework is the partially linear regression (PLR) model. The PLR model assumes that the response variable $Y$ not only has a linear relationship with certain covariates but also is regressed by another covariate $z$ with an unknown smooth function. More concretely, the PLR model allows both parametric and nonparametric specifications in the regression function as

$$ Y = \beta^\top x + \psi(z) + \epsilon, $$

where $\psi(\cdot)$ is an unknown smooth function. It can be observed from (1) that the covariates are separated into parametric component, $x$, and nonparametric term, $z$. The parametric part of the model can be interpreted as a linear function, while the nonparametric part frees the rest of the model from any structural assumption. The estimator of $\beta$ is accordingly less affected by the model bias. Since the introduction of PLR model by [3], it has received attention in economics, social and biological sciences. See for instance [12,14,30] and [11] among others. Moreover, various alternative approaches to the penalized least-squares method, considered by [3], were proposed for estimating the PLR models. For instance, Robinson [26] exploited the profile estimator for $\beta$ and the Nadaraya–Watson kernel estimate [25,29] for the unknown function. Severini and Staniswalis [28] also proposed a quasi-likelihood estimation method. Castro et al. [1] postulated a left-censored PLR based on the SMN family of distributions and achieved a fully Bayesian inference using the spline smooth functions to approximate the nonparametric function. Recently, [15] provided a classical inference on the model proposed by [1]. By approximating the nonparametric part by using the natural cubic spline, Lemus et al. [15] estimated the parameters via an EM-type algorithm with the penalized log-likelihood where they imposed restrictions on the model to avoid over-fitting and non-identification.

In this article, we propose three aspects. The first and main objective of this contribution is to develop a PLR model in which the error term is followed by an SMN distribution. The SMN family of distributions is an extension of the normal model with fat tails. It contains the Student’s-$t$, slash, Laplace and contaminated normal as special cases. Comprehensive surveys of the SMN family of distributions are available in [7,23] and [8], among others. Our PLR model secondly implements the basis splines estimator as a powerful nonparametric approach of kernel estimation. As discussed in Section 2, the spline functions are piecewise polynomial functions where the weights in the sum are parameters that should be estimated. The basis functions of the spline will allow us to build a flexible model across the whole range of the data without any restriction on the model. The regression models with censored dependent variable have been considered in fields such as econometric and engineering analysis, clinical essays, and medical surveys. Lastly, the new PLR model with B-spline consideration is enriched by assuming the interval-censoring scheme on the response variable, as an extension of [7,8].

The rest of the paper is therefore organized as follows. In Section 2, we provide an overview on the SMN family of distributions as well as B-spline approximation. Section 3 then presents the scale-mixture of normal partial linear regression model with interval-censored data, hereafter PLR-SMN-IC model. We also discuss on the implementation of the expectation-conditionally maximization either (ECME) algorithm [20] for obtaining
2. Background and notation

In this section, we briefly review the SMN family of distributions and polynomial basis spline function. For the sake of notation, let $\phi(\cdot; \mu, \sigma^2)$ and $\Phi(\cdot; \mu, \sigma^2)$ represent the probability density function (pdf) and cumulative distribution function (cdf) of the normal distribution with mean $\mu$ and variance $\sigma^2$, denoted by $N(\mu, \sigma^2)$.

2.1. An overview on the scale-mixture of normal family of distributions

Formally, the SMN distribution is generated by scaling the variance of a normal distribution with a positive weighting random variable $U$. More specifically, a random variable $V$ is said to have an SMN distribution, denoted by $\text{SMN}(\mu, \sigma^2, \nu)$, if it is generated by the two-level hierarchical representation

$$V | U = u \sim N(\mu, u^{-1}\sigma^2), \quad U \sim H(u; \nu),$$

(2)

where $H(\cdot; \nu)$ indicates the cdf of random scale-mixing factor $U$ with parameter index $\nu$. Accordingly, the pdf of the random variable $V$ can be expressed as

$$f_{\text{SMN}}(v; \mu, \sigma^2, \nu) = \int_0^{\infty} \phi(v; \mu, u^{-1}\sigma^2) \ dH(u; \nu), \quad v \in \mathbb{R}.$$

In what follows, $f_{\text{SMN}}(\cdot; \nu)$ and $F_{\text{SMN}}(\cdot; \nu)$ will be used to denote the pdf and cdf of the standard SMN distribution ($\mu = 0, \sigma^2 = 1$). Depending on the specification of $H(\cdot; \nu)$, many special cases of the general SMN family of distribution can be generated by (2). However, we focus on a few commonly used particular examples of the SMN family of distributions as follows:

- **Normal (N) distribution**: The SMN family of distributions contains the normal model when $H(\cdot; \nu)$ is degenerate with $u = 1$.
- **Student’s-t (T) distribution**: If $U \sim \text{Gamma}(\nu/2, \nu/2)$, the random variable $V$ then follows the Student’s-t distribution, $V \sim T(\mu, \sigma^2, \nu)$. Here, $\text{Gamma}(\alpha, \beta)$ represents the gamma distribution with shape and scale parameters $\alpha$ and $\beta$, respectively.
- **Slash (SL) distribution**: Let $U$ in (2) follow $\text{Beta}(v, 1)$, where $\text{Beta}(a, b)$ signifies the beta distribution with parameters $a$ and $b$. Then, $V$ is distributed by the slash model, denoted by $V \sim \text{SL}(\mu, \sigma^2, \nu)$, with pdf

$$f_{\text{SL}}(v; \mu, \sigma^2, \nu) = v \int_0^{1} u^{v-1} \phi(v; \mu, u^{-1}\sigma^2) \ du, \quad v \in \mathbb{R}.$$
Contaminated-normal (CN) distribution: Let $U$ be a discrete random variable with pdf

$$h(u; v, \gamma) = v I_A(u) + (1 - v) I_B(u), \quad v, \gamma \in (0, 1),$$

where $I_A(\cdot)$ represents the indicator function of the set $A$. The random variable $V$ in (2) then marginally follows the contaminated-normal distribution, $V \sim CN(\mu, \sigma^2, v, \gamma)$, which has the pdf

$$f_{CN}(v; \mu, \sigma^2, v, \gamma) = v \phi(v; \mu, \gamma^{-1} \sigma^2) + (1 - v) \phi(v; \mu, \sigma^2), \quad v \in \mathbb{R}.$$

### 2.2. B-spline function description

The basis spline function, B-spline in short, is a numerical tool that was originally introduced by [2] and recently received considerable attention in the statistical analysis of density estimation. For a given degree, smoothness and domain partition, the B-spline provides a sophisticated approach to approximate the unknown function $\psi(\cdot)$. Let $d$ be the fix degree of the B-spline and $m$ denotes the number of interior knots on the interval $[a, b]$, say $a = z_1 < z_2 < \cdots < z_{m+2} = b$. Then, $\psi(z)$ over the interval $[a, b]$ can formally be represented as a unique linear combination

$$\psi(z) = \sum_{i=1}^{m+d} \alpha_i B_i^d(z), \quad (3)$$

where the B-spline basis functions are defined recursively by

$$B_i^d(z) = \frac{z - z_i}{x_i + d - z_i} B_i^{d-1}(z) + \frac{z_{i+d+1} - z}{z_{i+d+1} - z_{i+1}} B_{i+1}^{d-1}(z), \quad B_i^0(z) = I_{[z_i, z_{i+1})}(z).$$

Although the piecewise linear approximation, case with $d = 1$, is attractively simple, it produces a visible roughness, unless the knots, $z_i$s, are close to each other.

To use a $d$-degree B-spline approximation with $m$ interior knots, $\psi(z)$ should be a $(d - 1)$ continuously differentiable function in each of the interval $(a, z_2), (z_2, z_3), \ldots, (z_{m+1}, b)$. This condition may complicate the approximation issue. However, practical studies confirm that quadratic and cubic B-splines usually provide robust platforms that have the minimal requirement, viz $\psi(z)$ should be at last twice continuously differentiable smooth function. Comprehensive review on the theory of spline function can be found in [27].

This contribution considers the cubic B-spline approximation. The number of interior knots is also chosen by either $m_1 = \lceil n^{1/3} \rceil + 1$ or $m_2 = \lceil n^{1/5} \rceil + 1$, where $\lceil a \rceil$ denotes the greatest integer less than or equal to $a$ and $n$ is the sample size. For the location of knots, two scenarios are considered: the equally spaced (ES), and the equally spaced quantile (ESQ). As investigated in our simulation studies, our strategy works well under these assumptions. However, for the practical studies if a large number of interior knots is required and there is not enough data on the boundary, one may need to obtain the knots through an unequally spaced technique to avoid singularity problems [13].
3. Proposed methodology

3.1. Model formulation

The partial linear regression model is considered in this part upon the assumption that the random error follows the SMN family of distributions. Let \( \mathbf{Y} = (Y_1, \ldots, Y_n)^\top \) be a vector of response variables and \( \mathbf{x}_i \) a vector of explanatory covariates corresponding to \( Y_i \). The PLR model based on the SMN distribution is defined as

\[
Y_i = \beta^\top \mathbf{x}_i + \psi(z_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} SMN(0, \sigma^2, \nu), \quad i = 1, 2, \ldots, n, \tag{4}
\]

where iid stands for the independent and identically distributed, and \( z_i \) is a univariate covariate.

Censored data is widely seen in practical studies, such as time-to-death in cancer research, time-to-infection in HIV, TB and COVID-19 studies. The interval-censoring scheme is the most general scenario that covers the typical right- and left-censoring as special cases. Interval-censored data can also be generated from other science fields, such as detection limits of quantification in environmental, toxicological and pharmacological studies. We assume that the set of joint variables \( \{W_i, \rho_i\} \) are observed where \( W_i \) represents the uncensored reading \( (W_i = Y_i^o) \) or interval-censoring \( (W_i = (c_{i1}, c_{i2})) \) and \( \rho_i \) is the censoring indicator; \( \rho_i = 1 \) if \( c_{i1} \leq Y_i \leq c_{i2} \) and \( \rho_i = 0 \) if \( Y_i = Y_i^o \). Note that in this setting if \( c_{i1} = -\infty \) (or \( c_{i2} = +\infty \)) the left-censoring (or right-censoring) is occurred. Our methodology is established based on the interval-censoring scheme; however, the left- and right-censored data are also investigated in the simulation and real-data analyses. The PLR model of censored data based on particular cases of the SMN family of distributions will be refereed as PLR-N-IC, PLR-T-IC, PLR-SL-IC and PLR-CN-IC for the N, T, SL and CN distributions, respectively.

For ease of exposition and based on the aforementioned setting, one can divide \( \mathbf{Y} \) to the sets of observed responses, \( \mathbf{Y}^o \), and censored cases. Therefore, \( \mathbf{Y} \) can be viewed as the latent variable since it is partially unobserved. Under these assumptions, the likelihood function for \( \theta = (\beta^\top, \sigma^2, \nu) \) of the PLR-SMN-IC model can be written as

\[
L(\theta) = \prod_{i=1}^n \left[ \sigma^{-1} f_{SMN}\left( \frac{y_i^o - \mu_i}{\sigma}; \nu \right) \right]^{1-\rho_i} \left[ F_{SMN}\left( \frac{c_{i2} - \mu_i}{\sigma}; \nu \right) - F_{SMN}\left( \frac{c_{i1} - \mu_i}{\sigma}; \nu \right) \right]^{\rho_i},
\]

(5)

where \( \mu_i = \beta^\top \mathbf{x}_i + \psi(z_i) \), \( \rho = (\rho_1, \ldots, \rho_n)^\top \) is the vector of censoring indicators and \( y_i^o \) denotes the realization of \( Y_i^o \).

Due to complexity of the likelihood function (5), there is no analytical solution to obtain the ML estimate of parameters and the smooth function. A numerical search algorithm should therefore be developed. With the embedded hierarchical representation (2) and B-spline function, an innovative EM-type algorithm is developed to fit the PLR-SMN-IC model to the data.

3.2. Estimation via an EM-type algorithm

In order to utilize an EM-type algorithm for calibrating the PLR-SMN-IC model to the data, we first replace the basis expansion of \( \psi(z_i) \) defined in Equation (3)
into the PLR-SMN-IC model (4). Consequently, it can be obtained that

\[ Y_i = \beta^T x_i + \sum_{j=1}^{m+d} \alpha_j B^d_j(z_i) + \epsilon_i = \beta^T x_i + \alpha^T B^d(z_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{SN}(0, \sigma^2, v), \]

where \( \alpha = (\alpha_1, \ldots, \alpha_{m+d})^T, \ B^d(z) = (B^d_1(z), \ldots, B^d_{m+d}(z))^T. \) For convenience, the obtained model in (6) can be rewritten as

\[ Y_i = \tilde{\beta}^T \tilde{x}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{SN}(0, \sigma^2, v), \]

where \( \tilde{x}_i = (x_i^T, B^d_1(z_i))^T \) is the pseudo covariate vector and \( \tilde{\beta} = (\beta^T, \alpha^T)^T \) the pseudo parameter. So, it can be concluded that \( Y_i \sim \mathcal{SN}(\tilde{\beta}^T \tilde{x}_i, \sigma^2, v), \) for \( i = 1, \ldots, n. \) Now using (2), the hierarchical representation of the PLR-SMN-IC model is

\[ Y_i | (x_i, U = u_i) \sim \mathcal{N}(\tilde{\beta}^T \tilde{x}_i, u_i^{-1} \sigma^2), \quad U_i \sim H(u_i; v). \quad (7) \]

For the vector of realizations \( y = (y_1, \ldots, y_n)^T \) and latent vector \( u = (u_1, \ldots, u_n)^T, \) let \( y_c = (w^T, \rho^T, y^T, u^T)^T \) represent the entire model values. In light of (7), the complete-data log-likelihood function for \( \Theta = (\tilde{\beta}^T, \sigma^2, v) \) associated with \( y_c \) is

\[
\ell_c(\Theta | y_c) = \sum_{i=1}^{n} \log \phi(y_i; \tilde{\beta}^T \tilde{x}_i, u_i^{-1} \sigma^2) + \sum_{i=1}^{n} \log h(u_i; v) \\
\approx -\frac{n}{2} \log \sigma^2 + \sum_{i=1}^{n} \log h(u_i; v) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} u_i(y_i - \tilde{\beta}^T \tilde{x}_i)^2, \quad (8)
\]

where \( h(\cdot; v) \) is the pdf of \( U_i, \) and \( w = (w_1, \ldots, w_n)^T \) denotes the realization of \( W = (W_1, \ldots, W_n)^T. \)

In the rest, an expectation conditional maximization either (ECM; [20]) algorithm is developed to estimate parameters of the PLR-SMN-IC model. As an extension of expectation conditional maximization (ECM; [22]), the ECME algorithm has stable properties (e.g. monotone convergence and implementation simplicity) and can be implemented faster than ECM. The ML parameter estimates via ECME algorithm are obtained by maximizing the constrained Q-function with some CM-steps that maximize the corresponding constrained marginal likelihood function, called CML-steps. The ECME algorithm for ML estimation of the PLR-SMN-IC model proceeds as follows:

- **Initialization**: Set the number of iteration to \( k = 0 \) and choose a relative starting point.
- **E-Step**: At iteration \( k, \) the expected value of the complete-data log-likelihood function (8), known as the Q-function, is computed as

\[
Q(\Theta | \hat{\Theta}^{(k)}) = -\frac{n}{2} \log \sigma^2 + \sum_{i=1}^{n} \tilde{\epsilon}_i^{(k)} - \frac{1}{2\sigma^2} \left( \tilde{\epsilon}_i^{(k)} + (\tilde{\beta}^T \tilde{x}_i)^2 \tilde{u}_i^{(k)} - 2\tilde{u}_i^{(k)} \tilde{\beta}^T \tilde{x}_i \right), \quad (9)
\]
where \( \hat{u}_{y_i}^{(k)} = E(U_i|Y_i, \rho_i, \hat{\Theta}^{(k)}), \hat{u}_i^{(k)} = E(U_i|w_i, \rho_i, \hat{\Theta}^{(k)}), \hat{\upsilon}_i^{(k)} = E(U_i Y_i|w_i, \rho_i, \hat{\Theta}^{(k)}) \), and \( \hat{\Upsilon}_i^{(k)} = E(\log h(U_i; \upsilon_j)|w_i, \rho_i, \hat{\Theta}^{(k)}) \). Depending on whether a subject is observed or censored, these conditional expectations are computed as:

(i) For the uncensored observations, i.e. \( \rho_i = 0 \) and \( w_i = y_i^0 \), the expectations are computed as: \( \hat{u}_i^{(k)} = E(U_i|Y_i = y_i^0, \hat{\Theta}^{(k)}), \hat{\upsilon}_i^{(k)} = y_i \hat{u}_i^{(k)} \), \( \hat{\upsilon}_i^{(k)} = y_i^2 \hat{u}_i^{(k)} \), and \( \hat{\Upsilon}_i^{(k)} = E(\log h(U_i; \upsilon_j)|y_i = y_i^0, \hat{\Theta}^{(k)}) \).

(ii) If the censoring is occurred, \( \rho_i = 1 \) and \( w_i = (c_{i1}, c_{i2}) \), one can compute the conditional expectations by

\[
\hat{\upsilon}_i^{(k)} = E(U_i|c_{i1} \leq Y_i \leq c_{i2}, \hat{\Theta}^{(k)}), \quad \hat{\upsilon}_i^{(k)} = E(U_i Y_i|c_{i1} \leq Y_i \leq c_{i2}, \hat{\Theta}^{(k)}),
\]

\[
\hat{\Upsilon}_i^{(k)} = E(\log h(U_i; \upsilon_j)|c_{i1} \leq Y_i \leq c_{i2}, \hat{\Theta}^{(k)}).
\]

The closed form of the conditional expectations for the particular cases of the SMN family of distributions is provided in Appendix with proof in [7].

- **CM-step:** The M-step consists of maximizing the Q-function with respect to \( \Theta^{(k)} \). The maximization of (9) over \( \beta \) and \( \sigma^2 \) lead to the following CM estimators:

\[
\hat{\beta}^{(k+1)} = \left( \sum_{i=1}^{n} \hat{u}_i^{(k)} \tilde{x}_i \tilde{x}_i^\top \right)^{-1} \sum_{i=1}^{n} \hat{\upsilon}_i^{(k)} \tilde{x}_i,
\]

\[
\hat{\sigma}^{2(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\upsilon}_i^{(k)} - 2 \hat{\upsilon}_i^{(k)} \hat{\beta}^{(k+1)} \tilde{x}_i + \hat{\upsilon}_i^{(k)} \left( \hat{\beta}^{(k+1)} \tilde{x}_i \right)^\top \right)^2.
\]

- **CML-step:** The update of \( \hat{\upsilon}^{(k)} \) crucially depends on the conditional expectation \( \hat{\Upsilon}_i^{(k)} \) which is quite complicated. However, it can be updated through maximizing the constrained actual log-likelihood function as

\[
\hat{\upsilon}^{(k+1)} = \arg \max_{\upsilon} \sum_{i=1}^{n} (1 - \rho_i) \log \left[ f_{SMN} \left( \frac{y_i^0 - \hat{\beta}^{(k+1)} \tilde{x}_i}{\hat{\sigma}^{(k+1)}}; \upsilon \right) / \hat{\sigma}^{(k+1)} \right] + \rho_i \log \left[ F_{SMN} \left( \frac{c_{i2} - \hat{\beta}^{(k+1)} \tilde{x}_i}{\hat{\sigma}^{(k+1)}}; \upsilon \right) - F_{SMN} \left( \frac{c_{i1} - \hat{\beta}^{(k+1)} \tilde{x}_i}{\hat{\sigma}^{(k+1)}}; \upsilon \right) \right].
\]  

The R function `nlminb(·)` is used to update \( \upsilon \) in the numerical parts of the current paper.

**Remark 3.1.** To facilitate the update of \( \upsilon = (\upsilon, \gamma) \) for the PLR-CN-IC model in the above ECME algorithm, one can introduce an extra latent binary variable \( B_i \) such that \( B_i = 1 \) if an observation \( y_i \) is a bad point (outlier) and \( B_i = 0 \) if \( y_i \) is a good point. The hierarchical
representation of the PLR-CN-IC model can therefore be written as
\[
Y_i | (\tilde{x}_i, U = u_i, B_i = 1) \sim \mathcal{N}(\hat{\beta}^T \tilde{x}_i, u_i^{-1} \sigma^2), \quad U_i | (B_i = 1) \sim h(u_i; v, \gamma), \quad B_i \sim \mathcal{B}(1, v),
\]
where \( \mathcal{B}(1, v) \) denotes the Bernoulli distribution with success probability \( v \). Consequently, by computing the Q-function based on (11), the update of \( \nu \) is
\[
\hat{\nu}^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{b}_i^{(k)},
\]
where \( \hat{b}_i^{(k)} \) is defined as
\[
\hat{b}_i^{(k)} = \begin{cases} 
\hat{\nu}^{(k)} \phi(y_i; \hat{\beta}^{(k)} \tilde{x}_i, \hat{\gamma}^{(k)} \sigma^{2(k)}) - (1 - \hat{\nu}^{(k)}) \phi(y_i; \hat{\beta}^{(k)} \tilde{x}_i, \hat{\gamma}^{(k)} \sigma^{2(k)}) & \text{if } \rho_i = 0, \\
\hat{\nu}^{(k)} \left( \Phi(c_{i2}; \hat{\beta}^{(k)} \tilde{x}_i, \hat{\gamma}^{(k)} \sigma^{2(k)}) - \Phi(c_{i1}; \hat{\beta}^{(k)} \tilde{x}_i, \hat{\gamma}^{(k)} \sigma^{2(k)}) \right) - F_{\text{CN}}(c_{i2}; \hat{\beta}^{(k)} \tilde{x}_i, \hat{\gamma}^{(k)} \sigma^{2(k)}) & \text{if } \rho_i = 1,
\end{cases}
\]
where \( F_{\text{CN}}(\cdot) \) denotes the cdf of CN distribution. Since there is no closed-form solution for \( \hat{\gamma}^{(k+1)} \), we update \( \gamma \) by maximizing the restricted actual log-likelihood function (10) as a function of \( \gamma \).

### 3.3. Computational aspects

#### 3.3.1. Initial values

The choice of starting points plays a critical role in speeding up parameter estimation via the EM-type algorithm. It also guarantees reaching stationarity in the ML solutions. As a convenient approach to generate sensible initial values, we fit a classical linear model to data and find the estimate of \( \beta \). To do this, the \texttt{lm} function is used. The parameter \( \sigma^2 \) is also set to the average of squared residuals of the classical linear model. The obtained parameter estimates of \( \beta \) and \( \sigma^2 \) are used as the initial points for implementing PLR-N-IC model. By calibrating the PLR-N-IC model to the data, the parameter estimates are exploited as the starting values for the PLR-T-IC, PLR-SL-IC, PLR-CN-IC models. We adapt the scale mixture factor parameter \( \hat{\nu}^{(0)} \) so that it corresponds to an initial assumption near normality. For instance, set \( \hat{\nu}^{(0)} = 20 \) in the PLR-T-IC and PLR-SL-IC models.

#### 3.3.2. Convergence

The process of EM algorithm should be iterated until a suitable convergence rule is satisfied. Herein, the incremental likelihood property of the EM-type algorithm is used to detect the convergence. The above ECME algorithm is terminated either the maximum number of iterations \( K_{\text{max}} = 2000 \) has been reached or
\[
\frac{\ell(\hat{\Theta}^{(k+1)}) - \ell(\hat{\Theta}^{(k)})}{|\ell(\hat{\Theta}^{(k)})|} \leq \varepsilon,
\]
where \( \ell(\cdot) \) is the logarithm of the likelihood function (5) and \( \varepsilon \) is a user specified tolerance. In our study, the tolerance \( \varepsilon \) is considered as \( 10^{-5} \).
3.3.3. Model selection

The models in competition in our data analysis are compared using the most commonly used information-based measures. Following [13], we vary the number of interior knots in a relatively large range and choose the one which minimizes the Akaike information criterion (AIC) or Bayesian information criteria (BIC) defined as

\[
AIC = 2(m + d + p + s + 1) - 2\ell_{\text{max}} \quad \text{and} \quad BIC = (m + d + p + s + 1) \log n - 2\ell_{\text{max}},
\]

where \(m, d, p\) and \(s\) denote, respectively, the number of interior knots, degree of the spline, number of covariates and number of parameters of the scale mixing factor \(U\), and \(\ell_{\text{max}}\) is the maximized log-likelihood value. Models with lower values of AIC or BIC are considered more preferable. It should be noted that \(m\) ranging from 3 to 10 are usually adequate and the results are quite stable when we vary the number of interior knots.

4. Simulation studies

In this section, four Monte–Carlo simulation studies are conducted in order to verify the asymptotic properties of the ML estimates, to assess the fitting performance of the model, and to check the robustness of the proposed model in dealing with highly peaked and heavily tailed data as well as its robustness in presence of outliers and noise points. For the sake of data generation, it should be noted that one of the simplest ways of interval-censored data generation is to consider \(C_{i1} = Y_i - U_i^{(1)}\) and \(C_{i2} = Y_i + U_i^{(2)}\) where \(U_i^{(1)}\) and \(U_i^{(2)}\) are two independent continuous variables followed by \(U(0, c)\) such that the non-informative condition (1.2) of [10] is fulfilled. Here \(U(a, b)\) represents the uniform distribution on interval \((a, b)\). Recommended by [10], a way to go around non-informative condition is to construct \(C_{i1} = \max(Y_i - U_i^{(1)}, Y_i + U_i^{(2)} - c)\) and \(C_{i2} = \min(Y_i + U_i^{(2)}, Y_i - U_i^{(1)} + c)\) for any constant \(c\). For the random samples \(y_1, \ldots, y_n\) generated from model (4), the following steps are used in our simulation studies to have a \(p\%\) interval-censored dataset.

(1) Compute the number of censored samples \(\mathcal{N}C = \lfloor n \times p \rfloor + 1\) and then, generate an index set, \(\mathcal{I}N\mathcal{D}\), as a sample of size \(\mathcal{N}C\) from \(\{1, 2, \ldots, n\}\) without replacement by using the \(R\) function sample(-).

(2) For \(i = 1, \ldots, n\), if \(i \in \mathcal{I}N\mathcal{D}\), we then generate two random samples \(u_i^{(1)}\) and \(u_i^{(2)}\) independently from \(U(0, c)\) with \(c = 1\). Finally, we have \(c_{i1} = \max(y_i - u_i^{(1)}, y_i + u_i^{(2)} - c)\) and \(c_{i2} = \min(y_i + u_i^{(2)}, y_i - u_i^{(1)} + c)\).

4.1. Finite sample properties of the ML estimates

To assess the performance of the ECME algorithm for obtaining ML estimates, an extensive simulation study is conducted under various scenarios. We generate the response variable from the following model

\[
Y_i = \beta^\top x_i + \psi(z_i) + \epsilon_i, \quad i = 1, 2, \ldots, n,
\]

where \(\epsilon_i\) is generated from either the N, T, SL or CN distributions for the sample size ranging from 50 to 800, \(\psi(z) = \exp(z/3) - 1\), and \(\beta^\top = (1, 2, -2)\). The regression covariates
vector is considered as \( x_i^* = (x_{1i}, x_{2i}, x_{3i}) \) where \( x_{1i}, x_{2i} \) and \( x_{3i} \) are respectively generated from \( \mathcal{N}(0, 1) \), \( \mathcal{B}(1, 0.5) \) and \( \mathcal{U}(-4, 1) \), and \( z_i \) form \( \mathcal{U}(-1, 2) \). Furthermore, we set \( \sigma^2 = 2 \), \( v = 3 \) for the T and SL distributions, and \((\nu, \gamma) = (0.4, 0.3)\) for the CN model. The considered levels of interval-censoring are 7.5% and 30%.

In each replication of 500 trials, we fit the special cases of the PLR-SMN-IC model to the data under both assumptions of the number of interior knots \( m_1 \) and \( m_2 \) explained in Section 2. The location of knots is also chosen based on the ESQ scenario, even though same results are observed based on the ES method. Based on the described results in Appendix, we only focused on \( m_2 \) approach in what follows of this simulation to shorten the length of the paper.

To investigate the estimation accuracies, we compute the bias and the mean squared error (MSE):

\[
\text{BIAS} = \frac{1}{500} \sum_{j=1}^{500} (\hat{\theta}_j - \theta_{\text{true}}) \quad \text{and} \quad \text{MSE} = \frac{1}{500} \sum_{j=1}^{500} (\hat{\theta}_j - \theta_{\text{true}})^2,
\]

where \( \hat{\theta}_j \) denotes the estimate of a specific parameter at the \( j \)-th replication. In addition, we are interested in examining the accuracy of the \( \psi(z) \) estimation in terms of the averaged integrated absolute bias (IABIAS) and mean integrated square error (MISE) defined as

\[
\text{IABIAS} = \frac{1}{n500} \sum_{j=1}^{500} \sum_{i=1}^{n} |\hat{\psi}_j(z_i) - \psi(z_i)| \quad \text{and} \quad \text{MISE} = \frac{1}{n500} \sum_{j=1}^{500} \sum_{i=1}^{n} (\hat{\psi}_j(z_i) - \psi(z_i))^2,
\]

where \( \hat{\psi}_j(z_i) = \hat{\alpha}_j^T \beta^3(z_i) \), is the estimate of \( \psi(z) \) at the \( j \)-th replication.

Figure 1 plots the BIAS and MSE of the regression parameter vector \( \beta \), \( \sigma^2 \), and the IABIAS and MISE of the estimated \( \psi(z) \) as a function of sample size for two levels of censoring. It can be observed that the estimates of the \( \beta \)'s have very small (around zero) BIAS for all sample sizes. Moreover, as \( n \) increases the BIAS of \( \hat{\sigma}^2 \) and the IABIAS of the estimated \( \hat{\psi}(z) \) tend to zero. The plots in Figure 1 also reveal that the MSE of the parameters and MISE of \( \hat{\psi}(z) \) tend to zero when the sample size increases. These results indicate that the estimator of the regression parameters and the estimation of \( \psi(z) \) are rather accurate.

### 4.2. Model comparison

This Monte–Carlo simulation experiment investigates the flexibility of the PLR-SMN-IC model when data are generated form some other distributions. We generate 200 samples of size 400 from the right-censored linear regression model

\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sin(\pi z) + \epsilon, \]

where \( \beta^T = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 2, -2, 1) \), \( \sigma^2 = 2 \). In the aforementioned PLR model the covariates are generated as \( x_1 \sim \mathcal{N}(0, 1) \), \( x_2 \sim \mathcal{B}(1, 0.5) \), \( x_3 \sim \mathcal{U}(-2, 2) \) and \( z \sim \mathcal{U}(-1, 7) \). The considered levels of censoring are 7.5%, 15%, and 30%. We generate the error terms under three different scenarios from the SMN representation (2). In the first scenario, the random variable \( U \) in (2) is generated from the exponential distribution with mean 2, where as the Birnbaum–Saunders (BS) and generalized inverse Gaussian (GIG)
models are used for the second and third scenarios, respectively. The parameters are set to \( \kappa = 1 \) (scale parameter) and \( \nu = 2 \) (shape parameter) for the BS distribution, and to \((\kappa, \chi, \psi) = (-0.5, 1, 3)\) for the GIG model.

Since the error term under these scenarios follows the Laplace, BS and GIG scale-mixture distributions, the notations PLR-L-IC, PLR-SBS-IC and PLR-SGIG-IC are used to denote the PLR model under these models. Bare in mind that the considered non-normal data generation offers desired level of leptokurtosis. It should also be noted that the PLR-L-IC, PLR-SBS-IC and PLR-SGIG-IC models are not considered in this paper since their conditional expectations involved in the ECME algorithm do not exist.

For each sample, we fit the PLR-N-IC, PLR-T-IC, PLR-SL-IC and PLR-CN-IC models to the data by assuming both \( m_1 \) and \( m_2 \) methods for choosing the number of interior knots, as well as ES and ESQ methods for the knots location. Table 1 depicts the average values of AIC and BIC measures over the 200 trials. Not surprisingly, under different levels of censoring, number of interior knots and location of knots, all criteria are in favor of the PLR censored models based on heavy-tailed distributions. As highlighted in Table 1, the PLR-T-IC model is the best in almost all cases. Furthermore, the outputs in this table suggest that models with number of interior knots \( m_2 \) perform significantly better than those with \( m_1 \). However, it can be seen that the values of AIC and BIC for two methods of knots position (in all models) are very closed. It indicates that the PLR-SMN-IC model is robust against the location of knots.
Table 1. Performance of special cases of PLR-SMN-IC model fitted to 200 simulated datasets from either PLR-L-IC, PLR-SBS-IC or PLR-SGIG-IC models.

| True model | Cens. | Criterion | knots loc. | PLR-N-IC | PLR-T-IC | PLR-CN-IC | PLR-SL-IC |
|------------|-------|-----------|------------|----------|----------|-----------|-----------|
|            |       |           |            | $m_1$    | $m_2$    | $m_1$    | $m_2$    |
| 7.5%       | AIC   | ES        | 1609.464   | 1605.538 | 1566.020 | 1562.102 | 1573.886 | 1562.102 |
|            |       | ESQ       | 1609.425   | 1605.473 | 1566.202 | 1561.543 | 1572.966 | 1562.212 |
|            | BIC   | ES        | 1673.310   | 1653.435 | 1637.867 | 1617.983 | 1645.733 | 1618.693 |
|            |       | ESQ       | 1673.288   | 1653.371 | 1638.048 | 1617.424 | 1644.812 | 1618.092 |
| PLR-L-IC   | 15%   | AIC       | 1478.256   | 1472.341 | 1442.978 | 1438.575 | 1459.893 | 1442.688 |
|            |       | ESQ       | 1477.029   | 1472.298 | 1442.344 | 1438.136 | 1458.009 | 1445.056 |
|            | BIC   | ES        | 1542.119   | 1520.238 | 1514.824 | 1513.739 | 1531.739 | 1549.895 |
|            |       | ESQ       | 1540.893   | 1520.195 | 1514.190 | 1494.017 | 1529.855 | 1522.911 |
| 30%        | AIC   | ES        | 1264.806   | 1264.870 | 1233.298 | 1226.613 | 1247.416 | 1239.220 |
|            |       | ESQ       | 1265.242   | 1264.150 | 1231.990 | 1226.302 | 1247.435 | 1244.752 |
|            | BIC   | ES        | 1328.669   | 1312.767 | 1305.144 | 1282.493 | 1319.262 | 1311.260 |
|            |       | ESQ       | 1329.106   | 1312.048 | 1303.836 | 1282.182 | 1319.212 | 1310.260 |
| 7.5%       | AIC   | ES        | 1723.514   | 1719.053 | 1587.860 | 1583.466 | 1590.737 | 1583.011 |
|            |       | ESQ       | 1723.027   | 1719.418 | 1587.746 | 1583.220 | 1590.390 | 1583.220 |
| PLR-SBS-IC | 15%   | AIC       | 1584.004   | 1580.361 | 1584.342 | 1581.378 | 1584.950 | 1581.382 |
|            |       | ESQ       | 1584.342   | 1581.378 | 1587.480 | 1583.504 | 1590.390 | 1583.220 |
|            | BIC   | ES        | 1647.867   | 1628.258 | 1528.072 | 1507.425 | 1539.775 | 1513.379 |
|            |       | ESQ       | 1648.205   | 1629.276 | 1527.696 | 1507.713 | 1539.239 | 1514.231 |
| 30%        | AIC   | ES        | 1399.779   | 1399.677 | 1399.677 | 1399.677 | 1399.677 | 1399.677 |
|            |       | ESQ       | 1399.677   | 1399.677 | 1399.677 | 1399.677 | 1399.677 | 1399.677 |
|            | BIC   | ES        | 1463.643   | 1447.575 | 1347.067 | 1323.609 | 1360.761 | 1333.115 |
|            |       | ESQ       | 1463.486   | 1448.103 | 1345.001 | 1322.976 | 1360.346 | 1333.283 |
| 7.5%       | AIC   | ES        | 1736.578   | 1734.462 | 1736.578 | 1734.462 | 1736.578 | 1734.462 |
|            |       | ESQ       | 1737.991   | 1734.704 | 1723.561 | 1720.301 | 1727.343 | 1722.575 |
| PLR-SGIG-IC| 15%   | AIC       | 1609.567   | 1606.316 | 1609.567 | 1606.316 | 1609.567 | 1606.316 |
|            |       | ESQ       | 1609.403   | 1605.598 | 1601.513 | 1597.056 | 1606.890 | 1600.087 |
|            | BIC   | ES        | 1673.430   | 1654.213 | 1673.430 | 1654.213 | 1673.430 | 1654.213 |
|            |       | ESQ       | 1673.267   | 1653.496 | 1673.359 | 1652.937 | 1678.736 | 1655.968 |
| 30%        | AIC   | ES        | 1393.320   | 1389.093 | 1388.290 | 1379.569 | 1395.803 | 1383.915 |
|            |       | ESQ       | 1394.001   | 1391.994 | 1387.456 | 1380.693 | 1395.832 | 1385.663 |
|            | BIC   | ES        | 1457.184   | 1436.990 | 1460.136 | 1435.449 | 1467.676 | 1439.795 |
|            |       | BIC       | 1457.864   | 1439.892 | 1459.302 | 1436.574 | 1467.676 | 1441.544 |

The lowest AIC and BIC scores, among all models and two knots positions, are indicated in bold.
4.3. **Imputation of censored observations in presence of noisy points**

We consider the following left-censored PLR model with censoring levels 10%, 20%,

\[ Y_i = 1 + 3x_{1i} + \psi(z_i) + \epsilon_i, \quad i = 1, \ldots, 200, \]

where the nonparametric component \( \psi(z_i) \) has the form \( 3 \sin(2z_i) + 10\xi \mathbb{I}_{(0,0.1)}(z_i) + \xi \mathbb{I}_{(0.1,\infty)}(z_i) \), in which \( \xi \) is set to vary among 0, 3 and 6. Figure 2 shows the plot of \( \psi(z) \) in which the jump in the function for different values of \( \xi \) can be observed. It is assumed that the realizations \((x_{1i}, z_i)\) are jointly generated from a bivariate normal distribution with mean zero, variance one and correlation coefficient \( \rho = 0.5 \). The error term \( \epsilon_i \) is also drawn from \( \mathcal{N}(0, 1) \). In this experiment, we are interested in predicting the censored realizations, \( y^c_i \), through computing the expectation \( \hat{y}^c_i = E(Y|w_i, \rho_i, \hat{\Theta}) \) at the last iteration of the ECME algorithm. Moreover, to check the influence of noise points in model performance and imputation of censored observations, some noisy points simulated from \( \mathcal{U}(-5, 5) \), \( \mathcal{U}(-3, 2) \), and \( \mathcal{U}(-2, 8) \) are added to the dataset, respectively for \( y, x_1 \) and \( z \). We generate 200 samples of size 200 and consider \( m_2 \) and ESQ approaches for the number of interior knots and their locations. By fitting the PLR-N-IC, PLR-T-IC, PLR-CN-IC and PLR-SL-IC models to the data, the BIC value of the models is recorded. To investigate the prediction performance of the censored realizations, we follow [21] and compute the mean absolute error (MAE) defined as

\[
\text{MAE} = \frac{1}{n_c \cdot 200} \sum_{j=1}^{200} \sum_{i=1}^{n_c} |\hat{y}_{ij} - y_{ij}|,
\]

where \( y_{ij} \) and \( \hat{y}_{ij} \) are the actual and predicted values of the \( i \)th realization at \( j \)th trail and \( n_c \) denotes the number of censored observations. Table 2 shows the average values of BIC and MAE over 200 replications for two censoring levels, three values of \( \xi \), and three levels of noise points. The results in Table 2 indicate that the value of MAE increases as the number of censored observations grows. As can be expected, the PLR-N-IC model performs well when the number of noise realizations is zero. However, adding the noise points reduces its flexibility in both fitting data and predicting censored observations. It can be observed that only the MAE of the PLR-N-IC model increases when the noise points are added to

![Figure 2](image-url). Plot of the function \( \psi(z_i) = 3 \sin(2z_i) + 10\xi \mathbb{I}_{(0,0.1)}(z_i) + \xi \mathbb{I}_{(0.1,\infty)}(z_i) \) for three values of \( \xi \).
Table 2. Model performance and evaluation of the prediction accuracy for the PLR-N-IC, PLR-T-IC, PLR-CN-IC and PLR-SL-IC models with two censoring levels and various numbers of noise point.

| Cens. | Noise | \( \xi \) | \( \text{BIC} \) | \( \text{MAE} \) | \( \text{BIC} \) | \( \text{MAE} \) | \( \text{BIC} \) | \( \text{MAE} \) | \( \text{BIC} \) | \( \text{MAE} \) |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|       | 0     | 0     | 556.012 | 0.669 | 566.464 | 0.674 | 566.410 | 0.671 | 561.050 | 0.671 |
|       | 3     | 0.668 | 565.887 | 0.672 | 565.849 | 0.669 | 560.449 | 0.669 | 560.209 | 0.668 |
| 0%    | 10    | 0     | 732.298 | 0.671 | 669.805 | 0.654 | 666.365 | 0.654 | 666.271 | 0.658 |
|       | 3     | 0.669 | 762.192 | 0.687 | 679.711 | 0.666 | 673.627 | 0.666 | 673.007 | 0.665 |
|       | 6     | 0.668 | 789.640 | 0.707 | 689.772 | 0.671 | 683.301 | 0.672 | 682.814 | 0.671 |
| 10%   | 10    | 0     | 860.391 | 0.730 | 763.462 | 0.679 | 757.945 | 0.678 | 759.861 | 0.685 |
|       | 3     | 0.671 | 899.970 | 0.687 | 777.642 | 0.699 | 772.075 | 0.696 | 771.062 | 0.709 |
|       | 6     | 0.668 | 920.889 | 0.704 | 878.971 | 0.712 | 872.039 | 0.704 | 871.512 | 0.728 |
| 20%   | 10    | 0     | 682.773 | 0.744 | 617.020 | 0.707 | 615.887 | 0.695 | 614.687 | 0.704 |
|       | 3     | 0.671 | 711.959 | 0.687 | 777.642 | 0.699 | 772.075 | 0.696 | 771.062 | 0.709 |
|       | 6     | 0.668 | 738.295 | 0.707 | 714.279 | 0.719 | 714.320 | 0.708 | 708.880 | 0.711 |

the data, showing its lack of robustness related to the contaminated dataset with the noise observations. One can see from the values of BIC that the PLR-SL-IC model is the best model for all scenarios with noise observations.

4.4. Robustness of the EM estimates

In this Monte–Carlo experiment, we are interested in comparing the models parameter estimate performance when data contain a single outlier on the response variable. In this regard, 500 Monte Carlo samples of size \( n = 300 \) are simulated from the interval-censored PLR model (4) where the errors are randomly generated from \( \mathcal{N}(0, 2) \). The imposed censoring levels are 10%, 20% and 30%. We set \( \beta^T = (1, 4, 2) \) and \( x_i^T = (1, x_{1i}, x_{2i}) \) where the covariates are generated independently from \( x_{1i} \sim \mathcal{U}(2, 20) \) and \( x_{2i} \sim \mathcal{B}(1, 0.6) \). The nonparametric part of the model is also simulated by the function \( \psi(z_i) = \cos(4\pi z_i) \exp(-z_i^2/2) \), where \( z_i \) comes from \( \mathcal{U}(0, 3) \). Furthermore, we add perturbations into the largest uncensored observation, namely \( y_{\text{max}}^* = y_{\text{max}} - \delta \) with \( \delta \) varying from 1 to 10. The parameter estimates are computed under the PLR-N-IC, PLR-T-IC, PLR-SL-IC and PLR-CN-IC models in each replication with and without contaminations, denoted by \( \hat{\theta}(\delta) \) and \( \hat{\theta} \), respectively. We use \( m_2 \) and ESQ approaches for number of interior knots and their locations.

To assess relative changes on the parameter estimates by the presence of outliers, the mean magnitude of relative error (MMRE; [4]) is calculated as

\[
\text{MMRE} = \frac{1}{2000} \sum_{l=1}^{500} \left\{ \sum_{j=1}^{3} \left( \frac{\hat{\beta}_{jl}(\delta) - \hat{\beta}_{jl}}{\hat{\beta}_{jl}} \right) + \frac{\hat{\sigma}^2(\delta) - \hat{\sigma}^2}{\hat{\sigma}^2} \right\}.
\]
Curves of the average MMRE as a function of contamination level $\delta$ are displayed in Figure 3. It could be seen that the influence of outlier in models parameter estimation increases as $\delta$ approaches large values. As one would expect, the heavy-tailed models, such as PLR-T-IC and PLR-SL-IC, are less adversely affected. It shows their robustness against the presence of outliers. On the other hand, an extreme observation seems to be much more effective on the PLR-N-IC, reflecting a lack of ability to reduce the effect of outliers.

5. Real-data analyses

In the following, we present two applications of the PLR-SMN-IC model to real-data for illustrative purposes. The first real dataset corresponds to the young married women’s labor force participation using the data extracted from the Canadian Survey of Labour and Income Dynamics (SLID). The complete data, reported in [6], consist of 6900 respondents of the married women aging from 18 to 65 years. However, the samples with missing data (near to 7%) are removed from the study. For each 6340 women in the remaining sample, four measures, namely working hours, family income, age and education, are recorded. By way of illustration, we consider the PLR model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \psi(z) + \epsilon,$$

where $y$ is the working hours scaled by 1000, $x_1$ age, $x_2$ family income scaled by 1000 and $z$ as education. Note that among 6340 women in the sample, only 4398 represent propensity to work outside the home, i.e. if that propensity is above the threshold 0, we observe positive hours worked. Therefore, the response variable is left-censored at the value 0.

By way of the second illustration, we consider extramarital Affairs data that is available on the ‘AER’ packages of R. The Affairs dataset, originally reported in [5], was recently re-analyzed by [9] in the censored regression framework. The variables involved in the study of the Affairs dataset are: the number of observations engaged in extramarital sexual intercourse during the past year ($y$), number of years married ($x_1$), occupation according to Hollingshead classification ($x_2$), self rating of marriage ($x_3$) and the age as the nonparametric component $z$. Recommended by [9], the response variable is highly
### Table 3. Model performance and evaluation of the prediction accuracy for the PLR-N-IC, PLR-T-IC, PLR-CN-IC and PLR-SL-IC models.

| Data   | Criterion | $m_1$       | $m_2$       | $m_1$       | $m_2$       | $m_1$       | $m_2$       | $m_1$       | $m_2$       |
|--------|-----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| SLID   | $\ell_{max}$ | -7990.325  | -8082.835  | -7888.270  | -8030.143  | -7941.59    | -8043.783  | -7879.242  | -8015.184  |
| AIC    |           | 16022.650  | 16191.670  | 15820.540  | 16088.290  | 15929.180   | 16117.570  | 15802.480  | 16058.370  |
| BIC    |           | 16164.500  | 16279.480  | 15969.140  | 16182.850  | 16084.540   | 16218.890  | 15951.090  | 16152.930  |
| Affairs| $\ell_{min}$ | -689.7863  | -693.7968  | -678.1139  | -671.7741  | -680.557    | -686.215   | -675.985   | -679.5878  |
| AIC    |           | 1407.573   | 1411.594   | 1386.228   | 1369.548   | 1393.114    | 1400.430   | 1381.970   | 1385.176   |
| BIC    |           | 1469.153   | 1464.377   | 1452.207   | 1426.730   | 1463.492    | 1462.011   | 1447.949   | 1442.357   |

The lowest AIC and BIC scores, among all models and two knots positions, are indicated in bold.
Table 4. ML parameter estimates of the PLR-SMN-IC models for two considered datasets.

| Parameter | SLID Affairs | SLID Affairs | SLID Affairs | SLID Affairs |
|-----------|--------------|--------------|--------------|--------------|
| \( \beta_0 \) | 2.1600       | 11.0125      | 2.2760       | 13.3676      |
| \( \beta_1 \) | -0.0180      | 0.3021       | -0.0179      | 0.3453       |
| \( \beta_2 \) | -0.0031      | 0.3432       | -0.0028      | 0.2699       |
| \( \beta_3 \) | -2.7982      | -2.7982      | -3.0777      | -3.0777      |
| \( \sigma \) | 1.1719       | 8.9666       | 0.9816       | 6.7422       |
| \( \nu \) | -           | -            | 6.8342       | 8.0434       |
| \( \gamma \) | -           | -            | -            | 0.6380       |

left-censored at zero (451 out of 601, or 75%) which may have significant effects on the coefficients.

To each of the dataset, the PLR-N-IC, PLR-T-IC, PLR-CN-IC and PLR-SL-IC models are fitted by implementing the proposed ECME algorithm in Section 4. We use both \( m_1 \) and \( m_2 \) approaches for choosing the number of interior knots. Exploiting the ES and ESQ methods for the location of knots suggests that the ESQ has better performance for analyzing these datasets. Table 3 presents the maximized log-likelihood function, and the values of AIC and BIC for all fitted models with both \( m_1 \) and \( m_2 \) number of interior knots. It can be observed that the PLR-SMN-IC models with heavy tails have better fit than the PLR-N-IC model. Results based on AIC and BIC indicate that the PLR-SL-IC and PLR-T-IC models provide a highly improved fit of the data over other models, for the SLID and Affairs datasets, respectively. Table 4 shows the parameter estimates of the best fitted PLR-SMN-IC sub-models with respect to the number of interior knots, for the SLID and Affairs datasets. It can be seen that all models offer similar estimates for the slope parameters (\( \beta_1, \beta_2 \) and \( \beta_3 \)). Moreover, the parameter estimate \( \nu \) is significantly different from zero and has large values for the PLR-T-IC, PLR-CN-IC, and PLR-SL-IC models, indicating the departure of the data from the normality assumption. Finally, we display in Figure 4 the estimated curve of the nonparametric component \( \psi(z) \) for the fitted PLR-SL-IC and PLR-T-IC models to the SLID and Affairs datasets, respectively.

Figure 4. The curves of the estimated unknown \( \psi(z) \) for the best fitted model to the SLID (left panel) and Affairs (right panel) datasets.
6. Conclusion

This paper proposed a semiparametric inference for partially linear regression model with interval-censored responses where the SMN family of distributions is assumed for the error term. The new model provided an alternative benchmark for the conventional choice of the normal distribution. Our proposed model extended the recent works by [7,8] to the partially linear regression framework. Using the basis spline function, B-spline, we defined the pseudo covariate and pseudo regression parameter vectors. The hierarchical representation of the SMN class of distributions was then exploited for developing a feasible ECME algorithm to obtain the ML parameter estimates.

Four simulation studies were conducted to examine the performance of the proposed model and its parameter estimation. Specifically, simulation studies aimed at checking the ability of the new methodology in parameter recovery, model comparisons and sensitivity analysis in presence of noise points and a single outlier. Finally, two real-world datasets, SLID and Affairs, were analyzed for the purpose of illustration. As was reported, the heavy-tailed PLR-SMN-IC models, such as PLR-T-IC, PLR-CN-IC and PLR-SL-IC, presented better results than the PLR-N-IC model for these real-data examples. All computations were carried out by \textit{R} software and the computer programs are available in the Online Supplement.

The methodology presented in this paper can be extended through the following open issues:

- Motivated from the SLID dataset, it is interesting to formulate a PLR model to simultaneously handle missing and censored observations [16,19].
- Motivated from the Affairs dataset and recommended by [9], one can construct a PLR-SMN model for analyzing doubly-censored data.
- Two important questions that might be raised are: (1) How can we handle multivariate covariates in the nonparametric part? and (2) How can we select the vector of explanatory covariates for both linear and nonparametric parts?. The former issue can be addressed either through the single-index model, generalized additive model, or multivariate B-spline, whereas the latter one can be answered by the variable selection studies.
- As future research, we are exploring in building a finite mixture of semiparametric partially linear regression models as an extension of [24,31] in both likelihood and Bayesian context [17,18].

Acknowledgments

We are very grateful to anonymous associate editor and reviewers for their valuable comments and insightful suggestions that greatly improved this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is based upon research supported by the South Africa National Research Foundation and South Africa Medical Research Council (South Africa DST-NRF-SAMRC SARChI Research
Chair in Biostatistics, Grant number 114613, and STATOMET), as well as by the National Research Foundation of South Africa [grant Number 127727]. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

ORCID

Mehrdad Naderi http://orcid.org/0000-0001-8195-9050
Elham Mirfarah http://orcid.org/0000-0002-8495-5132
Ding-Geng Chen http://orcid.org/0000-0002-3199-8665

References

[1] L.M. Castro, V.H. Lachos, G.P. Ferreira, and R.B. Arellano-Valle, Partially linear censored regression models using heavy-tailed distributions: A Bayesian approach, Stat. Methodol. 18 (2014), pp. 14–31.
[2] H.B. Curry and I.J. Schoenberg, On polya frequency functions iv: The fundamental spline functions and their limits, Tech. Rep., Wisconsin Univ Madison Mathematical Research Center, 1965.
[3] R.F. Engle, C.W. Granger, J. Rice, and A. Weiss, Semiparametric estimates of the relation between weather and electricity sales, J. Am. Stat. Assoc. 81 (1986), pp. 310–320.
[4] R.A. Fagundes, R.M. De Souza, and F.J.A. Cysneiros, Robust regression with application to symbolic interval data, Eng. Appl. Artif. Intell. 26 (2013), pp. 564–573.
[5] R.C. Fair, A theory of extramarital affairs, J. Polit. Econ. 86 (1978), pp. 45–61.
[6] J. Fox, Applied Regression Analysis and Generalized Linear Models, Sage Publications, California, 2015.
[7] A.M. Garay, V.H. Lachos, H. Bolfarine, and C.R.B. Cabral, Linear censored regression models with scale mixtures of normal distributions, Stat. Pap. 58 (2015), pp. 247–278.
[8] A.M. Garay, V.H. Lachos, and T.I. Lin, Nonlinear censored regression models with heavy-tailed distributions, Stat. Interface 9 (2016), pp. 281–293.
[9] W.H. Greene, Econometric Analysis, Pearson Education India, Chennai, 2003.
[10] G. Gómez, M.L. Calle, R. Oller, and K. Langohr, Tutorial on methods for interval-censored data and their implementation in R, Stat. Model. 9 (2009), pp. 259–297.
[11] W. He and G.Y. Yi, Parametric and semiparametric estimation methods for survival data under a flexible class of models, Lifetime Data. Anal. 26 (2019), pp. 369–388.
[12] A.D. Holland, Penalized spline estimation in the partially linear model, J. Multivar. Anal. 153 (2017), pp. 211–235.
[13] J.Z. Huang and L. Liu, Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form, Biometrics 62 (2006), pp. 793–802.
[14] S. Kim and H.R. Cho, Efficient estimation in the partially linear quantile regression model for longitudinal data, Electron. J. Stat. 12 (2018), pp. 824–850.
[15] M.N. Lemus, V.H. Lachos, C.E. Galarza, and L.A. Matos, Estimation and diagnostics for partially linear censored regression models based on heavy-tailed distributions, Stat. Interface 14 (2021), pp. 165–182.
[16] T.I. Lin, V.H. Lachos, and W.L. Wang, Multivariate longitudinal data analysis with censored and intermittent missing responses, Stat. Med. 37 (2018), pp. 2822–2835.
[17] T.I. Lin, J.C. Lee, and H.F. Ni, Bayesian analysis of mixture modelling using the multivariate t distribution, Stat. Comput. 14 (2004), pp. 119–130.
[18] T.I. Lin and T.C. Lin, Robust statistical modelling using the multivariate skew t distribution with complete and incomplete data, Stat. Model. Int. J. 11 (2011), pp. 253–277.
[19] T.I. Lin and W.L. Wang, Multivariate-t linear mixed models with censored responses, intermittent missing values and heavy tails, Stat. Methods Med. Res. 29 (2019), pp. 1288–1304.
[20] C. Liu and D.B. Rubin, The ecm algorithm: A simple extension of em and ecm with faster monotone convergence, Biometrika 81 (1994), pp. 633–648.
Appendices

Appendix 1. Conditional expectations in implementation of the ECME algorithm

Uncensored observations: For the uncensored data \( y \), we have \( \rho = 0 \). Therefore, the only necessary conditional expectation \( \hat{u} = E(U|Y = y, \hat{\Theta}) \) for the considered models can be computed as follows.

- If \( Y \sim N(\mu, \sigma^2) \), in this case, \( U = 1 \) with probability one, and so \( \hat{u} = 1 \).
- If \( Y \sim T(\mu, \sigma^2, \nu) \), we have \( \hat{u} = (\hat{\nu} + 1)/\left(\hat{\nu} + \delta(y, \hat{\mu}, \hat{\sigma})\right) \), where \( \delta(y, \mu, \sigma) = (y - \mu)^2/\sigma^2 \).
- If \( Y \sim SL(\mu, \sigma^2, \nu) \), then

\[
\hat{u} = \frac{2\delta^{-1}(y, \hat{\mu}, \hat{\sigma}) \Gamma\left(\hat{\nu} + 1.5, 0.5\delta(y, \hat{\mu}, \hat{\sigma})\right)}{\Gamma\left(\hat{\nu} + 0.5, 0.5\delta(y, \hat{\mu}, \hat{\sigma})\right)}.
\]

- If \( Y \sim CN(\mu, \sigma^2, \nu, \gamma) \), we have

\[
\hat{u} = \frac{1 - \hat{\nu} + \hat{\nu}\hat{\gamma}^{1.5} \exp\left\{0.5(1 - \hat{\gamma})\delta(y, \hat{\mu}, \hat{\sigma})\right\}}{1 - \hat{\nu} + \hat{\nu}\hat{\gamma}^{0.5} \exp\left\{0.5(1 - \hat{\gamma})\delta(y, \hat{\mu}, \hat{\sigma})\right\}}.
\]

Censored cases: In the censored cases, we have \( \rho = 1 \). For the sake of notation, let

\[
T = \frac{Y - \hat{\mu}}{\hat{\sigma}} \sim SMN(0, 1, \hat{\nu}), \quad \hat{t}_1 = \frac{c_1 - \hat{\mu}}{\hat{\sigma}}, \quad \hat{t}_2 = \frac{c_2 - \hat{\mu}}{\hat{\sigma}}.
\]
Therefore, the necessary conditional expectations for the considered special models can be computed as follows:

\[ \hat{u} = E\left(U|\tilde{t}_1 \leq T \leq \tilde{t}_2, \Theta\right) = \frac{E_{\Phi}\left(1, \hat{t}_2\right) - E_{\Phi}\left(1, \hat{t}_1\right)}{F_{FSMN}\left(\hat{t}_2; \hat{v}\right) - F_{FSMN}\left(\hat{t}_1; \hat{v}\right)}, \]

\[ \hat{u} = \mu \hat{u} + \sigma_j^{(k)} E\left(U|T|\tilde{t}_1 \leq T \leq \tilde{t}_2, \Theta\right) = \mu \hat{u} + \sigma_j^{(k)} \left(\frac{E_{\Phi}\left(0.5, \hat{t}_1\right) - E_{\Phi}\left(0.5, \hat{t}_2\right)}{F_{FSMN}\left(\hat{t}_2; \hat{v}\right) - F_{FSMN}\left(\hat{t}_1; \hat{v}\right)}\right), \]

\[ \hat{u} = \mu \hat{u} + \sigma^2 \left(\frac{E_{\Phi}\left(1, \hat{t}_2\right) - E_{\Phi}\left(1, \hat{t}_1\right) + \hat{t}_1 E_{\Phi}\left(0.5, \hat{t}_1\right) - \hat{t}_2 E_{\Phi}\left(0.5, \hat{t}_2\right)}{F_{FSMN}\left(\hat{t}_2; \hat{v}\right) - F_{FSMN}\left(\hat{t}_1; \hat{v}\right)}\right), \]

where \(E_{\Phi}(r, h) = E(U^r \phi(h \sqrt{U}))\), and \(E_{\Phi}(r, h) = E(U^r \Phi(h \sqrt{U}))\). In the following, the closed forms of \(E_{\Phi}(r, h)\) and \(E_{\Phi}(r, h)\) for the special cases of SMN class of distributions are presented.

- For the normal distribution, we have
  \[ E_{\Phi}(r, h) = \phi(h) \] and \[ E_{\Phi}(r, h) = \Phi(h). \]

- In the case of Student’s-t distribution, we have
  \[ E_{\Phi}(r, h) = \frac{\Gamma\left(\frac{\hat{v} + 2r}{2}\right)}{\sqrt{2\pi} \Gamma(\hat{v}/2)} \left(\frac{\hat{v}}{2}\right) \left(\frac{2}{\hat{h}^2 + \hat{v}}\right) \frac{\hat{v} + 2r}{2}, \]

  \[ E_{\Phi}(r, h) = \Gamma\left(\frac{\hat{v} + 2r}{2}\right) \left(\frac{2}{\hat{v}}\right)^r F_{PVII}\left(h; \hat{v} + 2r, \hat{v}\right) / \Gamma\left(\frac{\hat{v}}{2}\right). \]

  where \(F_{PVII}(\cdot; \nu, \delta)\) denotes the cdf of Pearson type \(VII\) distribution.

- For the slash model, we have
  \[ E_{\Phi}(r, h) = \frac{\hat{v}}{\sqrt{2\pi}} \left(\frac{2}{\hat{h}^2}\right)^{\hat{v}} \Gamma(\hat{v} + r, \frac{\hat{h}^2}{2}) \] and \[ E_{\Phi}(r, h) = \frac{\hat{v}}{\hat{v} + r} F_{SL}\left(h; \hat{v} + r\right). \]

- For the contaminated-normal distribution, we have
  \[ E_{\Phi}(r, h) = (\hat{\gamma})^r \hat{\nu} \phi\left(h \sqrt{\hat{\nu}}\right) + (1 - \hat{\nu}) \phi(h), \]

  \[ E_{\Phi}(r, h) = (\hat{\gamma})^r F_{CN}\left(h; \hat{\nu}, \hat{\gamma}\right) + (1 - (\hat{\gamma})^r) (1 - \hat{\nu}) \Phi(h). \]

**Appendix 2. Additional results of simulation sec4.1**

In order to choose the best approach of the number of interior knots, the model selection criteria can be exploited. Table A1 depicts the average values of the AIC and BIC across all generated samples in Section 4.1. As can be expected, the values of AIC and BIC are increased by exceeding the percentage of censoring. Results in Table A1 suggest that the chosen optimal number of interior knots depends on the level of censoring and the considered model. For instance, the BIC values of the PLR-T-IC model highlight the outperformance of the \(m_2\) approach. On the other hand, for the PLR-CN-IC model, one can conclude that the optimal number of interior knots for a 7.5% censoring level is obtained via \(m_1\) and for 30% via \(m_2\). The number of outperformance of the \(m_2\) approach is, however, significantly higher than \(m_1\) in this simulation study.
Table A1. The mean of AIC and BIC for various sample size and two approaches of the number of interior knots.

| Cens. | \( n \) | \( m_1 \) | \( m_2 \) | \( m_1 \) | \( m_2 \) | \( m_1 \) | \( m_2 \) | \( m_1 \) | \( m_2 \) |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| PLR-N-IC | | | | | | | | | |
| 50    | 189.252 | 188.859 | 221.472 | 221.504 | 251.299 | 251.435 | 209.108 | 209.038 |
| 100   | 372.238 | 370.656 | 438.647 | 437.863 | 497.653 | 497.330 | 412.790 | 410.142 |
| 200   | 736.095 | 731.455 | 865.569 | 859.853 | 982.907 | 985.717 | 817.521 | 808.516 |
| 400   | 1453.720| 1452.680| 1719.688| 1708.525| 1947.138| 1956.133| 1618.665| 1605.993|
| AIC   | 200    | 778.943 | 764.437 | 911.746 | 896.134 | 1032.381| 1025.297| 863.696 | 844.797 |
| 50    | 210.284 | 207.978 | 244.416 | 242.536 | 276.155 | 274.379 | 232.052 | 230.070 |
| 100   | 403.500 | 396.708 | 472.514 | 466.520 | 534.128 | 528.592 | 446.657 | 438.799 |
| 400   | 1513.671| 1496.586| 1783.551| 1756.422| 2014.992| 2008.021| 1682.528| 1653.890|
| 800   | 2942.385| 2942.482| 3514.596| 3462.390| 3953.266| 3957.871| 3324.615| 3255.959|
| BIC   | 50     | 198.278 | 198.278 | 222.721 | 223.604 | 270.302 | 269.439 | 222.132 | 220.470 |
| 100   | 392.823 | 386.803 | 446.061 | 440.721 | 534.041 | 536.024 | 434.102 | 432.319 |
| 400   | 1510.605| 1532.380| 1733.630| 1716.690| 2053.333| 2120.194| 1743.172| 1710.321|
| AIC   | 200    | 768.888 | 777.524 | 873.019 | 862.663 | 1052.913| 1070.035| 868.217 | 857.214 |
| 50    | 223.255 | 217.399 | 245.665 | 244.636 | 295.158 | 292.383 | 245.076 | 241.502 |
| 100   | 424.114 | 412.854 | 479.928 | 470.377 | 570.513 | 567.286 | 467.969 | 460.976 |
| 400   | 1570.476| 1576.286| 1797.493| 1764.587| 2121.188| 2172.082| 1807.035| 1758.219|
| 800   | 3015.453| 3105.677| 3557.307| 3463.809| 4179.317| 4329.259| 3558.681| 3477.770|