Journal of Mathematical Extension
Vol. 20, No. 3 (2026) (4) 1-27
ISSN: 1735-8299
URL: http://doi.org/10.30495/JME.2026.3677
Original Research Paper

Robust Asymmetric Censored Regression Model and Outlier Contaminated Data

O. Moatani
Marv.C., Islamic Azad University

K. Zare*
Marv.C., Islamic Azad University

M. Maleki
University of Isfahan

D. Wraith
Queensland University of Technology

Abstract. This study introduces a robust regression modeling framework for censored observations based on the two-piece scale mixture of normal (TP-SMN) class. Using this distributional family enables exact maximum likelihood inference while simultaneously accommodating skewness, the presence of outliers, and censoring mechanisms. The effectiveness of the proposed methodology is assessed through comprehensive simulation experiments and an empirical analysis using real data, which indicate superior estimation precision and enhanced computational performance. These results highlight the advantages of the proposed modeling censored data with symmetric and asymmetric, as well as light- and heavy-tailed, behavior, offering both methodological flexibility and practical applicability.

Received: February 2026; Accepted: April 2026
*Corresponding Author

AMS Subject Classification: 62J05; 62J07
Keywords and Phrases: Censored regression model, ECME algorithm, ML estimate, TP-SMN.

1 Introduction

Linear regression (LR) models providing a interpretability and flexibility of modeling the responses and covariates (Hastie and Tibshirani [6]). This framework allows certain covariates to have a linear effect, and making LR models widely applicable in economics, biostatistics, and social sciences, particularly when modeling real-world data with heterogeneous effects. In ordinary LR models, the normality assumption is common but often too restrictive in the presence of outliers. To address this issue, Rubio and Genton [12] studied a Bayesian linear regression with skew-symmetric error distributions with applications to survival analysis, using the scale mixture of normal (SMN; proposed by Andrews and Mallows [2]) distributions in LR model to handle light- heavy-tailed errors. The SMN family includes important cases light- heavy-tailed distributions, providing a flexible framework for modeling outliers and achieving robust inference. The asymmetric nature of datasets is another major reason to avoid the normality assumption in statistical models, particularly in LR model. Flexible classes of distributions that are both asymmetric and heavy-tailed can simultaneously address the issues of outliers and asymmetry. In addition to the presence of outliers and asymmetric behavior, which depend on the LR model's error distribution, censored and missing observations are also common challenges in the analysis of real-world data. Censoring is a natural phenomenon that occurs in many datasets for various reasons, such as limitations of measurement instruments that cannot record values beyond certain thresholds. LR model with censored observations are encountered across numerous applied fields such as engineering, economics, biomedical studies, medical surveys, and astronomy. In this context [9] investigated LR model for censored datasets under the presence of outliers by adopting a symmetric heavy-tailed distribution family. Their approach provides robust estimation under such conditions; however, asymmetric behavior in the data was not addressed, which could be incorporated in future

studies to simultaneously handle both outliers and skewness. In this study, we adopt such an approach by considering the highly flexible TP-SMN distribution family, introduced by Maleki and Mahmoudi [8], to address the aforementioned issues. The TP-SMN family also encompasses the symmetric SMN distributions and their asymmetric counterparts, which is an advantage, in addition to covering both light- and heavy-tailed distributions. The remainder of this article is structured as follows. Section 2 outlines key characteristics of the TP–SMN family and develops censored LR (CLR) model derived from this framework, along with an EM-type algorithm for computing maximum likelihood (ML) estimators and the corresponding observed information matrix. Section 3 assesses the performance of the proposed methodology through simulation studies and an empirical example, with comparative analysis against well-known censored regression approaches. Section 4 concludes with some final remarks.

2 TP–SMN Censored Linear Regression Models

This section formulates CLR model with TP–SMN error structures and describes the use of an EM-type algorithm to carry out maximization of the complete-data likelihood function. First, we provide a brief review of key aspects of the TP-SMN distribution family. The classical SMN family proposed by Andrews and Mallows [2], denoted as $SMN(\mu, \sigma, \nu)$, encompasses the symmetric light-tailed normal (N) as well as several symmetric heavy-tailed alternatives, including the Student-t (T) (with the Cauchy distribution obtained when $\nu = 1$), the Slash (SL), and the contaminated normal (CN) distributions. This SMN framework serves as the foundation for constructing the asymmetric TP–SMN family introduced by Maleki and Mahmoudi [8]. Within this extended class, notable members include the asymmetric light-tailed two-piece normal (TP–N) distribution and a variety of asymmetric heavy-tailed counterparts such as the two-piece Student-t (TP–T), which reduces to the two-piece Cauchy when $\nu = 1$, along with the two-piece Slash (TP–SL) and two-piece contaminated normal (TP–CN) distributions. Collectively, these models form a flexible family denoted by TP–SMN$(\mu, \sigma, \gamma, \boldsymbol{\nu})$. It

is worth noting that setting $\gamma = 0.5$ yields the corresponding symmetric SMN distributions as special cases. For a random variable X following the TP–SMN$(\mu, \sigma, \gamma, \boldsymbol{\nu})$ distribution, its probability density function (pdf) and cumulative distribution function (cdf)—both of which are essential components in the development of censored modeling frameworks—are expressed as follows

$$\mathscr{g}(x|\mu,\sigma,\gamma,\boldsymbol{\nu}) = \begin{cases} 2(1-\gamma) f_{smn}(x|\mu,\sigma(1-\gamma),\boldsymbol{\nu}) & x \leqslant \mu, \\ 2\gamma f_{smn}(x|\mu,\sigma\gamma,\boldsymbol{\nu}) & x > \mu, \end{cases}$$

and

$$\mathcal{G}(x|\mu,\sigma,\gamma,\boldsymbol{\nu}) = \begin{cases} 2(1-\gamma) F_{smn}(x|\mu,\sigma(1-\gamma),\boldsymbol{\nu}) & x \leqslant \mu, \\ 2\gamma F_{smn}(x|\mu,\sigma\gamma,\boldsymbol{\nu}) & x > \mu, \end{cases}$$

where $\mu \in \mathbb{R}$ denotes the location parameter, $\sigma > 0$ is the scale parameter, $0 < \gamma < 1$ controls the degree of skewness. The parameter ν corresponds to the degrees of freedom that vary across the members of the family, and $f_{smn}(.|.)$ and $F_{smn}(.|.)$ denote the probability density and cumulative distribution functions of the underlying SMN distributions, respectively. A detailed discussion of the properties and construction of this family can be found in Ghasami et al. [5].

2.1 The model definition

A TP-SMN-CLRM is specified as

$$y_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim TP - SMN(0,\sigma,\gamma,\boldsymbol{\nu}), \quad i = 1,\ldots,n, \tag{1}$$

where y_i denotes the left-censored response from the i-th experimental unit with censoring threshold κ_i defined by

$$y_i = \begin{cases} \kappa_i & y_i \leqslant \kappa_i, \\ y_i & y_i > \kappa_i; \end{cases} \tag{2}$$

where $\boldsymbol{x_i}$ represents a $p \times 1$ vector of explanatory variable values, $\boldsymbol{\beta}$ is a $p \times 1$ fixed parameter vector and the error terms ε_i , $i = 1,\ldots,n$

are assumed to be independent and identically distributed according to the TP–SMN family. The model (1) can be re-expressed the matrix form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ is an $n \times 1$ vector of observed responses, $\boldsymbol{X}$ is an $n \times p$ design matrix with rows $\boldsymbol{x}_i$. Finally, the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ consists of independent random components drawn from the TP–SMN family of distributions. Let $\boldsymbol{\Theta} = (\boldsymbol{\beta}^\top, \sigma, \gamma, \boldsymbol{\nu})^\top$ denote the complete vector of unknown parameters associated with the TP–SMN–CLR models. Assume that, among the n observations, a subset of size m is subject to censoring. Accordingly, the observed response vector $\boldsymbol{y}_{obs}$ can be decomposed into two groups consisting of m censored observations and $n-m$ uncensored ones, such that $\boldsymbol{y}_{obs} = \{\kappa_1, \ldots, \kappa_m; y_{m+1}, \ldots, y_n\}$. Under this setup, the log-likelihood function based on the observed data takes the form

$$\ell(\boldsymbol{\Theta}|\boldsymbol{y_{obs}}) = \log\left[\Pi_{i=1}^{n}\{\mathcal{G}(\kappa_i|\mu_i, \sigma, \gamma, \boldsymbol{\nu})\}^{I_i}\{\mathscr{G}(\kappa_i|\mu_i, \sigma, \gamma, \boldsymbol{\nu})\}^{1-I_i}\right]$$

where $\mu_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}$, and I_i is an indicator variable such that $I_i = 1$ if $y_i \leqslant \kappa_i$, and $I_i = 0$ otherwise. Owing to the computational complexity of direct optimization, the maximum likelihood (ML) estimators of the TP–SMN–CLR model parameters are obtained through the ECME algorithm (Liu and Rubin, [7]), as direct maximization can be cumbersome.

2.2 ML estimation of TP–SMN–CLR model parameters

This section presents the implementation details of the ECME algorithm employed to compute the ML estimators for the parameters of the TP–SMN–CLR model.

2.2.1 ECME algorithm

The hierarchical representation of the TP–SMN–CLR model facilitates ML estimation through the ECM-E algorithm, where some CM-steps use the actual likelihood instead of the expected complete-data log-likelihood. By exploiting the stochastic representation of TP–SMN random variables proposed by Bazrafkan et al. [3], one can derive a hierarchical structure for the TP–SMN–CLR model, summarized in the proposition below.

Proposition 2.1. *Let the TP–SMN–CLR model be defined as in equations (1) and (2). Then, for $i = 1, \ldots, n$, the model admits the following hierarchical representation:*

$$
\begin{aligned}
&Y_i|U_i = u_i, \quad W_i = w_i \overset{iid}{\sim} TN(\boldsymbol{x}_i^\top \boldsymbol{\beta}, u_i^{-1} w_i^2 \sigma^2) I_B(y_i),\\
&P(W_i = w_i) = \gamma^{\frac{w_i+|w_i|}{2w_i}} (1-\gamma)^{\frac{w_i+|w_i|}{2w_i}}; \quad w_i = \gamma, -(1-\gamma),\\
&U_i \overset{iid}{\sim} H(u_i, \boldsymbol{\nu}),
\end{aligned}
$$

where $TN(.)I_B(.)$ denotes the truncated normal distribution on the interval B. Specifically, the truncation set is defined as $B = (\kappa_i, \boldsymbol{x}_i^\top \boldsymbol{\beta}]$ if $w_i < 0$, and $B = (\boldsymbol{x}_i^\top \boldsymbol{\beta}, +\infty)$ if $w_i \geqslant 0$.

Let $\boldsymbol{Z} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_M, u_1, \ldots, u_n, w_1, \ldots, w_n)^\top$ denote the complete data vector, where $\boldsymbol{y}_M = \{y_1, \ldots, y_m\}$ corresponds to the missed (censored) component of the response. Then, the associated complete-data log-likelihood function, ignoring constants, can be written as

$$
\begin{aligned}
\ell_c(\boldsymbol{\Theta}|\boldsymbol{Z}) = &-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{u_i}{w_i^2}e_i^2 + \sum_{i=1}^{n}\frac{w_i+|w_i|}{2w_i}\log\gamma\\
&+\sum_{i=1}^{n}\frac{w_i+|w_i|}{2w_i}\log(1-\gamma),
\end{aligned}
$$

where $e_i = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}$. Throughout the sequel, the notation (r) is used to denote the value of a parameter estimate at the r-th iteration of the algorithm. Within the E-step, the corresponding Q-function is constructed as

$$
Q\left(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = E\left[\ell_{cp}(\boldsymbol{\Theta}|\boldsymbol{Z})|\boldsymbol{y}_{obs}, \widehat{\boldsymbol{\Theta}}^{(r)}\right].
$$

Starting from an initial value $\widehat{\boldsymbol{\Theta}}^{(0)}$, the Q-function is computed using $\widehat{\boldsymbol{\Theta}}^{(r)}$ for $\boldsymbol{\Theta}$, where $\widehat{\boldsymbol{\Theta}}^{(r)} = \left(\hat{\boldsymbol{\beta}}^{\top(r)}, \hat{\sigma}^{(r)}, \hat{\gamma}^{(r)}, \hat{\boldsymbol{\nu}}^{(r)}\right)$. For this purpose, define

$$
\alpha_{ki}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right) = E_{\widehat{\boldsymbol{\Theta}}^{(r)}}\left[U_i W_i^{-2} Y_i^k|\boldsymbol{y}_{obs(i)}\right], \quad k = 0, 1, 2,
$$

and

$$
\zeta_i^{(r)} = E_{\widehat{\boldsymbol{\Theta}}^{(r)}}\left[\frac{|W_i|}{W_i}|\boldsymbol{y}_{obs(i)}\right],
$$

for $i = 1, \ldots, n$. With these quantities, the E-step is completed by expressing the Q-function as

$$\begin{aligned} Q\left(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = &-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big[\alpha_{0i}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)(\boldsymbol{x}_i^\top\boldsymbol{\beta})^2 \\ &- 2\alpha_{1i}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)(\boldsymbol{x}_i^\top\boldsymbol{\beta}) + \alpha_{2i}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)\Big] \\ &+ \sum_{i=1}^{n}\frac{1+\zeta_i\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)}{2}\log\gamma + \sum_{i=1}^{n}\frac{1-\zeta_i\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)}{2}\log(1-\gamma) \end{aligned}$$

At each step, for the censored ith observation ($i = 1, \ldots, m$), we have

$$\alpha_{ki}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right) = E_{\widehat{\boldsymbol{\Theta}}^{(r)}}[U_i W_i^{-2} Y_i^k | Y_i \leqslant \kappa_i], \quad k = 0, 1, 2,$$

and $\zeta_i\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right) = E_{\widehat{\boldsymbol{\Theta}}^{(r)}}[|W_i|/W_i | Y_i \leqslant \kappa_i]$, which are derived by the truncated TP–SMN distributions.

At follows for uncensored ith observation ($i = m+1, \ldots, n$), we have $\boldsymbol{y}_{obs(i)} = Y_i \sim TP - SMN(\boldsymbol{x}_i^\top\boldsymbol{\beta}, \sigma, \gamma, \boldsymbol{\nu})$. Therefore,

$$\alpha_{ki}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right) = y_i^k E_{\widehat{\boldsymbol{\Theta}}^{(r)}}[U_i W_i^{-2} | y_i], \quad k = 0, 1, 2,$$

where the $E_{\widehat{\boldsymbol{\Theta}}^{(r)}}[U_i W_i^{-2}|y_i]$ and $\zeta_i\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right) = E_{\widehat{\boldsymbol{\Theta}}^{(r)}}[|W_i|/W_i]$ are derived by the TP–SMN distributions.

E-step: Given the current parameter estimates $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}^{(r)}$, the conditional expectations $\alpha_{ki}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)$ for $k = 0, 1, 2$, and $i = 1, \ldots, n$ are evaluated using the quantities derived above. Subsequently, the maximization of the Q-function is carried out within the EM framework by solving the system of equations obtained from the first-order optimality condition $\nabla Q\left(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = 0$, where ∇ denotes the gradient operator. This procedure leads to the following parameter update expressions:

CM-step: The parameter vector is updated from $\widehat{\boldsymbol{\Theta}}^{(r)}$ to $\widehat{\boldsymbol{\Theta}}^{(r+1)}$ by maximizing $Q\left(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)}\right)$ with respect to $\boldsymbol{\Theta}$. The resulting parameter

updates are obtained as follows:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \left(\sum_{i=1}^{n} \alpha_{0i} \widehat{\boldsymbol{\Theta}}^{(r)} \boldsymbol{x}_i \boldsymbol{x}_i^{\top}\right)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \alpha_{1i} \widehat{\boldsymbol{\Theta}}^{(r)},$$

$$\hat{\sigma}^{(r+1)} = \sqrt{\frac{\sum_{i=1}^{n} \left[\alpha_{0i}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right) \mu_i^{(r+1)^2} - 2\alpha_{1i}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right) \mu_i^{(r+1)} + \alpha_{2i}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)\right]}{n}},$$

$$\hat{\gamma}^{(r+1)} = \frac{1}{2n} \sum_{i=1}^{n} \left(1 + \zeta_i^{(r)}\right),$$

CML-step: In the final step, the degrees-of-freedom parameter $\boldsymbol{\nu}$ is updated by maximizing the observed-data log-likelihood directly, leading to the following estimate:

$$\hat{\boldsymbol{\nu}}^{(r+1)} = \text{argmax}_{\boldsymbol{\nu}} \ell\left(\widehat{\boldsymbol{\Theta}}_{-\boldsymbol{\nu}}^{(r+1)} | \boldsymbol{y}_{obs}\right) \tag{3}$$

where $\widehat{\boldsymbol{\Theta}}_{-\boldsymbol{\nu}}^{(r+1)}$ denotes the updated $\boldsymbol{\Theta}$ without $\boldsymbol{\nu}$ at the $(r+1)$-th stage of algorithm. A more computationally efficient implementation of the CML-step (3) can be achieved using standard optimization routines available in R, such as optimize or optimx (R Core Team [11]). The algorithm proceeds by alternating between the E-step and the CML-step until convergence is attained. Convergence is determined when a suitable measure of change in the observed-data log-likelihood between successive iterations, for example $\| \ell(\widehat{\boldsymbol{\Theta}}^{(r+1)} | \boldsymbol{Y}, \boldsymbol{X}) / \ell(\widehat{\boldsymbol{\Theta}}^{(r)} | \boldsymbol{Y}, \boldsymbol{X}) - 1 \|$, falls below a pre-specified tolerance threshold

2.3 Approximation of standard errors

This section focuses on deriving standard error estimates for the ML estimators associated with the proposed TP–SMN–CLR model. Under the usual regularity assumptions, the asymptotic covariance matrix of the parameter vector $\boldsymbol{\Theta}$ can be approximated by the inverse of an empirical information matrix, defined as

$$I_0 = \sum_{i=1}^{n} \boldsymbol{s}\left(y_i|\boldsymbol{\Theta}\right) \boldsymbol{s}^{\top}(y_i|\boldsymbol{\Theta}) - n^{-1} \left(\sum_{i=1}^{n} \boldsymbol{s}(y_i|\boldsymbol{\Theta})\right) \left(\sum_{i=1}^{n} \boldsymbol{s}(y_i|\boldsymbol{\Theta})\right)^{\top},$$

where $\boldsymbol{s}(y_i|\boldsymbol{\Theta})$ denotes the score contribution associated with the i-th observation and is given by

$$Q\left(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \sum_{i=1}^{n} Q_i\left(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)}\right).$$

By replacing $\boldsymbol{\Theta}$ with its ML estimate $\widehat{\boldsymbol{\Theta}}$ in I_0, we obtain an approximation in the form of $\hat{I}_0 = \sum_{i=1}^{n} \hat{\boldsymbol{s}}_i \hat{\boldsymbol{s}}_i^{\top}$, where $\hat{\boldsymbol{s}}_i = \boldsymbol{s}(y_i|\widehat{\boldsymbol{\Theta}})$ denotes the individual score contribution evaluated at the estimated parameter values. The score vector can be partitioned as $\hat{\boldsymbol{s}}_i = (\hat{\boldsymbol{s}}_{i,\boldsymbol{\beta}^{\top}}, \hat{\boldsymbol{s}}_{i,\boldsymbol{\sigma}}, \hat{\boldsymbol{s}}_{i,\boldsymbol{\gamma}}, \hat{\boldsymbol{s}}_{i,\boldsymbol{\nu}^{\top}})$, with components given explicitly by

$$\begin{aligned}
\hat{\boldsymbol{s}}_{i,\boldsymbol{\beta}^{\top}} &= -\frac{1}{\sigma^2}\left[\alpha_{0i}(\widehat{\boldsymbol{\Theta}})\boldsymbol{x}_i^{\top}\boldsymbol{\beta} - \alpha_{1i}(\widehat{\boldsymbol{\Theta}})\right]\boldsymbol{x}_i^{\top}, \\
\hat{\boldsymbol{s}}_{i,\boldsymbol{\sigma}} &= -\frac{1}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\left[\alpha_{0i}(\widehat{\boldsymbol{\Theta}})(\boldsymbol{x}_i^{\top}\boldsymbol{\beta})^2 - 2\alpha_{1i}(\widehat{\boldsymbol{\Theta}})\boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}} + \alpha_{2i}(\widehat{\boldsymbol{\Theta}})\right], \\
\hat{\boldsymbol{s}}_{i,\boldsymbol{\gamma}} &= \frac{1+\zeta_i(\widehat{\boldsymbol{\Theta}})}{2\hat{\gamma}} - \frac{1-\zeta_i(\widehat{\boldsymbol{\Theta}})}{2(1-\hat{\gamma})} \\
\hat{\boldsymbol{s}}_{i,\boldsymbol{\nu}^{\top}} &= \frac{\partial}{\partial\boldsymbol{\nu}} E_{\widehat{\boldsymbol{\Theta}}}\left[\log h(u_i;\boldsymbol{\nu})|\boldsymbol{y}_{obs(i)}\right].
\end{aligned}$$

Note that the block part of $\hat{\boldsymbol{s}}_{i,\boldsymbol{\nu}^{\top}}$ involves integral expressions that do not admit closed-form solutions and are therefore evaluated numerically. Finally, standard error estimates for $\widehat{\boldsymbol{\Theta}}$ are obtained as the square roots of the diagonal elements of $\hat{I}_0$.

2.4 Diagnostic analysis

In practice, the estimates from the regression models may be unduly influenced by outliers. After estimating the proposed CLR model, we examine the fitted model to detect outliers and influential observations. To do so, first we obtain the Hessian Matrix of the model, and then use it in a case-deletion approach in Cook [4].

2.4.1 Hessian matrix

By taking derivatives of Q-function at $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}$, we obtain the Hessian matrix

$$\ddot{Q}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial^2 Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^\top}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}}$$

with following non-zero elements:

$$\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial^2 Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}} = -\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\alpha_{0i}(\widehat{\boldsymbol{\Theta}})\boldsymbol{x}_i\boldsymbol{x}_i^\top,$$

$$\ddot{Q}_{\boldsymbol{\nu}\boldsymbol{\nu}}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial^2 Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})}{\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^\top}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}} = \sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^\top}E_{\widehat{\boldsymbol{\Theta}}}\left[\log h(u_i;\boldsymbol{\nu})|\boldsymbol{y}_{obs(i)}\right],$$

$$\ddot{Q}_{\sigma\sigma}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial^2 Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})}{\partial\sigma\partial\sigma}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}} = \frac{n}{\hat{\sigma}^2} - \frac{3}{\hat{\sigma}^4}$$
$$\sum_{i=1}^{n}\left[\alpha_{0i}(\widehat{\boldsymbol{\Theta}})(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})^2 - 2\alpha_{1i}(\widehat{\boldsymbol{\Theta}})(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}) + \alpha_{2i}(\widehat{\boldsymbol{\Theta}})\right],$$

$$\ddot{Q}_{\gamma\gamma}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial^2 Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})}{\partial\gamma\partial\gamma}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}} = \sum_{i=1}^{n}\frac{1-\zeta_i(\widehat{\boldsymbol{\Theta}})}{2(1-\hat{\gamma})^2} - \sum_{i=1}^{n}\frac{1+\zeta_i(\widehat{\boldsymbol{\Theta}})}{2\hat{\gamma}^2},$$

$$\ddot{Q}_{\boldsymbol{\beta}\sigma}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial^2 Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})}{\partial\boldsymbol{\beta}\partial\sigma}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}}$$
$$= \frac{2}{\hat{\sigma}^3}\sum_{i=1}^{n}\left[\alpha_{0i}(\widehat{\boldsymbol{\Theta}})\boldsymbol{x}_i(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} + \boldsymbol{n}_i^\top\hat{\boldsymbol{f}}) - \alpha_{1i}(\widehat{\boldsymbol{\Theta}})\boldsymbol{x}_i\right].$$

2.4.2 Case deletion diagnostic method

Case-deletion is a classical diagnostic technique extensively employed to assess the sensitivity of inferential conclusions to the removal of individual observations. To formalize this procedure, the notation $[-i]$ is used to indicate quantities computed from the dataset with the i-th observation excluded. For instance, the vector $\boldsymbol{y}_{[-i]} = (y_1, \ldots, y_{n-1})^\top$ denotes the $(n-1)\times 1$ response vector obtained after deleting the i-th case. Let $\ell_c(\boldsymbol{\Theta}|\boldsymbol{Z}_{[-i]})$ represent the complete-data log-likelihood evaluated on the reduced dataset. The corresponding ML estimator, obtained

by maximizing the reduced Q-function, is denoted by

$$\widehat{\boldsymbol{\Theta}}_{[-i]} = \left(\hat{\boldsymbol{\beta}}_{[-i]}^{\top}, \hat{\sigma}_{[-i]}, \hat{\gamma}_{[-i]}, \hat{\boldsymbol{\nu}}_{[-i]}^{\top}\right)^{\top}$$

where

$$Q_{[-i]}\left(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = E\left(\ell_c(\boldsymbol{\Theta}|\boldsymbol{Z}_{[-i]})|\boldsymbol{y}_{obs[-i]}, \widehat{\boldsymbol{\Theta}}^{(r)}\right).$$

Here $\widehat{\boldsymbol{\Theta}} = \left(\hat{\boldsymbol{\beta}}^{\top}, \hat{\sigma}, \hat{\gamma}, \hat{\boldsymbol{\nu}}\right)^{\top}$ denotes the ML estimate of $\boldsymbol{\Theta}$ obtained from the full dataset via the ECME algorithm. The difference between $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Theta}}_{[-i]}$ serves as a measure of the influence exerted by the i-th observation in the ML estimates of $\boldsymbol{\Theta}$, such that large values indicate that the i-th case is potentially influential which needs special attention. However, in the case of large data, the procedure requires a large computational effort and raises difficulties. Zhu et al. [13] proposed a computationally efficient one-step pseudo case-deletion approximation to the leave-one-out estimator $\widehat{\boldsymbol{\Theta}}_{[-i]}$, constructed directly from the full-sample ML estimate $\widehat{\boldsymbol{\Theta}}$. Specifically, the approximation is defined as

$$\widehat{\boldsymbol{\Theta}}^{*}_{[-i]} = \widehat{\boldsymbol{\Theta}} + \left[\ddot{Q}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right)\right]^{-1} \dot{Q}_{[-i]}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right)$$

where $\ddot{Q}(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)})$ denotes the Hessian matrix of the Q-function with respect to $\widehat{\boldsymbol{\Theta}}$, evaluated at the ML estimate, and

$$\dot{Q}_{[-i]}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial Q_{[-i]}(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}}$$

is the individual score vector associated with the reduced dataset.
Note that the elements of the individual score vector evaluated at $\widehat{\boldsymbol{\Theta}}$ are

obtained as follows:

$$\dot{Q}_{[-i]\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial Q_{[-i]}(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}}$$
$$= \frac{1}{\hat{\sigma}^2}\sum_{j\neq i}\left[\alpha_{1j}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)\boldsymbol{x}_j - \alpha_{0j}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)\left(\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}\right)\boldsymbol{x}_j\right],$$

$$\dot{Q}_{[-i]\sigma}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial Q_{[-i]}(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}})}{\partial \sigma}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}}$$
$$= -\frac{n-1}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{j\neq i}\left[\alpha_{1j}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)\boldsymbol{x}_j - \alpha_{0j}\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)\left(\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}\right)\boldsymbol{x}_j\right],$$

$$\dot{Q}_{[-i]\gamma}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial Q_{[-i]}(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}})}{\partial \gamma}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}} = \sum_{j\neq i}\frac{1-2\hat{\gamma}+\zeta_i\left(\widehat{\boldsymbol{\Theta}}^{(r)}\right)}{2\hat{\gamma}(1-\hat{\gamma})},$$

$$\dot{Q}_{[-i]\boldsymbol{\nu}}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) = \left.\frac{\partial Q_{[-i]}(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\nu}}\right|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}} = \sum_{j\neq i}\frac{\partial}{\partial\boldsymbol{\nu}}E_{\widehat{\boldsymbol{\Theta}}}[\log h(u_j;\boldsymbol{\nu})|\boldsymbol{y}_{obs[j]}].$$

A simple measure of the distance between $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Theta}}_{[-i]}$ for detecting influential observations is the **Q-displacement** (Zhu et al., [13]):

$$\mathrm{QD}_i = 2\left\{Q\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right) - Q\left(\widehat{\boldsymbol{\Theta}}_{[-i]}|\widehat{\boldsymbol{\Theta}}^{(r)}\right)\right\}, \quad i=1,\ldots,n.$$

Zhu et al. [13] also introduced the generalized Cook's distance:

$$\mathrm{GD}_i = \left(\widehat{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}_{[-i]}\right)^\top\left[-\ddot{Q}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right)\right]\left(\widehat{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}_{[-i]}\right), \quad i=1,\ldots,n.$$

which can be approximated by (for $i=1,\ldots,n$):

$$\mathrm{AGD}_i = \dot{Q}_{[-i]}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right)^\top\left[-\ddot{Q}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right)\right]^{(-1)}\dot{Q}_{[-i]}\left(\widehat{\boldsymbol{\Theta}}|\widehat{\boldsymbol{\Theta}}^{(r)}\right).$$

3 Numerical Results

This section examines the empirical behavior of the proposed TP–SMN–CLR model through an extensive Monte Carlo simulation study, complemented by an analysis of a real dataset previously investigated in the

literature. All computational procedures, including the implementation of the ECME algorithm, were carried out in the R computing environment (R Core Team, [11]), using a workstation equipped with an Intel Core i7-760 processor operating at 2.8 GHz.

3.1 Simulation study

The finite-sample properties of the proposed TP–SMN–CLR model and the associated estimation procedure are investigated through a comprehensive Monte Carlo simulation framework. Specifically, three simulation experiments are conducted under the data-generating mechanism defined in model (1). Left-censored samples are generated across multiple experimental settings, considering censoring proportions of $0\%, 10\%$, 20%, and 30%, together with sample sizes $n \in \{200, 300, 400, 600\}$. For each combination of sample size and censoring rate, 500 independent datasets are simulated from the TP–SMN–CLR model under four distinct distributional configurations for the error term, namely: the two-piece normal CLR (TP–N–CLR), the two-piece Student-t CLR with $(\nu = 4)$ (TP–T–CLR), the two-piece slash CLR with $(\nu = 4)$ (TP–SL–CLR), and the two-piece contaminated-normal CLR with contamination parameters $(\boldsymbol{\nu} = (0.2, 0.2)^\top)$, (TP–CN–CLR) models. We performed all Monte Carlo (MC) simulations by setting $(\boldsymbol{\beta} = (1, 5)^\top), \gamma = 0.45.$, (weak asymmetry) and $\gamma = 0.9$ (strong asymmetry) and $\sigma = 1$, with $\boldsymbol{x}_i = (x_{i1}, x_{i2})^\top$ for which $x_{i1} \sim U(0, 1)$ and $x_{i2} \sim U(1, 2)$ are generated independently.

3.1.1 Recovery of parameters and asymptotic properties

This section summarizes the results of the Monte Carlo simulations conducted for the TP–SMN–CLR model with different censored levels of the observations in each dataset. In each generated sample $k = (1, \ldots, 500)$ with length 300, we obtain the proposed ML estimates of θ (denoting $\boldsymbol{\beta}, \sigma, \gamma$ and $\boldsymbol{\nu}$) denoted by $\hat{\theta}_k$. To assess the empirical behavior of the estimators, we compute the Monte Carlo mean and dispersion measures

across replications. Specifically, the Monte Carlo average is defined as

$$\text{MC-Mean} = \hat{\theta} = \frac{1}{500}\sum_{k=1}^{500}\hat{\theta}_k,$$

while the associated Monte Carlo variability is quantified through the empirical standard deviation

$$\text{SD} = \sqrt{\frac{1}{500}\sum_{k=1}^{500}(\hat{\theta}_k - \hat{\theta})^2}.$$

The numerical summaries corresponding to the various TP–SMN–CLR model specifications are reported in Tables 1 and 2. These results highlight the estimation accuracy and robustness of the proposed models across scenarios exhibiting both weak and strong skewness, as well as across different sample sizes $n = 200, 400$, and 600.

3.1.2 Asymptotic properties

This part of the simulation study is devoted to assessing the asymptotic properties of the proposed estimation procedure for the TP–SMN–CLR model. We generated left-censored samples from the proposed TP–SMN–CLR model with at various censoring levels and moderate asymmetry $\gamma = 0.75$, with several sample sizes ($n = 200, 300, 400$ and 600), over the 500 replications. For each replication, maximum likelihood estimates of the regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ are obtained. The bias and mean squared error (MSE) are then employed as summary measures of estimator accuracy. Specifically, for $i = 1, 2$, these quantities are defined as

$$\text{Bias}(\beta_i) = \frac{1}{500}\sum_{k=1}^{500}(\beta_i - \hat{\beta}_{ik}), \quad \text{MSE}(\beta_i) = \frac{1}{500}\sum_{k=1}^{500}(\beta_i - \hat{\beta}_{ik})^2, \quad i = 1, 2.$$

where $\hat{\beta}_{ik}$ denotes the ML estimate of β_i obtained from the k-th simulated sample.

According to Figures 1 and 2, as the sample size increased, it is naturally observed that, in all of the censoring levels, the Bias and MSE of

ML estimates approach to zero. This convergence toward zero provides empirical support for the consistency and favorable asymptotic behavior of the proposed ML estimators within the TP–SMN–CLR framework.

3.1.3 Robustness of estimates to the outlier

It is important to highlight the robustness property of the TP–SMN–CLR model for attenuating the effects of the outlier observations. We generate 300 samples of TP–N–CLR model from the proposed TP–SMN–CLR model with sample sizes $n = 250$ with moderate censoring levels 10% and 20% and moderate asymmetry $\gamma = 0.75$. In order to detect the outliers, in each sample, we ignored observation Y_{150} and considered its replacement by $Y^*_{150} = Y_{150} + \pi$ with $\pi = 1, 2, \ldots, 10$. Then, for each original and modified dataset, the proposed TP–SMN–CLR model were fitted and the ML estimates of θ (i.e. $\boldsymbol{\beta}, \sigma$ and γ) in generated sample $k = (1, \ldots, 500)$ were denoted by $\hat{\theta}_k$ and $\hat{\theta}_k(\pi)$ average of relative changes (ARC), given by

$$\mathrm{ARC}_\pi(\theta) = \frac{1}{500} \sum_{k=1}^{500} \left| \frac{\hat{\theta}_k(\pi) - \hat{\theta}_k}{\hat{\theta}_k} \right|,$$

was evaluated, this measure increases for larger outliers (See Figure 3). In other words, the TP–T–CLR, TP–SL–CLR and TP–CN–CLR models (as heavy-tailed distributions) are more robust and reasonable than TP–N–CLR model to accommodate atypical observations. Note that we checked the robustness for various values of censoring and skewness and found similar results.

3.2 Real application: Wage rate dataset

In this section, we analyze a real data set which has previously been used in the literature (Mroz [10]) to examine models allowing for censoring. The empirical illustration is based on data drawn from the Panel Study of Income Dynamics (PSID), conducted by the University of Michigan. The sample comprises 753 married White women observed in 1975, whose ages range from 30 to 60 years. Among these individuals, 428 reported positive labor market participation during the reference year.

Table 1: Monte-Carlo average (MC-Mean) and standard deviation (S.D.) of the ML estimates TP–SMN–CLR model (with $\gamma = 0.45$; weakly skewed), in 500 simulated samples for different censored levels (C.L.)

C.L.	Parameter	TP–N–CLR		TP–T–CLR		TP–SL–CLR		TP–CN–CLR	
		MC-Mean	S.D.	MC-Mean	S.D.	MC-Mean	S.D.	MC-Mean	S.D.
0%	$\beta_1(1)$	1.1066	0.1085	1.1933	0.1311	1.1039	0.1103	1.1004	0.1197
	$\beta_2(5)$	5.1450	0.2125	5.1333	0.1836	5.0363	0.2001	5.1635	0.2093
	$\sigma(1)$	1.0137	0.0143	1.0878	0.0221	1.1029	0.0321	1.0074	0.0144
	$\gamma(0.45)$	0.4485	0.0021	0.4520	0.0023	0.4541	0.0043	0.4488	0.0020
	ν	–	–	4.1008	0.1003	4.0821	0.0897	0.2008	0.0090
	τ	–	–	–	–	–	–	0.2092	0.0091
10%	$\beta_1(1)$	0.9655	0.1311	1.1038	0.1103	1.1502	0.1103	1.1211	0.1301
	$\beta_2(5)$	4.9091	0.1321	5.1131	0.1100	5.1210	0.1165	5.1043	0.1059
	$\sigma(1)$	0.9964	0.0143	1.0100	0.0222	1.0198	0.0132	1.0837	0.0199
	$\gamma(0.45)$	0.4486	0.0054	0.4584	0.0078	0.4532	0.0065	0.4566	0.0111
	ν	–	–	4.1094	0.1100	4.1199	0.1045	0.2109	0.0143
	τ	–	–	–	–	–	–	0.2100	0.0136
20%	$\beta_1(1)$	0.8903	0.1012	1.1104	0.0932	1.0201	0.1001	1.1033	0.1011
	$\beta_2(5)$	4.9765	0.1033	5.0543	0.0896	5.1023	0.0876	5.0766	0.1009
	$\sigma(1)$	0.9943	0.0073	1.0105	0.0032	0.9942	0.0093	0.9953	0.0059
	$\gamma(0.45)$	0.4443	0.0018	0.4562	0.0021	0.4542	0.0031	0.4542	0.0035
	ν	–	–	4.0931	0.0931	4.0867	0.0929	0.2034	0.0081
	τ	–	–	–	–	–	–	0.2083	0.0092
30%	$\beta_1(1)$	0.9105	0.1200	1.0977	0.1196	1.0274	0.1117	1.0544	0.1211
	$\beta_2(5)$	5.0231	0.1065	5.0454	0.1533	5.0543	0.1082	5.1029	0.1237
	$\sigma(1)$	0.9982	0.0076	1.0100	0.0036	0.9942	0.0091	0.9973	0.0059
	$\gamma(0.45)$	0.4482	0.0015	0.4582	0.0021	0.4542	0.0035	0.4541	0.0037
	ν	–	–	4.0610	0.1021	3.9011	0.0921	0.1866	0.0121
	τ	–	–	–	–	–	–	0.2056	0.0086

Table 2: Monte-Carlo average (MC-Mean) and standard deviation (S.D.) of the ML estimates TP–SMN–CLR model (with $\gamma = 0.9$; strongly skewed), in 500 simulated samples for different censored levels (C.L.)

C.L.	Parameter	TP–N–CLR		TP–T–CLR		TP–SL–CLR		TP–CN–CLR	
		MC-Mean	S.D.	MC-Mean	S.D.	MC-Mean	S.D.	MC-Mean	S.D.
0%	$\beta_1(1)$	1.0543	0.0987	1.1088	0.0876	0.9346	0.0980	1.0104	0.1200
	$\beta_2(5)$	5.0453	0.1090	5.0876	0.0831	5.0212	0.0897	4.9511	0.1312
	$\sigma(1)$	1.0576	0.0049	1.0201	0.0066	0.9922	0.0054	1.0548	0.0090
	$\gamma(0.45)$	0.9103	0.0032	0.8909	0.0041	0.9132	0.0053	0.9200	0.0040
	ν	–	–	4.1098	0.1032	3.8803	0.1100	0.2155	0.0145
	τ	–	–	–	–	–	–	0.2103	0.0187
10%	$\beta_1(1)$	1.0643	0.1032	1.0843	0.1070	1.1043	0.1153	0.8901	0.1056
	$\beta_2(5)$	5.0654	0.1303	5.1065	0.1179	5.0876	0.1311	5.1211	0.1200
	$\sigma(1)$	1.0142	0.0122	1.0210	0.0149	1.0131	0.0180	0.9807	0.0120
	$\gamma(0.45)$	0.9022	0.0054	0.8978	0.0060	0.9032	0.0079	0.9099	0.0131
	ν	–	–	4.0987	0.1102	3.9065	0.1121	0.2097	0.0162
	τ	–	–	–	–	–	–	0.2088	0.0179
20%	$\beta_1(1)$	1.0170	0.1109	1.0132	0.1098	0.9708	0.1100	1.0345	0.1087
	$\beta_2(5)$	4.9763	0.1037	5.0641	0.1060	5.0723	0.1182	5.0736	0.1201
	$\sigma(1)$	1.0149	0.0080	1.0281	0.0049	0.9968	0.0061	1.0276	0.0115
	$\gamma(0.45)$	0.9001	0.0086	0.9103	0.0121	0.8901	0.0032	0.9102	0.0073
	ν	–	–	4.0871	0.0669	3.9208	0.0801	0.2076	0.0112
	τ	–	–	–	–	–	–	0.2120	0.0093
30%	$\beta_1(1)$	1.0736	0.1165	1.1103	0.1102	1.0201	0.1211	1.0300	0.1102
	$\beta_2(5)$	5.0957	0.1059	5.0721	0.1033	5.1134	0.1543	4.9102	0.1543
	$\sigma(1)$	1.0322	0.0137	0.9841	0.0187	1.0133	0.0101	1.0322	0.0167
	$\gamma(0.45)$	0.9102	0.0040	0.8893	0.0037	0.9072	0.0038	0.8901	0.0054
	ν	–	–	4.0353	0.0601	4.0837	0.0547	0.2001	0.0132
	τ	–	–	–	–	–	–	0.2098	0.0101

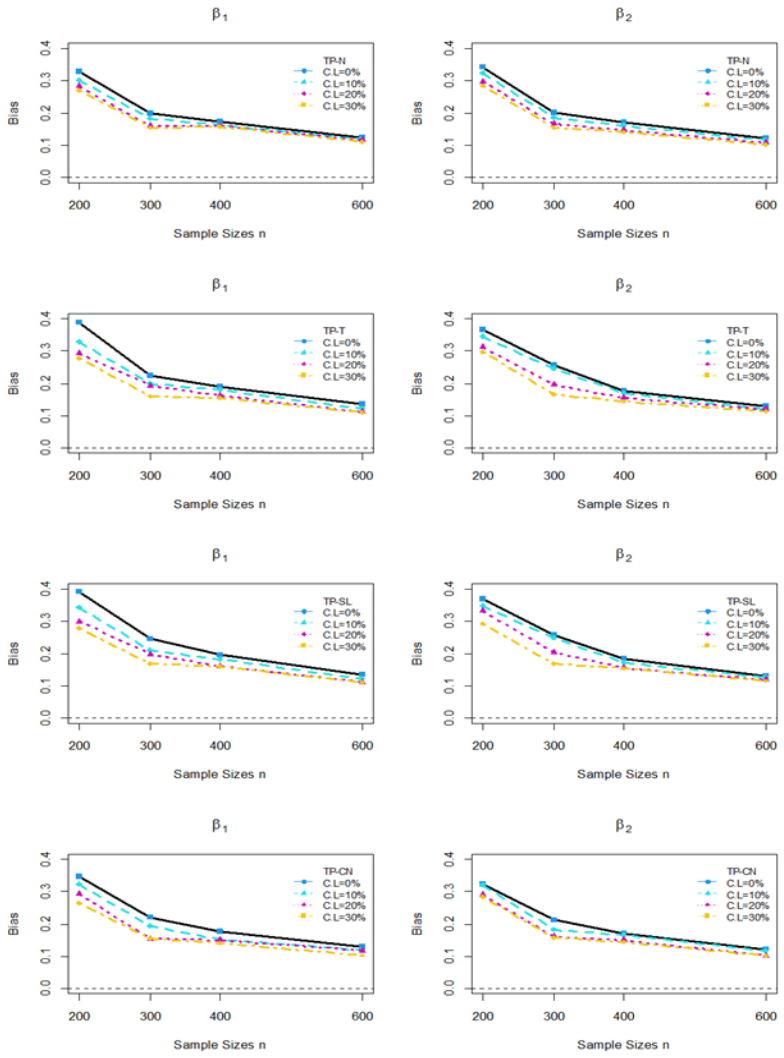

Figure 1: Bias for β_1 and β_2 for different sample sizes and levels of censoring in TP–SMN–CLR model.

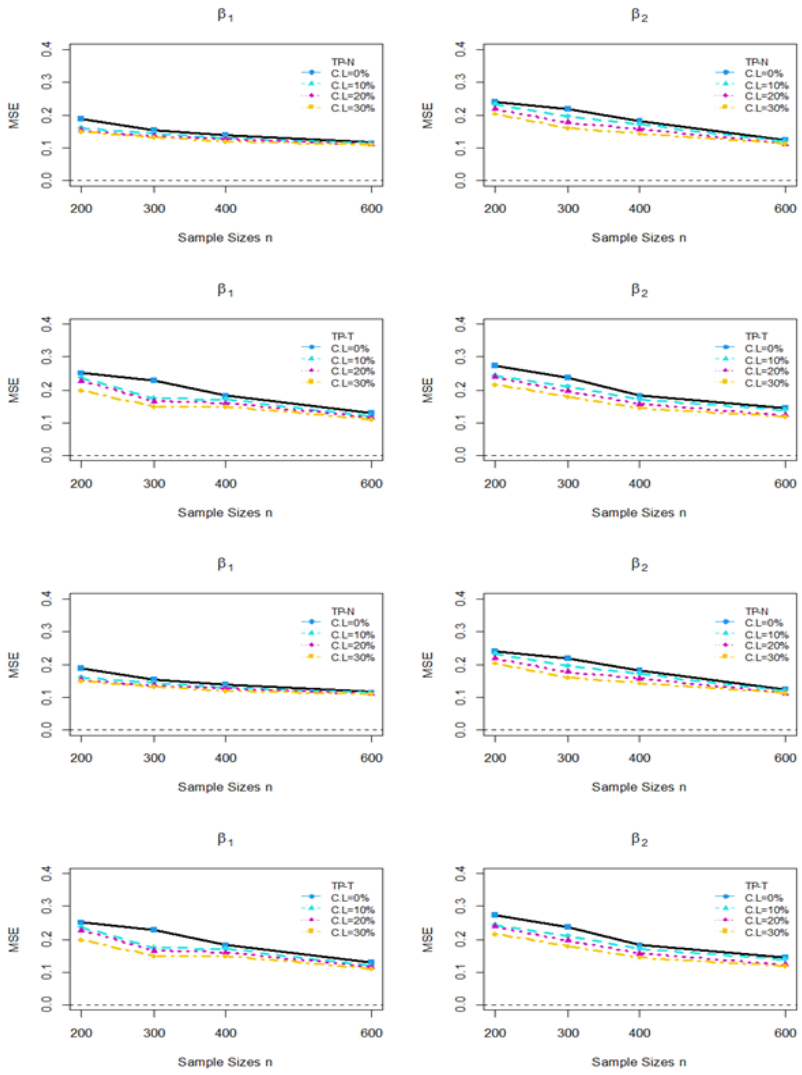

Figure 2: MSE for β_1 and β_2 for different sample sizes and levels of censoring in TP–SMN–CLR model.

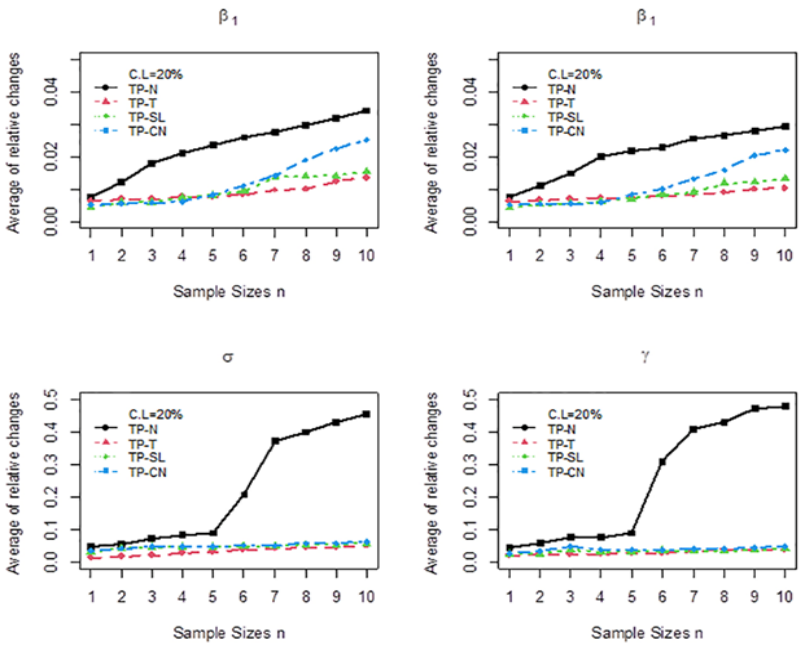

Figure 3: Average relative changes of the ML estimates of the TP–SMN–CLR model parameters at censored level 20%.

The outcome of interest is the wage rate, measured as average hourly earnings. For women who did not participate in paid work during 1975, the recorded wage equals zero, resulting in a naturally left-censored response variable. In addition to the wage outcome, several socioeconomic and demographic covariates are considered, including:

- y : Wage rates (response variable);
- x_1 : Wife's age;
- x_2 : Years of schooling;
- x_3 : Husband's hours worked;
- x_4 : Husband's wage in dollars;

- x_5 : Tax rate faced by the wife;
- x_6 : Number of children younger than six years old in the household;
- x_7 : Number of children between the ages of six and nineteen;

The relation between the wage rate and the number of years the wife worked since age 18 exhibits can be seen in Figure 4.

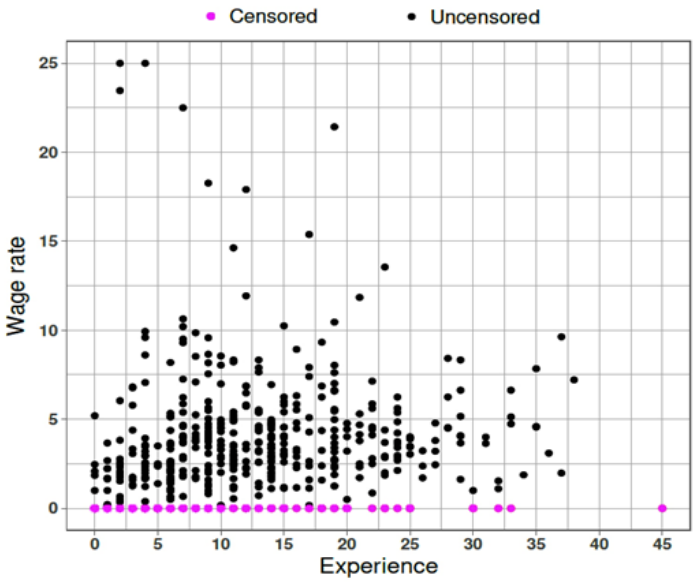

Figure 4: Wage rates vs. number of years the wife worked (Experience).

3.2.1 Analysis of model fitting

We fitted robust and symmetrically/asymmetrically models to this data with the proposed estimates and corresponding standard errors calculated via the empirical information matrix, provided in Tables 3 and 4. The $\ell(\widehat{\boldsymbol{\Theta}})$ and AIC ([**?**]) values in Tables 3 and 4, show that asymmetrical members of the TP–SMN–CLR model are preferred compared to

Table 3: ML estimates of fitted SMN–CLR model to the wage rate data.

Parameter	N–CLR		T–CLR		SL–CLR		CN–CLR	
	ML est.	SE.	ML est.	SE.	ML est.	SE.	ML est.	SE.
β_1	0.7512	(0.0800)	0.6691	(0.0654)	0.6492	(0.0731)	0.6621	(0.0728)
β_2	-0.0599	(0.0211)	-0.0711	(0.0211)	-0.0728	(0.0199)	-0.0707	(0.0200)
β_3	-0.0009	(0.0004)	-0.0005	(0.0003)	-0.0004	(0.0003)	-0.0004	(0.0003)
β_4	-0.1796	(0.0713)	-0.1422	(0.0539)	-0.1501	(0.0600)	-0.1601	(0.0443)
β_5	-8.2911	(2.9736)	-5.6532	(3.1029)	-5.1082	(3.2212)	-6.2221	(3.1001)
β_6	-1.8409	(0.3726)	-1.7042	(0.3025)	-1.7019	(0.3208)	-1.8376	(0.3211)
β_7	0.3401	(0.1594)	0.1998	(0.1019)	0.1938	(0.1071)	0.2101	(0.1037)
σ	3.9842	(1.2401)	2.3432	(0.6301)	2.0012	(0.3028)	2.6102	(0.6809)
ν	-	-	2.8651	(0.4213)	1.1231	(0.1265)	0.1101	(0.0031)
τ	-	-	-	-	-	-	0.1201	(0.0110)
$\ell(\widehat{\boldsymbol{\Theta}})$	-1346.5		-1299.1		**-1293.4**		-1294.6	
AIC	2709		2616.2		**2604.8**		2609.2	

the corresponding symmetrical members (i.e., SMN–CLR model). The proposed TP–T–CLR model has the best performance between all of the proposed symmetrical and asymmetrical models highlighting that it is important to consider models that allow for both heavy tailed and asymmetric behavior.

3.2.2 Diagnostic checking

To assess influential observations, we apply the diagnostic measures from Subsection 2.4 to the dataset. In the ECME algorithm, the scale mixers are estimated as $\alpha_{0i}(\widehat{\boldsymbol{\Theta}}) = \hat{u}_i \hat{w}_i^{-2}$, and plotted against the Mahalanobis distance $d_i^2 = (y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})^2/\hat{\sigma}^2$ for $i = 1, \ldots, 753$ (See Figure 5). For the light-tailed Normal–CLR model, the weights remain close to 1 (red segments in Figure 5). The figure shows that the weights decrease as the Mahalanobis distance increases; large d_i^2 produce small weights. Thus, the heavier-tailed TP–T–CLR assigns smaller weights to potentially influential observations, reducing their impact on parameter estimation. In terms of detecting influential observations, the index plots for the approximate generalized Cook's distance AGD_i, $i = 1, \ldots, 753$, are shown in Figure 6. High values of AGD_i suggest that the i-th observation has

Table 4: ML estimates of fitted TP–SMN–CLR model to the wage rate data.

Parameter	TP–N–CLR		TP–T–CLR		TP–SL–CLR		TP–CN–CLR	
	ML est.	SE.	ML est.	SE.	ML est.	SE.	ML est.	SE.
β_1	0.5908	(0.0733)	0.4653	(0.0987)	0.4421	(0.0701)	0.4701	(0.0726)
β_2	0.0034	(0.0008)	0.0020	(0.0005)	0.0021	(0.0008)	0.0022	(0.0007)
β_3	0.0009	(0.0000)	0.0002	(0.0000)	0.0002	(0.0000)	0.0002	(0.0000)
β_4	-0.0043	(0.0007)	-0.0029	(0.0007)	-0.0031	(0.0006)	-0.0028	(0.0007)
β_5	-5.1310	(1.0209)	-3.5647	(0.8635)	-3.7252	(0.9817)	-3.7625	(0.8999)
β_6	-0.0312	(0.0076)	-0.0290	(0.0061)	-0.0254	(0.0051)	-0.0276	(0.0061)
β_7	0.4107	(0.0332)	0.1208	(0.0123)	0.1209	(0.0141)	0.12130	(0.0142)
σ	4.0726	(1.0716)	2.7827	(0.8262)	2.9723	(0.8010)	2.8972	(0.9001)
γ	0.3526	(0.0091)	0.3016	(0.0065)	0.3100	(0.0098)	0.3122	(0.0099)
ν	-	-	2.0811	(0.3901)	1.1198	(0.4001)	0.1002	(0.0098)
τ	-	-	-	-	-	-	0.1187	(0.0102)
$\ell(\widehat{\boldsymbol{\Theta}})$	-1098.4		**-976.3**		-997.2		-1001.3	
AIC	2214.8		**1972.6**		2014.4		2024.6	

an impact on the ML estimates. According to the top left panel for the TP–N–CLR model, observations #185, #210, #349, #357, #366, #369, #394, #408 and #692 are potentially influential in the ML estimates, but for distributions with heavy tails, i.e., TP–T–CLR, TP–SL–CLR and TP–CN–CLR models, these observations are no longer influential.

4 Conclusion

This work introduces a class of CLR models for censored responses under the TP–SMN family of distributions, which provides a highly adaptable framework capable of accommodating both symmetric and asymmetric behaviors, as well as light- and heavy-tailed error structures. For parameter estimation of the proposed TP–SMN–CLR model, an ECME–type algorithm is carried out that yields analytical forms for parameter estimation and greater accuracy which is a main advantage of the proposed models. The empirical performance of the proposed models is demonstrated through extensive Monte Carlo simulations and an application to real data, highlighting their robustness and practical effectiveness across a variety of scenarios. An R package is currently in development for the

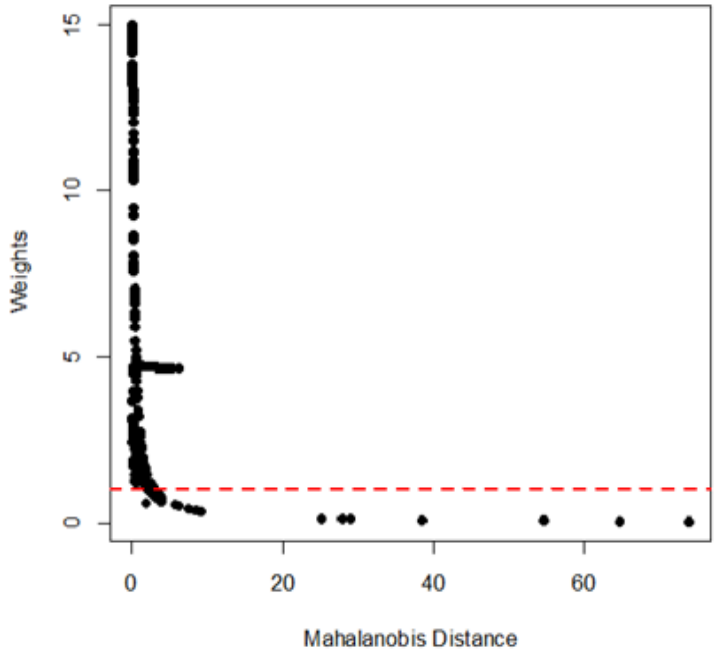

Figure 5: Estimated weights $u_i w_i^{-2}$ versus Mahalanobis distance d_i^2 for fitted TP–T–CLR model on the Wage rate data.

proposed models to allow users to model censored data with greater accuracy and without approximations for the estimation of the model parameters. Future work will extend this framework by formulating a Bayesian version of the TP–SMN–CLR model, with the aim of further increasing modeling flexibility while offering potential gains in computational efficiency and inferential richness.

Compliance with Ethical Standards

Funding:
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest:
The authors declare that they have no conflict of interest.

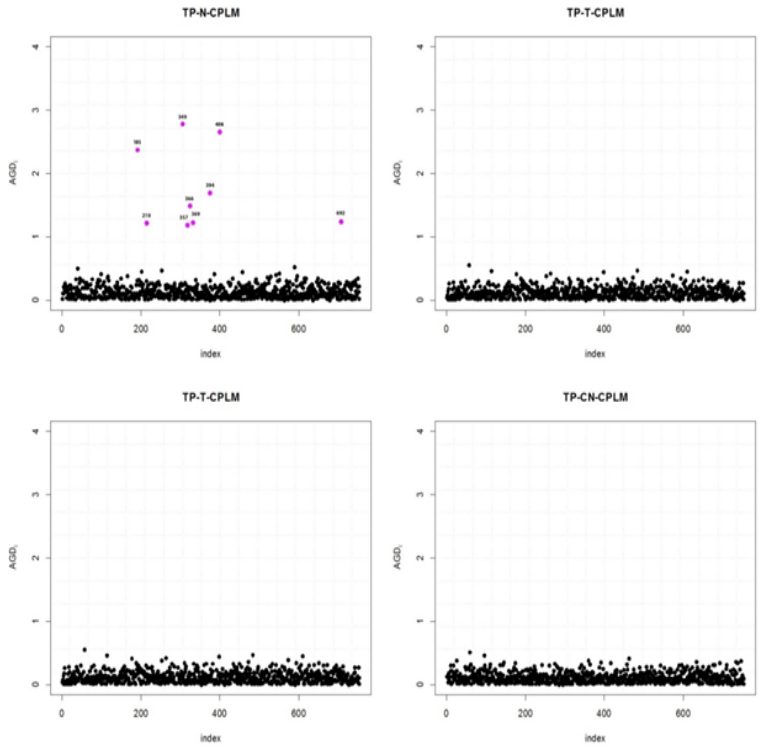

Figure 6: Approximate generalized Cook's distance (AGD_i) for several TP–SMN–CLR model.

Ethical Conduct:
This article does not contain any studies with human participants or animals performed by any of the authors.

Data Availability Statement:
The datasets analyzed in the current study are publicly available within R software packages. Specifically, the data are included in the R packages **AER** (Mroz), **sampleSelection** (Mroz87), **wooldridge** (mroz), and **survival**, all of which are freely accessible through the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org.

References

[1] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Autom. Control*, 19 (1974), 716–723.

[2] D. F. Andrews and C. L. Mallows, Scale mixtures of normal distributions, *J. R. Stat. Soc. Ser. B*, 36 (1974), 99–102.

[3] M. Bazrafkan, K. Zare, M. Maleki and Z. Khodadadi, Partially linear models based on heavy-tailed and asymmetrical distributions, *Stoch. Environ. Res. Risk Assess.*, 36 (2022), 1243–1253.

[4] R. D. Cook, Detection of influential observation in linear regression, *Technometrics*, 19 (1977), 15–18.

[5] S. Ghasami, M. Maleki and Z. Khodadadi, Leptokurtic and platykurtic class of robust symmetrical and asymmetrical time series models, *J. Comput. Appl. Math.*, 376 (2020), 112806.

[6] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, Monographs on Statistics and Applied Probability, vol. 43. Chapman and Hall, London, 1990.

[7] C. Liu and D. B. Rubin, The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence, *Biometrika*, 81 (1994), 633–648.

[8] M. Maleki and M. R. Mahmoudi, Two-piece location-scale distributions based on scale mixtures of normal family, *Commun. Stat. Theory Methods*, 46 (2017), 12356–12369.

[9] M. B. Massuia, A. M. Garay, C. R. B. Cabral, and V. H. Lachos, Bayesian analysis of censored linear regression models with scale mixtures of skew-normal distributions, *Statistics and Its Interface*, 10 (2017), 425–439.

[10] T. A. Mroz, The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, *Econometrica*, 55 (1987), 765–799.

[11] R Core Team, *R: A Language and Environment for Statistical Computing*, Version 4.3.3. R Foundation for Statistical Computing, Vienna, Austria, 2024. Available online: https://www.R-project.org/ (accessed on 29 February 2024).

[12] F. J. Rubio and M. G. Genton, Bayesian linear regression with skew-symmetric error distributions with applications to survival analysis, *Stat. Med.*, 35 (2026), 2441–2454.

[13] H. Zhu, S. Y. Lee, B. C. Wei and J. Zhou, Case-deletion measures for models with incomplete data, *Biometrika*, 88 (2001), 727–737.

Omolbanin Moatani
Department of Statistics,
Ph.D. Student of Statistics,
Marv.C., Islamic Azad University, Marvdasht, Iran.
E-mail: om.moatani@iau.ac.ir

Karim Zare
Department of Statistics,
Associate Professor of Statistics,
Marv.C., Islamic Azad University, Marvdasht, Iran.
E-mail: karim.zare@iau.ac.ir

Mohsen Maleki
Department of Statistics, Faculty of Mathematics and Statistics,
Assistant Professor of Statistics,
University of Isfahan, Isfahan, Iran.
E-mail: m.maleki.stat@gmail.com

Darren Wraith
Institute of Health and Biomedical Innovation,
Associate Professor of Statistics,
Queensland University of Technology, Queensland, Australia.
E-mail: d.wraith@qut.edu.au