UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Influence Diagnostics for Linear Censored Regression Models with Skew-Scale Mixtures of Normal Distributions

Daniel Camilo Fuentes Guzman

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Daniel Camilo Fuentes Guzman

Influence Diagnostics for Linear Censored Regression Models with Skew-Scale Mixtures of Normal Distributions

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree Doctorate of the Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
July 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

F954d

FUENTES GUZMAN, DANIEL CAMILO
Diagnóstico de Influência para Modelos de Regressão Linear Censurada com Misturas de Escala Assimétrica de Distribuições Normais / DANIEL CAMILO FUENTES GUZMAN; orientador FRANCISCO LOUZADA NETO. -- São Carlos, 2024.
91 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Censura. 2. Algoritmo EM. 3. Diagnóstico de Influência. 4. Modelos de Regressão Linear. 5. Distribuições Assimétricas. I. LOUZADA NETO, FRANCISCO , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Daniel Camilo Fuentes Guzman

Diagnóstico de Influência para Modelos de Regressão Linear Censurada com Misturas de Escala Assimétrica de Distribuições Normais

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutor em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
Julho de 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Daniel Camilo Fuentes Guzman, realizada em 24/05/2024.

Comissão Julgadora:

Prof. Dr. Francisco Louzada Neto (USP)

Prof. Dr. Carlos Alberto Ribeiro Diniz (UFSCar)

Profa. Dra. Camila Borelli Zeller (UFJF)

Prof. Dr. Oilson Alberto Gonzatto Junior (USP)

Prof. Dr. Pedro Luiz Ramos (PUC-Chile)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

In memory of my eternally beloved sister Luisa Paola Fuentes Guzman.

ACKNOWLEDGEMENTS

This moment is both unique and profoundly special in my life, symbolizing the culmination of a journey that commenced in my early childhood when I took my first steps at the Yuldaima neighborhood school in my hometown, Ibagué (Tolima-Colombia). It is paramount to convey my gratitude with genuine emotion, recognizing the multitude of factors and individuals who have enabled me to navigate a path of hope, success, and personal growth, ultimately reaching and completing a milestone like this.

Above all, I extend my deep gratitude to God for bestowing upon me the gifts of life, health, a loving family, dear friends, and opportunities throughout this odyssey. My mother, Ofir Antonia Guzman, stands as a fundamental pillar—an unwavering example and my greatest strength. To my sister, Maria Paula, and my nephew, Juan Diego, who are essential components of my daily strength and motivation. I carry eternal gratitude and love for my dear sister, Luisa Paola, now in heaven, my childhood and adolescent companion. She took a part of me with her, leaving everlasting longing and love. I am thankful for her existence and the profound lessons in unconditional love, spiritual strength, and resilience in the face of life's challenges.

To my father, whose influence, even when not present, guided me towards becoming a better and more humane person. His absence provided strength, and his abandonment subtly imparted the lesson of cherishing those around me, fostering within me an unwavering loyalty to those by my side.

Gratitude extends to my extensive and welcoming family, full of uncles, aunts, and cousins, for their company, confidences, and years of shared joys. Special appreciation to my uncles, Guillermo and Javier, for protecting and caring for the family over the years. Fond memories of my grandparents, Gabriel and Ernesto, who are no longer in this world, and my dear grandmothers, Mildred and Evelia.

A heartfelt acknowledgment to my beloved fiancée, Elizabeth Ciampi, for her enduring patience, love, and unwavering support in every aspect—mind and soul. To my dear mother-in-law, Cristina Ciampi, a great friend and, at times, a mother to me in Brazil. Gratitude to my brother-in-law, Igor, for being present in every moment, and to my dear friend and father-in-law, José Roque, now in heaven. Special thanks to Elite for their meals and care at the Ciampi residence.

I extend profound gratitude to my professors at PIPGES, with special acknowledgment to my advisor, Francisco Louzada. His invitation to join his research team opened the door to the

intricate world of project management in data science, exposing me to the challenges of statistics in both Brazilian industry and academia. I am indebted to my academic mentor, Camila Borelli Zeller, for her unwavering support, serving as a constant source of strength, light, and inspiration throughout my postgraduate years in Brazil. My appreciation extends to Professor Carlos Diniz for his insightful suggestions in my studies and invaluable career advice.

Additionally, I hold deep admiration for the professors at UFSCar, including Rafael Izbicki, Rafael Stern, Luis Milan, Vera Tomazella, Alexsandro Gallo, and Daiane Zuanetti, as well as those at ICMC USP—Cibele Russo, Francisco Rodrigues, Jorge Bazán, Ricardo Ehlers, and Josemar Rodrigues. I extend my heartfelt appreciation to Professor Reiko Aoki for her displayed exceptional kindness, guidance and mentorship during my role as a monitor in the Statistics and Regression Models disciplines at ICMC-USP amidst the challenging times of the COVID-19 pandemic in 2020

I extend thanks to my dear friends, Jardel and Oilson, for their companionship and sincere friendship, offering advice, rides, and assistance in moments of need. Gratitude to friends Hans Montcho and Michel Lima for their constant affection. Appreciation to colleagues Átila Correia, Gilberto Pereira, and Rafael Rocha for sharing academic experiences. Special acknowledgment to Giovanni Piccirilli for the rides between UFSCar and ICMC-USP. Heartfelt thanks to my doctoral colleagues—Pedro Ramos, Éder Brito, Milton Miranda Neto, Alex Mota, Marcílio Cardial, Osafu Augustine, Isaac Cortés Olmos, Alex de la Cruz, Marina Gandolfi, Naiara Caroline, Patrícia Stülp, Jessica Barragan, Renato Fernandes and especially Vitor Amorim—for the valuable lessons in analysis and probability during our extensive study journeys.

I express deep gratitude to all the professors and staff at PIPGES, the university dining halls, and the secretariats at both UFSCar and ICMC-USP, for their unwavering support, guidance, and companionship throughout these doctoral years. Thanks to MECAI and CeMEAI for the roles I've undertaken as a tutor, monitor, or researcher in various semesters during my PIPGES doctoral studies. Special thanks to CAPES of Brazil for the financial support for my studies and to PIPGES for the opportunity.

To all of you, my profound gratitude. Your contributions were fundamental to both my academic and personal journey.

This study received partial financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) under Finance Code 001.

“-Eppur si muove-.("And yet it moves.")”
(Galileo Galilei)

“If I have seen further, it is by standing on the shoulders of giants.”
(Isaac Newton)

“It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change.”
(Charles Darwin)

“Nothing in life is to be feared, it is only to be understood.”
(Marie Curie)

“Imagination is more important than knowledge.”
(Albert Einstein)

“Wir müssen wissen, wir werden wissen.”
(David Hilbert)

“Make it a rule never to give a child a book you would not read yourself.”
(Charles Sanders Peirce)

RESUMO

DANIEL CAMILO FUENTES GUZMAN. **Diagnóstico de Influência para Modelos de Regressão Linear Censurada com Misturas de Escala Assimétrica de Distribuições Normais**. 2024. 91 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Nesta pesquisa, conduzimos estudos de diagnóstico de influência local e global para modelos de regressão linear com censura e misturas de escala assimétrica de distribuições normais, propostos por Guzman, Ferreira and Zeller (2020) e denotados como SSMN-CR. Inicialmente, discutimos métodos para gerar dados censurados, apresentando especificamente métodos para gerar dados censurados aleatórios com censura unilateral e intervalar. Posteriormente, abordamos a exclusão de casos e o diagnóstico de influência local com base na função Q, inspirada nas descobertas de Zhu *et al.* (2001) e Zhu and Lee (2001). Para analisar a sensibilidade dos estimadores de máxima verossimilhança dos parâmetros do modelo SSMN-CR a pequenas perturbações nos pressupostos e/ou dados, consideramos vários esquemas de perturbação, como ponderação de casos, variáveis explicativas, variáveis resposta e perturbações nos parâmetros de escala e assimetria. Para ilustrar a utilidade da metodologia proposta, apresentamos a análise de um conjunto de dados reais e três estudos de simulação.

Palavras-chave: Censura, Algoritmo EM, Diagnóstico de Influência, Modelos de Regressão Linear, Distribuições Assimétricas.

ABSTRACT

DANIEL CAMILO FUENTES GUZMAN. **Influence Diagnostics for Linear Censored Regression Models with Skew-Scale Mixtures of Normal Distributions**. 2024. 91 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

In this research, we conducted studies on local and global influence diagnostics for Censored Linear Regression Models with Skew Scale Mixtures of Normal Distributions (SSMN-CR), proposed by Guzman, Ferreira and Zeller (2020). Initially, we discussed methods for generating censored data, specifically presenting methods to generate randomly censored data with both unilateral and interval censoring. Subsequently, we addressed case deletion and local influence diagnostics based on the Q function, inspired by the findings of Zhu *et al.* (2001) and Zhu and Lee (2001). To analyze the sensitivity of the maximum likelihood estimators of the SSMN-CR model parameters to small perturbations in assumptions and/or data, we considered various perturbation schemes, such as case weighting, explanatory variables, response variables, and perturbations in scale and skewness parameters. To illustrate the usefulness of the proposed methodology, we presented the analysis of a real dataset and three simulation studies.

Keywords: Censoring, EM Algorithm, Influence Diagnostics, Linear Regression Models, Skewed Distributions.

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF SOURCE CODES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CR	Censored linear regression models
EM	Expectation-Maximization
MSOM	Mean Shift Outliers Models
SCN	Skew-Contaminated Normal
SMN	Scale Mistures of Normal
SMSN	Scale Mixtures of Skew-Normal
SN	Skew-Normal
SSL	Skew-Slash
SSMN	Skew Scale Mixtures of Normal
SSMN-CR	Censored Linear Regression Models with Skew Scale Mixtures of Normal Distributions
ST	Skew Student-T-Normal

CONTENTS

CHAPTER

1

INTRODUCTION

The analysis of linear regression models with censored data and the detection of outliers are crucial for ensuring the robustness and accuracy of statistical inferences. Censored linear regression models (CR), particularly those dealing with errors distributed according to families that include asymmetry and heavy tails, are of great importance in fields such as health and economics, where censorship and the presence of outliers are common. These models are especially valuable in contexts where measurement limitations or the nature of the phenomena studied result in censored data, often complicating statistical analysis.

The family of Skew Scale Mixtures of Normal (SSMN) distributions, proposed by Ferreira, Bolfarine and Lachos (2011), offers a flexible approach for modeling data with asymmetry and heavy tails. Such characteristics are frequently observed in real-world data that are not well-represented by symmetric distributions. The SSMN-CR model, developed by Guzman, Ferreira and Zeller (2020), extended this approach by incorporating censorship into linear regression and estimating parameters using the Expectation-Maximization (EM) algorithm. The Q-function, derived from the Expectation step of the EM algorithm, forms the basis for the influence diagnostics conducted in this thesis.

This research focuses on global and local influence diagnostics for SSMN-CR models, following the methodology proposed by Zhu and Lee (2001). This methodological choice is strategic, as the Q-function, already available from Guzman's previous work, allows for an in-depth influence analysis without the need to develop new methods. The thesis is distinguished by integrating influence diagnostics techniques with outlier detection methods specifically adapted for SSMN-CR models.

The work significantly advances the analysis of censored data and the robustness of statistical models, providing effective tools for influence diagnostics and outlier detection in SSMN-CR models. Additionally, it addresses the development of mechanisms for generating censored data and applies these models to real and simulated datasets.

The main objectives of this thesis are: To advance the understanding and development of techniques for influence diagnostics and outlier detection in SSMN-CR models. To explore methods for global and local influence diagnostics, integrating measures based on the Q-function and perturbation techniques for sensitivity analysis. To validate the effectiveness of the developed algorithms through simulation experiments, demonstrating their applicability in real-world scenarios.

This work represents a significant advancement in the analysis of censored data and the robustness of statistical models, providing precise and reliable tools for analysis in various application areas, and contributing to the improvement of statistical practices in challenging contexts.

1.1 Background

Regression models have become a cornerstone of statistical modeling across various scientific disciplines. Their strength lies in their broad applicability, allowing researchers to analyze data from diverse phenomena. The research process for regression models typically involves two key steps.

The first step is model inference. This initial phase focuses on calibrating the model. Researchers estimate the model's parameters and evaluate its performance through simulations. This ensures the model's ability to represent the data and the underlying phenomenon accurately.

Following the initial inference, model diagnostics become crucial. This step involves further investigation to assess the model's sensitivity and robustness. Influence analysis, as emphasized by Fung *et al.* (2002) and Zeller *et al.* (2010), plays a vital role here. It helps identify observations that may disproportionately influence the model results, ensuring reliable conclusions.

The methodology for influence diagnostics is well-established, with numerous references demonstrating its application in both symmetric and asymmetric models (Zhu, He and Fung (2003), Zeller *et al.* (2010), Zeller, Lachos and Vilca-Labra (2011), Ferreira, Lachos and Bolfarine (2015), Massuia *et al.* (2015), Matos *et al.* (2019) and Louredo, Zeller and Ferreira (2021)).

However, defining a linear regression model involves establishing various aspects, including the nature of variables, parameters, error terms, and their corresponding distributions. These initial assumptions are critical for accurate analysis.

Real-world phenomena often present challenges, particularly when dealing with censored response variables. Censoring occurs when the complete response variable cannot be observed, either due to limitations in measurement instruments or inherent characteristics of the phenomenon under study. An example is measuring viral load in a living organism, where values below or above a certain detection limit are not quantifiable (ZELLER *et al.*, 2019). Censoring

can arise for various reasons (WU, 2010), and the censored data may also exhibit extreme values, asymmetries, and varying levels of censoring.

The development of robust statistical models for censored data has gained significant traction in recent years. Numerous approaches have been proposed to address the complexities of real-world data from diverse fields, including health, technology, agriculture, and social sciences, among others, such as the works of Arellano-Valle *et al.* (2012) and Garay *et al.* (2017).

The SSMN-CR model, which has the possibility of the presence of censorship in the response variable and errors distributed in the family of SSMN distributions, allows the adequate modeling of phenomena that present outliers and/or asymmetries, as well as having a good hierarchical representation, which allows for easy implementation of inference and the influence diagnosis procedure based on Zhu and Lee's approach. Based on this approach, there are influence analysis studies developed for different types of models. For example, Zeller *et al.* (2010) for skew-normal/independent linear mixed models, Li, Chen and Xie (2012) for heterogeneous log-Birnbaum-Saunders regression models, Ferreira, Lachos and Garay (2020) for heteroscedastic nonlinear regression models under skew-scale mixtures of normal distributions, Ferreira, Zeller and Garcia (2022) for a partially linear heteroscedastic model under skew-normal distribution, and Ferreira, Paula and Lana (2022) for partially linear models with first-order autoregressive skew-normal errors.

The SSMN-CR model shares similarities with skew-elliptical regression models, particularly in their ability to handle censored data. Therefore, investigating influence diagnostics for the SSMN-CR model is a natural progression. This thesis focuses on applying global and local influence analysis approaches, based on Zhu and Lee's methods, to the SSMN-CR model. Influence analysis is a crucial step in data analysis, and this study aims to equip practitioners with tools to identify potentially influential observations.

By leveraging the hierarchical structure of the SSMN-CR model, we propose diagnostic measures derived from the Q-function calculated during the E-step of the EM-type algorithm. This approach avoids the complexities of using the log-likelihood function directly. The proposed methods include case-deletion diagnostics and local influence analysis. These techniques will help researchers identify observations that may significantly impact the analysis and assess the sensitivity of parameter estimates to data perturbations.

The results of this research are organized into four chapters.

In Chapter 2, all the basic concepts necessary to understand the thesis are defined, and the random sampling process of the SSMN-CR models is described.

The Chapter 3, focuses on the study of influence diagnosis techniques for the SSMN-CR model, including several simulation studies and the analysis of a real data set.

In Chapter 4, the final considerations derived from this thesis are presented, and possible directions for future research are suggested, such as Mean Shift Outliers (MSOM) models and

outlier testing techniques for detecting outliers in SSMN-CR models.

CHAPTER

2

RANDOM SAMPLES WITH SSMN-CR MODELS

This chapter defines the theoretical foundations necessary to understand the statistical models covered in this thesis. First, the fundamental concepts are outlined, highlighting their importance in statistical modeling and their applicability across various research contexts. The family of SSMN distributions, which underpins the models studied, is then explored in detail. Special attention is given to the hierarchical and stochastic representation of SSMN distributions, allowing for flexible and adaptable modeling of different types of data.

The chapter then introduces the SSMN-CR Model, detailing its structure and parameter estimation procedures based on the work of Guzman, Ferreira and Zeller (2020). This model provides a robust framework for handling censored data and complex distributions, forming the foundation for the studies conducted in this thesis.

Finally, the chapter discusses approaches to statistical influence diagnostics, laying a solid groundwork for the subsequent chapters, which delve deeper into the analysis of censored data and outlier detection.

2.1 Fundamental Concepts

2.1.1 Statistical censorship

Censorship is a critical concept in statistical analysis, particularly prevalent in survival studies, reliability analysis, economics, and epidemiology. It occurs when complete information about the timing of an event is unavailable due to limitations or restrictions in the study design. This can happen due to factors such as loss to follow-up, study termination before event occurrence, or events not happening within the observation period, as described by Ramos *et al.* (2020). Ignoring censorship can lead to biased analyses and inaccurate conclusions about

the phenomenon under investigation. Understanding and appropriately handling censorship is essential for drawing valid inferences from data.

Several types of statistical censorship exist:

Right Censoring: Right censoring occurs when data is censored after a certain follow-up time or when observations exceed a certain value. For example, in a study of patient survival times, right censoring occurs when a patient is still alive at the end of the study period. Mathematically, right censoring can be represented as $T = \min(T^*, C)$, where T is the observed time, T^* is the actual time until the event of interest occurs, and C is the censoring time. **Example:** In a study on machine failure time, if some machines are still functioning well at the end of the observation period, those observations would be right-censored.

Left Censoring: Left censoring occurs when data is censored before a certain follow-up time or when observations do not reach a certain value. An example would be a study on the lifespan of a product, where products that fail before the start of the study are subject to left censoring. Mathematically, left censoring can be represented as $T = \max(T^*, C)$, where T is the observed time, T^* is the actual time until the event of interest occurs, and C is the censoring time. **Example:** In a study on disease detection time, if some patients have already been diagnosed with the disease before the start of the study, their observations would be left-censored.

Interval Censoring: Interval censoring occurs when data is censored within a specific time or value interval. For example, in a study on injury recovery time, interval censoring may occur when the injury heals between two scheduled measurements. Mathematically, interval censoring can be represented as $T = [L, R]$, where T is the observed time, L is the left endpoint of the interval, and R is the right endpoint of the interval. **Example:** In a study on response time to a medication, if patients are evaluated only at fixed time intervals and the exact response time is not known, the observations are subject to interval censorship.

Type I Censoring: Type I censoring occurs when an experiment is terminated at a predetermined time or after a predetermined number of events. For example, in a reliability study of light bulbs, the experiment may be terminated after a certain number of bulbs fail, and the remaining bulbs are right-censored. **Example:** In a study on the time until a computer program crashes, if the experiment is terminated after a fixed duration, any observations beyond that duration would be right-censored.

Type II Censoring: Type II censoring occurs when an experiment is terminated after a predetermined number of events or failures occur. For example, in a study on the time until a battery fails, the experiment may be terminated after a certain number of failures have occurred, and the remaining batteries are left-censored. **Example:** In a study on the lifespan of animals in a controlled environment, if the experiment is terminated after a fixed number of deaths, any animals that have not died by that point would be left-censored.

2.1.2 Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm is an iterative statistical method used to find maximum likelihood estimates of parameters when data is incomplete or has missing values. It is particularly useful in models with latent variables, which are unobserved variables that influence the observed data. Proposed by Dempster (1977), EM has applications in various fields, including pattern recognition and bioinformatics.

The EM algorithm consists of two steps:

- **Expectation (E-step):** In this step, the algorithm computes the expected value of the complete-data log-likelihood function, conditional on the observed data and current parameter estimates. This involves calculating the conditional expectations of the missing data. Here, the algorithm computes the conditional expectations of the latent variables Z, given the observations Y and the current parameters of the model θ. This is done using the likelihood function $L(\theta;Y,Z)$, which is the joint density function of the observations and the latent variables. Mathematically, the Expectation step is expressed as:

$$Q(\theta|\theta^{(t)}) = E[log\, L(\theta;Y,Z)|Y,\theta^{(t)}],$$

  this step computes an "expectation" of the log-likelihood function conditional on the distribution of the latent variables, with the current parameters of the model $\theta^{(t)}$.

- **Maximization (M-step):** In this step, the model parameters θ are updated to maximize the expected likelihood calculated in the previous step. This is done by adjusting the parameters to increase the joint probability of the observed data Y and the latent variables Z. Mathematically, the Maximization step is expressed as:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}).$$

The process is repeated iteratively until convergence is achieved, i.e., until the parameter values θ do not change significantly between consecutive iterations.

2.1.3 The SSMN-CR model

2.1.3.1 The SSMN distributions

It is important to note that there are some important differences between the classes of SSMN and Scale Mixtures of Skew-Normal (SMSN) distributions; see, for example, Branco and Dey (2001) and Ferreira, Bolfarine and Lachos (2011). The SSMN distributions were defined by Ferreira, Bolfarine and Lachos (2011) through the probability density function (pdf)

$$f_{SSMN}(y|\mu,\sigma^2,\lambda,H) = 2\Phi\left(\lambda\sqrt{d}\right)\int_0^{\infty}\frac{\kappa^{-1/2}(u)}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\kappa^{-1}(u)d\right\}dH(u|\tau), \qquad (2.1)$$

where $d = (y-\mu)^2/\sigma^2$ is the Mahalanobis distance, useful to check the validity of the model and to detect outliers, $\Phi(x|\mu,\sigma^2)$ is the cumulative distribution function (cdf) of the $N(\mu,\sigma^2)$ distribution evaluated at x, $H(u|\tau)$ is the cdf of a positive random variable U indexed by the parameter vector τ, that controls the tails of the distributions, and $\kappa(.)$ is a strictly positive function. In this work, we will consider the Skew-Normal (SN), the Skew Student-T-Normal (ST), the Skew-Slash (SSL) and the Skew-Contaminated Normal (SCN), i.e., when $\kappa(u) = u^{-1}$, whose properties have been widely discussed in Ferreira, Bolfarine and Lachos (2011). Note that when $\lambda = 0$ the PDF (2.1) reduces to the pdf obtained assuming SMN distributions; see Lange and Sinsheimer (1993) for more details. For a random variable with PDF as in (2.1), we use the notation $Y \sim SSMN(\mu,\sigma^2,\lambda,H)$. This class of distributions has a nice hierarchical representation, given by $Y|U=u \sim SN(\mu,\sigma^2u^{-1},\lambda u^{-1/2})$ and $U \sim H(\tau)$, which allows an easy implementation of inference and the influence diagnostic procedure based on Zhu and Lee's approach; see Zhu and Lee (2001) and Zhu *et al.* (2001).

2.1.3.2 The model

In this section, consider the linear regression model with the distributed errors in the family of skew-scale mixtures of normal distributions, as follows

$$Y_i = \mu_i + \xi_i, \quad \xi_i \overset{iid}{\sim} SSMN(0,\sigma^2,\lambda,H),\ i = 1,\ldots,n, \tag{2.2}$$

where the response variable Y_i is continuous for each individual i, ξ_i is a random error, $\mu_i = \mathbf{x}_i^\top\beta$, with β being a p-dimensional vector of unknown regression coefficients and $\mathbf{x}_i$ is assumed a vector of covariates $p\times 1$ known. In addition, ins this paper, we describe the censored regression model in the left censored scenario, that is, the observations are of the form:

$$V_i = \begin{cases} c_i, & \text{if } \rho_i = 1 \text{ (i.e. } Y_i \le c_i), \\ Y_i, & \text{if } \rho_i = 0 \text{ (i.e. } Y_i > c_i), \end{cases} \tag{2.3}$$

for some known threshold point to $c_i, i = 1,\ldots,n$, and ρ_i is the censoring indicator. However, the right censored scenario can be analyzed just by transforming the response V_i to $-V_i$. The model defined in the Equations (2.2) and (2.3) is called of the SSMN-CR model. More details about this model are provided in Guzman, Ferreira and Zeller (2020).

The log-likelihood function of the SSMN-CR model is given by

$$\ell(\theta|\mathbf{v},\rho) = \sum_{i=1}^{n}\rho_i \log\left[F\left(\frac{v_i-\mu_i}{\sigma}\right)\right] + \sum_{i=1}^{n}(1-\rho_i)\log[f_{SSMN}(v_i|\theta,H)], \tag{2.4}$$

where $\theta = (\beta^\top,\sigma^2,\lambda,\tau^\top)^\top$, $\mathbf{v} = (v_1,v_2,\ldots,v_n)$ is the observed sample of $\mathbf{V} = (V_1,V_2,\ldots,V_n)$, $\rho = (\rho_1,\rho_2,\ldots,\rho_n)$, and $\mathrm{F}(\cdot)$ denotes the cdf of the $\mathrm{SSMN}(0,1,\lambda,H)$ distribution. More details on additional properties of this class of distributions can be found in the work of Ferreira, Bolfarine and Lachos (2011).

The SSMN-CR model can be formulated by following hierarchical representation:

$$\begin{array}{rcl} Y_i|U_i=u_i,T_i=t_i & \overset{ind}{\sim} & N\left(\mu_i+\dfrac{\sigma\lambda}{(u_i(u_i+\lambda^2))^{1/2}}t_i,\dfrac{\sigma^2}{u_i+\lambda^2}\right) \\ U_i & \overset{iid}{\sim} & H(\tau) \\ T_i & \overset{iid}{\sim} & TN(0,1;(0,+\infty)),\ i=1,\ldots,n, \end{array} \tag{2.5}$$

all independent, where $TN(r,s;(a,b))$ denotes the univariate normal distribution $(N(r,s))$, truncated on the interval (a,b). It is important to point out that we exploit the hierarchical representation of the SSMN-CR model to derive diagnostic measures, constructed from the Q-function determined in the E-step of the EM-type algorithm instead of the more complicated $\ell(\theta|\mathbf{v},\rho)$. Thus, in this article, for performing influence diagnostics in SSMN-CR model, we use the EM-type algorithm, specifically, the MCEM algorithm.

In the MCEM estimation procedure, let the complete data be $\mathbf{y}_c=(\mathbf{v}^\top,\rho^\top,\mathbf{y}^\top,\mathbf{t}^\top,\mathbf{u}^\top)$, with $\mathbf{y}=(y_1,\ldots,y_n)^\top$, $\mathbf{t}=(t_1,\ldots,t_n)^\top$ and $\mathbf{u}=(u_1,\ldots,u_n)^\top$, where $\mathbf{y}$, $\mathbf{t}$ and $\mathbf{u}$ are treated as hypothetical missing data. Let $\widehat{\theta}^{(k)}=(\widehat{\beta}^{(k)\top},\widehat{\sigma^2}^{(k)},\widehat{\lambda}^{(k)},\widehat{\tau}^{(k)\top})^\top$ denote the estimates of θ at the k-th iteration. Given the current estimate $\widehat{\theta}^{(k)}$ at the kth iteration, we obtain the conditional expectation of the log-likelihood function of complete data given the observed $\mathbf{v}$ and ρ, thus defining the Q-function, whose structure for the model under study is given by $Q(\theta|\widehat{\theta}^{(k)})=\sum_{i=1}^n Q_i(\theta|\widehat{\theta}^{(k)})$, where, excluding unimportant constants,

$$\begin{array}{rcl} Q_i(\theta|\widehat{\theta}^{(k)}) & = & -\log\sigma^2-\dfrac{1}{2\sigma^2}[\widehat{uy^2}_i^{(k)}-2\mu_i\widehat{uy}_i^{(k)}+\mu_i^2\widehat{u}_i^{(k)}+\widehat{t^2}_i^{(k)}-2\lambda\widehat{ty}_i^{(k)} \\ & & +2\lambda\mu_i\widehat{t}_i^{(k)}+\lambda^2(\widehat{y^2}_i^{(k)}-2\mu_i\widehat{y}_i^{(k)}+\mu_i^2)], \end{array} \tag{2.6}$$

where $\widehat{uy}_i^{(k)}=\mathrm{E}[U_iY_i|v_i,\rho_i,\widehat{\theta}^{(k)}]$, $\widehat{uy^2}_i^{(k)}=\mathrm{E}[U_iY_i^2|v_i,\rho_i,\widehat{\theta}^{(k)}]$, $\widehat{u}_i^{(k)}=\mathrm{E}[U_i|v_i,\rho_i,\widehat{\theta}^{(k)}]$, $\widehat{t^2}_i^{(k)}=\mathrm{E}[T_i^2|v_i,\rho_i,\widehat{\theta}^{(k)}]$, $\widehat{ty}_i^{(k)}=\mathrm{E}[T_iY_i|v_i,\rho_i,\widehat{\theta}^{(k)}]$, $\widehat{t}_i^{(k)}=\mathrm{E}[T_i|v_i,\rho_i,\widehat{\theta}^{(k)}]$, $\widehat{y}_i^{(k)}=\mathrm{E}[Y_i|v_i,\rho_i,\widehat{\theta}^{(k)}]$ and $\widehat{y^2}_i^{(k)}=\mathrm{E}[Y_i^2|v_i,\rho_i,\widehat{\theta}^{(k)}]$.

Note that E-step of the developed EM-type algorithm is composed of two parts, one associated with uncensored data and another for censored data. The M-step requires the maximization of $Q(\theta|\widehat{\theta})$ with respect to θ, which leads to closed-form equations.

It is important to point out the special cases, the SN-CR, ST-CR, SSL-CR and SCN-CR models (based on skew-normal, skew Student-t-normal, skew-slash and skew-normal contaminated distributions, respectively). Now, since ξ_i is distributed according to a asymmetric distribution, then the SSMN-CR model may be used along the same line as the skew-elliptical regression models, in the context of censorship. Therefore, a study of analysis diagnostics in SSMN-CR model is a natural way to follow and develop. A influence diagnostics in this model is interesting because the SMN-CR model (linear censored regression model with scale mixtures of normal distributions) is particular case too; see Arellano-Valle *et al.* (2012) and Garay *et al.* (2017). In the section 3.1, we propose influence diagnostics for the SSMN-CR model.

2.1.4 Influence Diagnostics Approaches

2.1.4.1 Global influence approach

For incomplete data problems, Zhu *et al.* (2001) proposed an approach based on the Q-function, using the generalized Cook distance and the Q distance defined by, respectively,

$$GD_{[i]}(\theta) = \left(\widehat{\theta}_{[i]} - \widehat{\theta}\right)^{\top}\{-\ddot{Q}(\widehat{\theta}|\widehat{\theta})\}\left(\widehat{\theta}_{[i]} - \widehat{\theta}\right) \quad \text{and} \quad QD_{[i]}(\theta) = 2\left[Q\left(\widehat{\theta}|\widehat{\theta}\right) - Q\left(\widehat{\theta}_{[i]}|\widehat{\theta}\right)\right], \tag{2.7}$$

where a quantity with a subscript "[i]" means the original quantity with the i-th case deleted, $\widehat{\theta}_{[i]}$ is the maximizer of the Q-function $Q_{[i]}(\theta|\widehat{\theta})$, $i = 1,\ldots,n$, and $\ddot{Q}(\widehat{\theta}|\widehat{\theta}) = \dfrac{\partial^2 Q(\widehat{\theta}|\widehat{\theta})}{\partial\theta\partial\theta^{\top}}$ is the Hessian matrix evaluated at $\theta = \widehat{\theta}$. The Hessian matrix has elements given in Subsection 3.1.1.1. According to Louredo, Zeller and Ferreira (2021), the GD and QD measures provide the same information, then, in this work, let's consider the generalized Cook distance for our purposes. The interest is to consider the influence of the ith observation on some subset of parameters, it can be obtained quite easily as follows:

$$GD_{[i]}(\alpha) = (\widehat{\alpha}_{[i]} - \widehat{\alpha})^{\top}\{\ddot{\mathbf{R}}_{\hat{\alpha}\hat{\alpha}}\}(\widehat{\alpha}_{[i]} - \widehat{\alpha}), \tag{2.8}$$

where $\ddot{\mathbf{R}}_{\hat{\alpha}\hat{\alpha}}$ indicates the entries of the matrix $\ddot{\mathbf{R}} = \{-\ddot{Q}(\widehat{\theta}|\widehat{\theta})\}$ corresponding to the $\alpha = \beta, \sigma^2$ or λ. Next, we describe measures of joint influence and conditional influence; see Lawrance (1995a) and Li, Xu and Zhu (2009).

2.1.4.1.1 Joint Influence

To assess the influence of the observations in set M on the ML estimate $\widehat{\theta}$, the basic idea is to compare the difference between $\widehat{\theta}_{[M]}$ and $\widehat{\theta}$. If deletion of the observations in the set M seriously influence the estimates, more attention should be paid to the obsrvations in M. Hence, if $\widehat{\theta}_{[M]}$ is far from $\widehat{\theta}$, then the observations in M are regarded as influential. The joint influence of the subset M on the $\widehat{\theta}$ can be assessed using

$$GD_{[M]}(\theta) = (\widehat{\theta}_{[M]} - \widehat{\theta})^{\top}\{-\ddot{Q}(\widehat{\theta}|\widehat{\theta})\}(\widehat{\theta}_{[M]} - \widehat{\theta}), \tag{2.9}$$

where $\widehat{\theta}_{[M]}$ are the estimates of θ obtained using the data without observations in M. Note that if, for example, $M = \{i, j\}$ and $GD_{[M]} > GD_{[j]}$ the observation i is said to have an enhancing effect relative to the observation j when θ is estimated, for $i = 1,\ldots,n$ and $i \neq j$. Otherwise, the term reducing effect will be used. More details about these terms are given in Lawrance (1995a) and Li, Xu and Zhu (2009).

2.1.4.1.2 Conditional Influence

Another measure of influence is the conditional influence. The generalized Cook distance for the subset M_2 after prior removal of the subset M_1 from the entire dataset is defined by

$$GD_{[M_2|M_1]}(\theta) = (\widehat{\theta}_{[M_1,M_2]} - \widehat{\theta}_{[M_1]})^{\top}\{-\ddot{Q}(\widehat{\theta}_{[M_1]}|\widehat{\theta}_{[M_1]})\}(\widehat{\theta}_{[M_1,M_2]} - \widehat{\theta}_{[M_1]}). \tag{2.10}$$

Note that if $GD_{[M_2|M_1]}(\theta) > GD_{[M_2]}(\theta)$, the subset M_2 is said to have a masking effect by the subset M_1 when θ is estimated. Otherwise, the term boosting effect will be used. See Lawrance (1995a) and Li, Xu and Zhu (2009) for more details about these terms.

2.1.4.2 Local influence approach

The general approach developed by Zhu and Lee (2001) for local influence analysis of statistical models, in the context of incomplete data, will be utilized to obtain the diagnostic measures for the SSMN-CR model. Let $\omega = (\omega_1, ..., \omega_g)^\top$ be a g-dimensional vector of perturbation varying in an open region $\Omega \subseteq \mathbb{R}^g$. The perturbed complete log-likelihood function is denoted by $\ell_c(\theta, \omega|\mathbf{y}_c) = \log f(\mathbf{y}_c, \theta, \omega)$, where $f(\mathbf{y}_c, \theta, \omega)$ is the probability density function for the complete-data. We assume that there is a ω_0 such that $\ell_c(\theta, \omega_0|\mathbf{y}_c) = \ell_c(\theta|\mathbf{y}_c)$ for all θ. Let $\widehat{\theta}(\omega)$ the maximum of the function $Q(\theta, \omega|\widehat{\theta}) = \mathrm{E}[\ell_c(\theta, \omega|\mathbf{y}_c)|\mathbf{y}, \widehat{\theta}]$. Then, the influence graph is defined as $\alpha(\omega) = (\omega^\top, f_Q(\omega))^\top$, where $f_Q(\omega)$ is the Q-displacement function defined as follows: $f_Q(\omega) = 2\left[Q\left(\widehat{\theta}|\widehat{\theta}\right) - Q\left(\widehat{\theta}(\omega)|\widehat{\theta}\right)\right]$. Following the approach developed by COOK, R. D. (1986) and Zhu and Lee (2001), the normal curvature $C_{f_Q,\mathbf{r}}$ of $\alpha(\omega)$ at ω_0 in the direction of some unit vector $\mathbf{r}$ is given by $C_{f_Q,\mathbf{v}} = -2\mathbf{r}^\top \ddot{Q}_{\omega_0}\mathbf{r}$ and $-\ddot{Q}_{\omega_0} = \Delta_{\omega_0}^\top \left\{ -\frac{\partial^2 Q(\theta|\widehat{\theta})}{\partial\theta\partial\theta^\top}\Big|_{\theta=\widehat{\theta}} \right\}^{-1} \Delta_{\omega_0}$, where $\Delta_{\omega_0} = \frac{\partial^2 Q(\theta, \omega|\widehat{\theta})}{\partial\theta\partial\omega^\top}\Big|_{\theta=\widehat{\theta},\ \omega=\omega_0}$. Since $C_{f_Q,\mathbf{r}}(\theta)$ may assume any value, Zhu and Lee (2001) considered the following conformal normal curvature $B_{f_Q,\mathbf{r}}(\theta) = C_{f_Q,\mathbf{r}}(\theta)/tr[-2\ddot{Q}_{\omega_0}]$, which has an interesting property $0 \leq B_{f_Q,\mathbf{r}}(\theta) \leq 1$, for any unitary direction $\mathbf{r}$. In addition, Zhu and Lee (2001) showed that for all i, $M(0)_i = B_{f_Q,\mathbf{r}_i}$, with $\mathbf{r}_i$ be a basic perturbation vector with ith entry 1 and zero elsewhere.

2.2 Data Generation and Censorship Mechanisms in SSMN-CR Models

To conduct simulation studies with SSMN-CR statistical models, precise data generation mechanisms are crucial for each distribution in the SSMN family: SN, ST, SSL, and SCN. These mechanisms are essential for simulating data used in regression models with errors from these distributions. This chapter outlines the procedure for generating random samples from the SSMN family, a fundamental step for simulating data in censored linear regression models with errors distributed according to the SSMN family (SSMN-CR).

Initially, the generation of random samples following the SSMN distribution is discussed, covering the SN, ST, SSL, and SCN distributions. These distributions are essential for the models studied. Subsequently, unilateral and interval censoring mechanisms are incorporated into the data generation process.

Understanding these mechanisms is vital for simulating realistic data, as censoring often

occurs in survival studies and data analysis when events of interest are not fully observed due to limitations in data collection or follow-up.

Section 2.1.1 provides definitions of censorship, with a focus on unilateral (left or right) and Interval censoring. The objective of this chapter is to offer a comprehensive understanding of data simulation principles for subsequent statistical analyses.

Additionally, the article "Sampling with censored data: a practical guide" Ramos *et al.* (2020), developed alongside this research, is highlighted for its detailed description of methods and techniques for generating random samples under various types of censorship. This study offers practical insights that complement the theoretical methods discussed in this thesis.

2.2.1 Data Generation In Asymmetric Distributions

In Subsection 2.1.3 the family of SSMN distributions was defined. Note that, If $\mu = 0$ and $\sigma^2 = 1$ we refer to this as a **SSMN** standard distribution and denote this by $SSMN(\lambda, H; k)$.

A random variable Y follows a Scale Mistures of Normal (SMN) distribution with the location parameter $\mu \in \mathbb{R}$ and a positive scale parameter σ^2 if your pdf takes the form

$$f_0 = (y; \mu, \sigma^2, \tau) = \int_0^\infty \phi(y; \mu, k(u)\sigma^2) dH(u, \tau), \tag{2.11}$$

where $H(u; \tau)$ is the cdf of a positive random variable U indexed by the parameter vector τ and $k(.)$ is a strictly positive function (See Andrews and Mallows (1974)).

When $\lambda = 0$ in some member of the **SSMN** family, we have the corresponding **SMN** distribution. For the simulated data generation procedure described in this study, it will be necessary to define the asymmetric normal distribution **SN** and its respective stochastic representation.

A random variable Y follows a univariate SN distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ if its pdf is given by

$$f(y) = 2\phi(y; \mu, \sigma^2)\Phi\frac{\lambda(y-\mu)}{\sigma}, \; y \in \mathbb{R}, \tag{2.12}$$

where $\phi(x; \mu, \sigma^2)$ and $\Phi(x; \mu, \sigma^2)$ are the probability density function (pdf) and cumulative distribution function (cdf), respectively, of the Normal distribution $N(\mu, \sigma^2)$ evaluated in x. In the case where $\lambda = 0$, the distribution **SN** becomes a usual Normal Distribution $(Y \sim N(\mu, \sigma^2))$. The marginal stochastic representation is given by

$$Y \stackrel{d}{=} \mu + \sigma[\delta \mid T_0 \mid + (1-\delta^2)^{1/2} T_1] \tag{2.13}$$

with $\delta = \dfrac{\lambda}{(1+\lambda^2)^{1/2}}$ where $\mid T_0 \mid$ stands for the absolute value of T_0, $T_0 \sim N(0,1)$ and $T_1 \sim N(0,1)$ are independent. Here $N(0,1)$ is the standard normal.

To generate data of **SSMN** distribution, we first generate the U distribution and then the $Y|U$ conditional distribution, as illustrated in the following proposition using the stochastic representation of the Asymmetric Normal (**SN**) defined in the Equation (2.13),

Proposição 1. Let $Y \sim SSMN(\mu, \sigma^2, \lambda, H; k)$. So, its stochastic representation is given by

$$\begin{aligned} Y|U=u &\sim SN(\mu, \sigma^2 k(u), \lambda k(u)^{1/2}) \\ U &\sim H(\tau) \end{aligned} \tag{2.14}$$

Further details are described in the Subsection 2.1.3.

2.2.2 Censorship Mechanisms

In this work, we are interested in the situation in which the response variable is not fully observed for all subjects i . Thus, for the $i-$th subject we can assume unilateral (left or right) or interval censoring.

2.2.2.1 One-sided Censorship

Assuming right censoring, $\mathbf{Y_i}$ is a latent variable and the observed data (V_i, ρ_i) take shape

$$V_i = \begin{cases} c_i, & \text{if } \rho_i = 1 \text{ (i.e. } Y_i \leq c_i), \\ Y_i, & \text{if } \rho_i = 0 \text{ (i.e. } Y_i > c_i), \end{cases} \tag{2.15}$$

for some known threshold point c_i, $i = 1, \ldots, n$. The censor indicator $\rho_i = 1$ (or $\rho_i = 0$) means that the $i-$th observation is censored (or uncensored). The extensions of our results to left censoring are immediate: just transform the answer $\mathbf{Y_i}$ and the level of censorship c_i for $-Y_i$ and $-c_i$.

2.2.2.2 Left Censoring

Algorithm 1 – Left Censoring

```
1: procedure LeftCensoring(y, perc)
2:     Sort y in ascending order
3:     n ← length(y)
4:     m ← round(n × perc)
5:     for i ← 1 to m do
6:         y[i] ← −∞
7:     end for
8: end procedure
```

In (1) of Appendix A is the R code implementing the mechanism for generating data with unilateral censoring, specifically left censoring.

2.2.2.3 Rigth Censoring

Algorithm 2 – Right Censoring

```
1: procedure RIGHTCENSORING(y, perc)
2:     Sort y in ascending order
3:     n ← length(y)
4:     m ← round(n × perc)
5:     for i ← 1 to m do
6:         y[n − i + 1] ← ∞
7:     end for
8: end procedure
```

The Code (2) of Appendix A is the R code implementing the generation mechanism with unilateral censoring, specifically right censoring.

2.2.2.4 Interval Censorship

In the case of Interval Censorship for some fixed threshold points c_{i1} and c_{i2}, we will have

$$V_i = \begin{cases} (c_{i1}, c_{i2}), & if\ \rho_i = 1\ (\text{i.}e.\ c_{i1} \leq Y_i \leq c_{i2}), \\ Y_i, & if\ \rho_i = 0\ (\text{i.}e.\ -\infty < Y_i < +\infty), \end{cases} \tag{2.16}$$

Algorithm 3 – Interval Censoring

```
1: procedure INTERVALCENSORING(y, perc)
2:     Sort y in ascending order
3:     n ← length(y)
4:     m ← round(n × perc)
5:     for i ← 1 to m do
6:         y[i] ← −∞
7:         y[n − i + 1] ← ∞
8:     end for
9: end procedure
```

The Code 3 of Appendix A is the R code implementing the mechanism for generating data with interval censoring, for more details see Mirfarah, Naderi and Chen (2021).

2.3 Random Sampling with SSMN Distributions

To simulate **SSMN-CR** data we initially define values for the parameters. The number of generated observations will be indicated by n. The covariates are simulated from a uniform $U(0,1)$. The errors have **SSMN** distribution. Regarding the family of SSMN distributions, there exists the package Sanchez and Ferreira (2016), which provides the density, distribution

function, quantile function, random number generator, likelihood function, direct algorithm, and Expectation-Maximization (EM) algorithm for Maximum Likelihood estimators for a given sample, all for regression models using Skew Scale Mixtures of Normal Distributions. In (4) is a general algorithm illustrating the process of generating random samples from the SSMN distributions see Appendix A.

2.3.1 Algorithm Description

2.3.1.1 Random Sampling with SN Distribution

The `rsnn` function generates random samples from the Skew Normal (SN) distribution, allowing users to specify the location, scale, and shape parameters. If the optional parameter vector `dp` is provided, it overrides the individual parameters. *Algorithm:* 1. Check if the parameters `dp` are provided. If yes, set the location, scale, and shape parameters accordingly. 2. Generate n standard normal variates (`u1` and `u2`). 3. Compute the random samples `y` using the SN distribution formula. The SN distribution is commonly used to model skewed data (LACHOS; CABRAL, 2017). The function utilizes the Box-Muller transformation to generate random samples from a standard normal distribution, which are then transformed to follow the SN distribution.

The Algorithmic (5) in the Appendix A illustrates the function in detail.

2.3.1.2 Random Sampling with ST Distribution

The `rstn` function generates random samples from the skew t-normal distribution. It allows for specifying the location, scale, shape, and degrees of freedom (`nu`) parameters of the distribution. *Algorithm:* 1. Check if the parameters `dp` are provided. If yes, set the location, scale, shape, and degrees of freedom parameters accordingly. 2. Generate n gamma variates (`u`) with degrees of freedom `nu`. 3. Compute the random samples `y` using the ST distribution formula. The skew t-normal distribution is a generalization of the skew normal distribution with heavier tails (SANTOS; LACHOS, 2021). Random samples are generated using the gamma distribution method, with additional transformation steps to incorporate skewness and adjust the scale.

The Algorithmic (6) in the Appendix A illustrates the function in detail.

2.3.1.3 Random Sampling with SSL Distribution

The `rssl` function generates random samples from the skew slash distribution. It allows for specifying the location, scale, shape, and degrees of freedom (`nu`) parameters of the distribution. *Algorithm:* 1. Check if the parameters `dp` are provided. If yes, set the location, scale, shape, and degrees of freedom parameters accordingly. 2. Generate n uniform variates (`v`). 3. Compute the random samples `y` using the SSL distribution formula. This distribution is characterized by a symmetric shape with longer tails. Random samples are generated using the uniform distribution method, followed by transformations to achieve the desired skewness and scale.

The Algorithmic (7) in the Appendix A illustrates the function in detail.

2.3.1.4 Random Samples with SCN Distribution

The `rscn` function generates random samples from the skew normal contaminated distribution. It allows for specifying the location, scale, shape, degrees of freedom (`nu`), and contamination parameter (`gama`) of the distribution. *Algorithm:* 1. Check if the parameters `dp` are provided. If yes, set the location, scale, shape, degrees of freedom, and contamination parameters accordingly. 2. Generate n binomial variates (`uu`) to introduce contamination. 3. Compute the random samples `y` using the SCN distribution formula. This distribution combines elements of the skew normal and contaminated normal distributions, allowing for the modeling of data with skewness and contamination. Random samples are generated using a binomial distribution approach, followed by transformations to introduce skewness and adjust the scale.

The Algorithmic (8) in the Appendix A illustrates the function in detail.

2.4 Simulation Considerations for SSMN-CR Models

In this chapter, we illustrated the generation of random samples from the family of Skew Scale Mixtures of Normal (SSMN) distributions and discussed the applicable censoring mechanisms. To generate random samples for the SSMN-CR model, which includes right-censoring, we performed the following steps:

1. We defined the model parameters, including regression coefficients and SSMN distribution parameters.
2. We generated explanatory variables from standard normal distributions.
3. We constructed the design matrix with the explanatory variables.
4. We calculated the model mean based on the design matrix and regression coefficients.
5. We generated model errors from random samples of the SSMN distribution.
6. We created the response variable by adding the model mean to the model errors.
7. We introduced right-censoring by adjusting values of Y less than the specified cutoff value to the cutoff value (and for left-censoring, we adjusted values of Y greater than the specified cutoff value to the negative cutoff value).

This process enables the generation of data for the SSMN-CR model, considering both the structure of the regression model and the applicable censoring mechanisms.

2.4.1 Diagnostic Analyses through Simulation in SSMN-CR Models

The simulation of data plays a crucial role in studies for SSMN-CR statistical models, allowing for the assessment of robustness and identification of potential weaknesses in the proposed models. Through simulation, it is possible to investigate how SSMN-CR models behave under different scenarios and conditions, especially regarding outlier detection, statistical diagnostic analysis, and bootstrap studies.

In particular, simulation is fundamental for evaluating of SSMN-CR models concerning the presence of outliers and influential observations. This involves generating simulated data with specific characteristics, such as asymmetric distributions and heavy tails, and introducing outliers and influential observations at different proportions and magnitudes. These studies enable understanding how the models respond to different data patterns and identifying potential limitations or vulnerabilities.

Bootstrap studies also benefit from data simulation, allowing for the evaluation of the accuracy and validity of confidence intervals and estimates obtained through this resampling technique. Through simulation, it is possible to investigate the performance of the bootstrap in different contexts and conditions, providing valuable insights for its application in SSMN-CR models.

In summary, data simulation is an essential tool in studies for SSMN-CR models, providing important insights into the effectiveness, and validity of the proposed models. Its use allows for exploring different scenarios and conditions, identifying potential weaknesses, and enhancing the understanding and interpretation of the results obtained.

CHAPTER

3

INFLUENCE DIAGNOSTICS FOR SSMN-CR MODELS

This chapter is organized as follows. In Section 2.1.3, the SSMN-CR model are defined together with the Q-function from the EM-type algorithm. This function is fundamental for diagnostic analysis proposed in Section 3.1, where we deduced the essential elements for calculating the case-deletion measures and the local influence measures for the SSMN-CR model. For the global influence analysis, we discussed the effects of influential observations on parameter estimates through individual exclusion, joint exclusion and conditional exclusion. Furthermore, for the local influence analysis, we proposed the following perturbation schemes: case-weight, explanatory variable, response variable and perturbations in scale and skewness parameters. In Sections 3.2 and 3.3, we present a simulation study and an application on a real dataset of the proposed methodology for the SSMN-CR model, respectively. Conclusions and final considerations are presented in Section 3.4.

3.1 Influence diagnostics

After the model be fitted a key step is a influence diagnostic; see Cook (1977), Cook and Weisberg (1982), COOK, R. D. (1986), Zhu *et al.* (2001) and Zhu and Lee (2001), for example. In this section, we propose the case-deletion technique and the local influence approach in order to identify and understand the observations that may affect the analysis for the SSMN-CR model.

In the next subsections, we introduce the essential elements for calculating the case-deletion measures and the local influence measures for the SSMN-CR model. We first consider the case-deletion measures, then the local influence and finally the perturbation schemes used.

In Section 2.1.4, we present the description of the general methodology for the analysis of influence diagnostics in the global and local contexts. In particular, the global influence, joint influence and conditional influence techniques are described as well as the local influence

approach considered in SSMN-CR model.

3.1.1 Global influence approach

In Section 2.1.4, we present the measures for performing the global influence analysis, known as the case deletion analysis (COOK, 1977), for the context of incomplete data problems. Note that the generalized Cook distance (GD) and the Q distance (QD) are defined in the same manner to the usual Cook's measures, we here use the Q-function in place of the genuine likelihood; see Zhu *et al.* (2001) and Zhu and Lee (2001). According to Louredo, Zeller and Ferreira (2021), the GD and QD measures provide the same information, then, in this work, let's consider the generalized Cook distance for our purposes. An observation is considered influential if its deletion generates considerable impact on the estimates; see Cook and Weisberg (1982). The notation $GD_{[i]}$ means that it is the GD measure calculated considering the exclusion of observation i from the dataset. In this work, we consider the i-th observation to be influential if $GD_{[i]}$ is larger than the cutoff value: $\overline{\mathbf{GD}} + c^* \times sd(\mathbf{GD})$, where $\mathbf{GD}$ is the vector with all the values of $GD_{[i]}$, c^* is a selected constant, $\overline{\mathbf{GD}}$ and $sd(\mathbf{GD})$ are the mean and standard deviation of $\{GD_{[i]}\ i = 1, \ldots, n\}$, respectively. The choice of c^* is subjective; here we consider $c^* = 2$.

Additionally, a joint global influence analysis is performed as well as a conditional global influence analysis for the SSMN-CR models. Inspired by the works of Lawrance (1995b), Xu, Lee and Poon (2006) and Li, Xu and Zhu (2009), we consider a more general approach of case deletion, with M being the set of indices of the selected observations on which we want to assess the influence. According to Lawrance (1995b), if the diagnostic analysis is combined, considering the joint influence and the conditional influence, then it will be more effective.

3.1.1.1 The hessian matrix

The hessian matrix is an essential element in the method developed by Zhu *et al.* (2001) in order to obtain the diagnostic measures. Hence, compute

$$\frac{\partial^2 Q(\theta|\widehat{\theta})}{\partial\theta\partial\theta^\top}\Big|_{\theta=\widehat{\theta}} = \sum_{i=1}^{n} \frac{\partial^2 Q_i(\theta|\widehat{\theta})}{\partial\theta\partial\theta^\top}\Big|_{\theta=\widehat{\theta}},$$

where

$$\frac{\partial^2 Q_i(\theta|\widehat{\theta})}{\partial\beta\partial\beta^\top}\Big|_{\theta=\widehat{\theta}} = -\frac{1}{\widehat{\sigma^2}}\mathbf{x}_i[\widehat{u}_i + \widehat{\lambda}^2]\mathbf{x}_i^\top,$$

$$
\begin{aligned}
\frac{\partial^2 Q_i(\theta|\widehat{\theta})}{\partial\sigma^2\partial\sigma^2}\Big|_{\theta=\widehat{\theta}} &= \frac{1}{\widehat{\sigma^2}^2} - \frac{1}{\widehat{\sigma^2}^3}[\widehat{uy^2}_i - 2\widehat{\mu}_i\widehat{uy}_i + \widehat{\mu}_i^2\widehat{u}_i + \widehat{t^2}_i - 2\widehat{\lambda}\widehat{ty}_i + 2\widehat{\lambda}\widehat{\mu}_i\widehat{t}_i] - \\
&\quad \frac{1}{\widehat{\sigma^2}^3}\widehat{\lambda}^2(\widehat{y^2}_i - 2\widehat{\mu}_i\widehat{y}_i + \widehat{\mu}_i^2), \\
\frac{\partial^2 Q_i(\theta|\widehat{\theta})}{\partial\lambda\partial\lambda}\Big|_{\theta=\widehat{\theta}} &= -\frac{1}{\widehat{\sigma^2}}[\widehat{y^2}_i - 2\widehat{\mu}_i\widehat{y}_i + \widehat{\mu}_i^2], \\
\frac{\partial^2 Q_i(\theta|\widehat{\theta})}{\partial\sigma^2\partial\beta^\top}\Big|_{\theta=\widehat{\theta}} &= -\frac{1}{\widehat{\sigma^2}^2}[\widehat{uy}_i - \widehat{\mu}_i\widehat{u}_i - \widehat{\lambda}\widehat{t}_i + \widehat{\lambda}^2(\widehat{y}_i - \widehat{\mu}_i)]\mathbf{x}_i^\top, \\
\frac{\partial^2 Q_i(\theta|\widehat{\theta})}{\partial\lambda\partial\beta^\top}\Big|_{\theta=\widehat{\theta}} &= \frac{1}{\widehat{\sigma^2}}[-\widehat{t}_i + 2\widehat{\lambda}(\widehat{y}_i - \widehat{\mu}_i)]\mathbf{x}_i^\top, \\
\frac{\partial^2 Q_i(\theta|\widehat{\theta})}{\partial\sigma^2\partial\lambda}\Big|_{\theta=\widehat{\theta}} &= \frac{1}{\widehat{\sigma^2}^2}[-\widehat{ty}_i + \widehat{\mu}_i\widehat{t}_i + \widehat{\lambda}(\widehat{y}_i^2 - 2\widehat{\mu}_i\widehat{y}_i + \widehat{\mu}_i^2)].
\end{aligned}
$$

3.1.2 Local influence approach

Alternatively to the global influence method that evaluates the influence of excluding one or a group of observations in the estimation process of the SSMN-CR model, the local influence approach will evaluate the sensitivity of the model under small perturbations in the model (or data).

In Section 2.1.4, we describe this influence method, for the context of incomplete data problems. Next, we deduce expressions of the matrices required to implement some meaningful perturbation schemes under the SSMN-CR model.

3.1.2.1 Perturbation schemes for the SSMN-CR model

For each perturbation scheme, one has the partitioned form $\Delta\omega_0 = (\Delta_\beta^\top, \Delta_{\sigma^2}^\top, \Delta_\lambda^\top)^\top$, for $\Delta\alpha = (\Delta_1^\top\alpha, \ldots, \Delta_g^\top\alpha)^\top$, $\alpha = \beta, \sigma^2$ or λ and $\Delta_l\alpha = \frac{\partial^2 Q(\theta,\omega|\widehat{\theta})}{\partial\alpha\partial\omega_l}\Big|_{\theta=\widehat{\theta},\ \omega=\omega_0}, l = 1,\ldots,g,$ where $\Delta_\beta = \frac{\partial^2 Q(\theta,\omega|\widehat{\theta})}{\partial\beta\partial\omega^\top}\Big|_{\theta=\widehat{\theta},\ \omega=\omega_0} \in \mathbb{R}^{p\times g}$, $\Delta_{\sigma^2} = \frac{\partial^2 Q(\theta,\omega|\widehat{\theta})}{\partial\sigma^2\partial\omega^\top}\Big|_{\theta=\widehat{\theta},\ \omega=\omega_0} \in \mathbb{R}^{1\times g}$, $\Delta_\lambda = \frac{\partial^2 Q(\theta,\omega|\widehat{\theta})}{\partial\lambda\partial\omega^\top}\Big|_{\theta=\widehat{\theta},\ \omega=\omega_0} \in \mathbb{R}^{1\times g}$ and g is the dimension of the perturbation vector ω. Note that for the local influence analysis in SSMN-CR model, we consider $g = n$. The quantity $M(0)_i = B_{f_Q,\mathbf{r}_i}$ proposed by Zhu and Lee (2001), described in Appendix A, allows us to compare the curvatures among different special schemes of perturbations of SSMN-CR model. The i-th case may be regarded as influential if $\{M(0)_i, i = 1,\ldots,n\}$ is larger than the cutoff value. Lee and Xu (2004) propose to use $1/n + c^* \times sd(\mathbf{M(0)})$ as a benchmark, where $\mathbf{M(0)}$ is the vector with all the values of $M(0)_i$, $sd(\mathbf{M(0)})$ is the standard deviation of $\{M(0)_i,\ i = 1,\ldots,n\}$ and here we consider $c^* = 2$.

(i) *C*ase-weight perturbation

This perturbation scheme may capture departures in general directions, such as observation that can exercise high influence on the $\widehat{\theta}$ due to its outstanding contribution of the log-likelihood function. First, we consider an arbitrary attribution of weights for the Q-function, given in the Equation (2.6), that results in $Q(\theta,\omega|\widehat{\theta}) = \sum_{i=1}^{n} \omega_i Q_i(\theta|\widehat{\theta})$, where $0 \leq \omega_i \leq 1$. In this case, $\omega_0 = \mathbf{1}_n$, where $\mathbf{1}_n$ is the n-vector of ones. For this perturbation scheme, we find

$$\begin{aligned}
\Delta_{i\beta} &= \frac{1}{\widehat{\sigma^2}}\mathbf{x}_i[\widehat{uy}_i - \widehat{\mu}_i\widehat{u}_i - \widehat{\lambda}\widehat{t}_i + \widehat{\lambda}^2(\widehat{y}_i - \widehat{\mu}_i)], \\
\Delta_{i\sigma^2} &= -\frac{1}{\widehat{\sigma^2}} + \frac{1}{2\widehat{\sigma^2}^2}[\widehat{uy^2}_i - 2\widehat{\mu}_i\widehat{uy}_i + \widehat{\mu}_i^2\widehat{u}_i + \widehat{t^2}_i \\
&\quad -2\widehat{\lambda}(\widehat{ty}_i - \widehat{\mu}_i\widehat{t}_i) + \widehat{\lambda}^2(\widehat{y^2}_i - 2\widehat{\mu}_i\widehat{y}_i + \widehat{\mu}_i^2)], \\
\Delta_{i\lambda} &= -\frac{1}{\widehat{\sigma^2}}[-\widehat{ty}_i + \widehat{\mu}_i\widehat{t}_i + \widehat{\lambda}(\widehat{y^2}_i - 2\widehat{\mu}_i\widehat{y}_i + \widehat{\mu}_i^2)].
\end{aligned}$$

Note that for $\omega_i = 0$ and $\omega_j = 1$, $j \neq i$, the ith observation is dropped from the estimation of θ. Alternatively, following Leiva *et al.* (2007), we can consider two other sub-perturbations: *N*on-censoring case perturbation: when $\omega_i = 1$ for $\rho_i = 0$ in the Equation (2.3), where only the non-censoring observations are perturbed and *C*ensoring case perturbation: when $\omega_i = 1$ for $\rho_i = 1$ in the Equation (2.3), where only the censoring observations are perturbed.

(ii) *R*esponse perturbation

The response perturbation can indicate observations with large influence on their own predicted values. In this perturbation and in (iii) *E*xplanatory perturbation, the response and explanatory variables are modified through additive ("a") and multiplicative ("m") perturbation schemes. According to Castro, Galea-Rojas and Bolfarine (2007), we can interpret additive and multiplicative disturbances as absolute and relative changes of the data, respectively. We return to the working model by taking $\omega_0 = \mathbf{0}_n$ in the additive case and $\omega_0 = \mathbf{1}_n$ in the multiplicative case, where $\mathbf{0}_n$ is the n-vector of zeros. A perturbation of the response variables $\mathbf{V} = (V_1,\ldots,V_n)^\top$ is introduced by replacing V_i by $V_i(\omega)$, such that $V_i(\omega) = V_i + \omega_i S_{\mathbf{V}}$ with additive perturbation and $V_i(\omega) = V_i\omega_i^{-1}S_{\mathbf{V}}^{-1}$ with multiplicative perturbation, where $S_{\mathbf{V}}$ is the standard deviation of $\mathbf{V}$. The perturbed Q-function is as in the Equation (2.6), switching V_i with $V_i(\omega)$, i.e., we can write the perturbed model as

$$\begin{cases} Y_i(\omega) \leq V_i, & if\ \rho_i = 1, \\ Y_i(\omega) = V_i, & if\ \rho_i = 0, \end{cases}$$

where $Y_i(\omega) = Y_i - \omega_i S_{\mathbf{V}}$ with additive perturbation and $Y_i(\omega) = Y_i\omega_i S_{\mathbf{V}}$ with multiplicative

perturbation. For these perturbation schemes, we find

$$\begin{aligned}
\Delta_{i\beta} &= -\frac{1}{\widehat{\sigma^2}} S_y \mathbf{x}_i (\widehat{u}_i + \widehat{\lambda}^2), \\
\Delta_{i\sigma^2} &= \frac{1}{\widehat{\sigma^2}^2} S_y [-\widehat{uy}_i + \widehat{u}_i S_y + \widehat{\mu}_i \widehat{u}_i + \widehat{\lambda}\widehat{t}_i - \widehat{\lambda}^2 (\widehat{y}_i - S_y - \widehat{\mu}_i)], \\
\Delta_{i\lambda} &= -\frac{1}{\widehat{\sigma^2}} S_y [\widehat{t}_i - 2\widehat{\lambda} (\widehat{y}_i - S_y - \widehat{\mu}_i)],
\end{aligned}$$

with additive perturbation and

$$\begin{aligned}
\Delta_{i\beta} &= \frac{1}{\widehat{\sigma^2}} \mathbf{x}_i S_y [\widehat{uy}_i + \widehat{\lambda}^2 \widehat{y}_i], \\
\Delta_{i\sigma^2} &= \frac{1}{\widehat{\sigma^2}^2} S_y [\widehat{uy^2}_i S_y - \widehat{\mu}_i \widehat{uy}_i - \widehat{\lambda}\widehat{ty}_i + \widehat{\lambda}^2 (\widehat{y^2}_i S_y - \widehat{\mu}_i \widehat{y}_i)], \\
\Delta_{i\lambda} &= -\frac{1}{\widehat{\sigma^2}} S_y [-\widehat{ty}_i + 2\widehat{\lambda} (\widehat{y^2}_i S_y - \widehat{\mu}_i \widehat{y}_i)],
\end{aligned}$$

with multiplicative perturbation.

(iii) *E*xplanatory perturbation

Here, we investigate the influence that perturbation in the specific continuous explanatory variable may produce on the parameter estimates. An additive perturbation of the explanatory variable $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is defined as $\mathbf{x}_i(\omega) = \mathbf{x}_i + \omega_i S_r \mathbf{I}_{r_0}$, $r \in 1, \ldots, p$, where S_r is the standard deviation of the explanatory variable x_r and $\mathbf{I}_{r_0}$ denotes a $p \times 1$ vector of zeros with one in the r-th position. Furthermore, a multiplicative perturbation of the explanatory variable $\mathbf{x}_i$ is defined as $\mathbf{x}_i(\omega) = \mathbf{x}_i \omega_i S_r$. In these cases, the perturbed Q-function is as in the Equation (2.6), switching $\mathbf{x}_i(\omega)$ with $\mathbf{x}_i$. It follows that the matrix $\Delta\omega_0$ has the following elements:

$$\begin{aligned}
\Delta_{i\beta} &= -\frac{1}{\widehat{\sigma^2}} S_r \mathbf{I}_{r_0} [-\widehat{uy}_i + \hat{u}_i \widehat{\mu}_i + \widehat{\lambda}\hat{t}_i - \widehat{\lambda}^2 (\widehat{y}_i - \widehat{\mu}_i)] - \\
&\quad \frac{1}{\widehat{\sigma^2}} \widehat{\beta}_r S_r \mathbf{x}_i [\widehat{u}_i + \widehat{\lambda}^2],
\end{aligned}$$

$$\begin{aligned}
\Delta_{i\sigma^2} &= \frac{1}{\widehat{\sigma^2}^2} \widehat{\beta}_r S_r [-\widehat{uy}_i + \widehat{u}_i \widehat{\mu}_i + \widehat{\lambda}\widehat{t}_i - \widehat{\lambda}^2 (\widehat{y}_i - \widehat{\mu}_i)], \\
\Delta_{i\lambda} &= -\frac{1}{\widehat{\sigma^2}} \widehat{\beta}_r S_r [\widehat{t}_i - 2\widehat{\lambda} (\widehat{y}_i - \widehat{\mu}_i)],
\end{aligned}$$

with additive perturbation and

$$\begin{aligned}
\Delta_{i\beta} &= \frac{1}{\widehat{\sigma^2}}\frac{x_{ir}}{S_r}\mathbf{I}_{r_0}[-\widehat{uy}_i+\widehat{\mu}_i\widehat{u}_i+\widehat{\lambda}\widehat{t}_i-\widehat{\lambda}^2(\widehat{y}_i-\widehat{\mu}_i)]+\\
&\quad \frac{1}{\widehat{\sigma^2}}\frac{\widehat{\beta}_r x_{ir}}{S_r}\mathbf{x}_i(\widehat{u}_i+\widehat{\lambda}^2),\\
\Delta_{i\sigma^2} &= -\frac{1}{\widehat{\sigma^2}^2}\frac{\widehat{\beta}_r x_{ir}}{S_r}[-\widehat{uy}_i+\widehat{\mu}_i\widehat{u}_i+\widehat{\lambda}\widehat{t}_i-\widehat{\lambda}^2(\widehat{y}_i-\widehat{\mu}_i)],\\
\Delta_{i\lambda} &= \frac{1}{\widehat{\sigma^2}}\frac{\widehat{\beta}_r x_{ir}}{S_r}[\widehat{t}_i-2\widehat{\lambda}(\widehat{y}_i-\widehat{\mu}_i)],
\end{aligned}$$

with multiplicative perturbation.

(iv) *S*cale perturbation

To investigate the effects of deviations from the assumption with respect to the scale parameter σ^2, a perturbation in this parameter can be introduced by replacing σ^2 by $\sigma^2(\omega_i)=\omega_i^{-1}\sigma^2$, for $i=1,\ldots,n$, in the Equation (2.6), getting perturbed Q-function. It is assumed that the non-perturbed model is obtained when $\omega_0=\mathbf{1}_n$. In this case, the matrix Δ_{ω_0} has the following elements:

$$\begin{aligned}
\Delta_{i\beta} &= \frac{1}{\widehat{\sigma^2}}\mathbf{x}_i[\widehat{uy}_i-\widehat{\mu}_i\widehat{u}_i-\widehat{\lambda}\widehat{t}_i+\widehat{\lambda}^2(\widehat{y}_i-\widehat{\mu}_i)],\\
\Delta_{i\sigma^2} &= -\frac{1}{\widehat{\sigma^2}^2}+\frac{1}{2\widehat{\sigma^2}^2}[\widehat{uy^2}_i-2\widehat{\mu}_i\widehat{uy}_i+\widehat{\mu}_i^2\widehat{u}_i+\widehat{t^2}_i-2\widehat{\lambda}\widehat{ty}_i+2\widehat{\lambda}\widehat{\mu}_i\widehat{t}_i+\\
&\quad \lambda^2(\widehat{y^2}_i-2\mu_i\widehat{y}_i+\mu_i^2)],\\
\Delta_{i\lambda} &= -\frac{1}{\widehat{\sigma^2}}[-\widehat{ty}_i+\widehat{\mu}_i\widehat{t}_i+\widehat{\lambda}(\widehat{y^2}_i-2\widehat{\mu}_i\widehat{y}_i+\widehat{\mu}_i^2)].
\end{aligned}$$

(v) *P*erturbation of the skewness parameter

To investigate the effects of deviations from the assumption with respect to the skewness parameter λ, a perturbation in this parameter can be introduced by replacing λ by $\lambda(\omega_i)=\omega_i^{-1}\lambda$, for $i=1,\ldots,n$, in the Equation (2.6), getting perturbed Q-function. It is supposed that the non-perturbed model is obtained when $\omega_0=\mathbf{1}_n$. In this case, the matrix Δ_{ω_0} has the following elements:

$$\Delta_{i\beta} = -\frac{1}{\widehat{\sigma^2}}\widehat{\lambda}\mathbf{x}_i[\widehat{t}_i-2\widehat{\lambda}(\widehat{y}_i-\widehat{\mu}_i)],$$

$$\begin{aligned}
\Delta_{i\sigma^2} &= \frac{1}{\widehat{\sigma^2}^2}\{\widehat{\lambda}[-\widehat{ty}_i+\widehat{\mu}_i\widehat{t}_i+\widehat{\lambda}(\widehat{y^2}_i-2\widehat{\mu}_i\widehat{y}_i+\widehat{\mu}_i^2)]\},\\
\Delta_{i\lambda} &= -\frac{1}{\widehat{\sigma^2}}[-\widehat{ty}_i+\widehat{\mu}_i\widehat{t}_i+2\widehat{\lambda}(\widehat{y^2}_i-2\widehat{\mu}_i\widehat{y}_i+\mu_i^2)].
\end{aligned}$$

In the next sections, a simulation study and an application to real data are presented in order to illustrate the performance of the developed methodology.

3.2 Performance of the Proposed Diagnostic Measures

The performance of the local influence diagnostic measure in detecting influential observations is explored through the analysis of empirical studies. Inspired by Schumacher *et al.* (2018), we generate 500 Monte Carlo samples of sizes $n = 100$ using a left-censored SSMN-CR model and level of censoring 15%, with parameters set at $\beta = (1,-1,2,-2)^\top, \sigma^2 = 1, \lambda = -3, \tau = 5$ for the ST-CR and SSL-CR models and $\tau = (0.3, 0.1)^\top$ for the SCN-CR model; and the explanatory variables $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^\top$, where $x_{ij} \sim N(0,1)$ for $i = 1,\ldots,n$ and $j = 1,2,3$. For each sample, We generate an atypical point in the following way: replacing the $50th$ response observation by $max(y) + k \times sd(y), k = 0.1, 0.3, 1$ and 3, where $max(\cdot)$ and $sd(\cdot)$ is the maximum and standard deviation of the variable, respectively.

The results are presented in Tables 1-4. The proportion of correct identifications increases as the magnitude of the outlier response value increases under the GD measure, case weight, scale, response, and asymmetry parameter perturbation in the SCN-CR model (see Table 1). Note that in the SSL-CR model this behavior occurs under the GD measure, response and perturbation of the asymmetry parameter (see Table 2). In the ST-CR model, we observed this pattern under the GD measure and response perturbation (see Table 3); and only in the GD measure for the SN-CR model (see Table 4). In general, the proportion of correct identifications in the SSMN-CR model varies between 0.75 and 1 with exceptions under the case weight and scale perturbations for the ST-CR model; and under the response perturbation for the ST-CR and SCN-CR models. In the SCN-CR model, it can be seen that the proportion is around 0.8 with $k = 3$ under the response perturbation. In the ST-CR model, note that in the case of weight and scale perturbations, with $k = 0.1$, the proportions are around 0.7 and tend to decrease to approximately 0.2, but under the response perturbation the behavior is reversed, as mentioned here previously. In this way, we can conclude that the proposed diagnostic methods exhibit reasonable performance in detecting influential observations in the SSMN-CR model.

Table 1 – Proportion of times that the observation 50 was identified as influential for each type of perturbation, including GD measure, under the SCN-CR model.

k	Case weight	Scale	Skewness	Response (“m”)	GD(θ)
0.10	0.84	0.84	0.77	0.08	0.95
0.50	0.84	0.84	0.80	0.12	0.96
1.00	0.85	0.85	0.84	0.22	1.00
3.00	0.86	0.86	0.86	0.83	1.00

Note that the results presented for this simulation cover only a few of the possible scenarios. This is due to the extensive range of disturbances tested in the local analysis, as well

Table 2 – Proportion of times that the observation 50 was identified as influential for each type of perturbation, including GD measure, under the SSL-CR model.

k	Case weight	Scale	Skewness	Response ("m")	GD(θ)
0.10	0.80	0.83	0.83	0.78	0.97
0.50	0.82	0.85	0.85	0.84	0.99
1.00	0.79	0.83	0.87	0.87	1.00
3.00	0.75	0.81	0.87	0.87	1.00

Table 3 – Proportion of times that the observation 50 was identified as influential for each type of perturbation, including GD measure, under the ST-CR model.

k	Case weight	Scale	Skewness	Response ("m")	GD(θ)
0.10	0.66	0.69	0.86	0.21	0.94
0.50	0.59	0.61	0.83	0.22	0.97
1.00	0.52	0.55	0.85	0.33	0.97
3.00	0.16	0.18	0.82	0.68	0.97

Table 4 – Proportion of times that the observation 50 was identified as influential for each type of perturbation, including GD measure, under the SN-CR model.

k	Case weight	Scale	Skewness	Response ("m")	GD(θ)
0.10	0.84	0.84	0.85	0.86	0.98
0.50	0.83	0.83	0.84	0.85	0.99
1.00	0.81	0.81	0.84	0.84	1.00
3.00	0.84	0.85	0.87	0.87	1.00

as variations in parameter intensities, such as different sample sizes and numbers of iterations. For example, additional anomalies could be explored by testing different values for the model parameters in the simulation. However, this experimental part was not feasible due to the potential overlap of computational time with the final submission date of this thesis.

3.3 Application

In this section, we present the application of diagnostic analysis for the SSMN-CR models to the Stellar abundance dataset. This dataset was taken from the work of Santos *et al.* (2002) and has been previously analyzed by Mattos, Garay and Lachos (2018) under the SMSN family of distributions. Recently, Guzman, Ferreira and Zeller (2020) analyzed this same dataset and pointed out that they are better suited for SSMN distributions with heavier tails than SMSN distributions. Then, we revisited this dataset in order to carry out diagnoses of global and local influence based on the approach of Zhu and Lee (2001) under the SSMN-CR model.

This data contains measurements for 68 solar-type stars, where the $\log N(Be)$ is the response variable, which represents the log of the abundance of the light element beryllium (Be) in stars scaled to the Sun's abundance (i.e. the Sun has $\log N(Be) = 0.0$) and the $Teff/1000$ is

the explanatory variable, which represents the effective stellar surface temperature (in kelvin). Moreover, we have 12 left-censored data points, i.e. 12 undetected beryllium measurements, which represents 19.35% of observations.

This illustrative application is organized as follows. First, we show the SSMN-CR models fitted to the stellar abundance dataset. We then compare the SSMN-CR models by examining various information selection criteria. Next, we identify influential observations in the stellar abundance dataset using the influence diagnostics analysis described in Section 3.1. The adjustment results, including log-likelihood, AIC, and BIC, are provided in Table 5. Both the AIC and BIC criteria favor the ST-CR model, which is characterized by heavy tails and asymmetric behavior. For more details on these selection criteria, see Akaike (1998), Schwarz (1978) and Gelman, Hwang and Vehtari (2014).

Table 5 – Stellar abundances dataset: Comparison of log-likelihood maximum, AIC and BIC for fitted various models using the stellar abundances data. Best fit indicated by (*1).

SSMN-CR models	log-likelihood	AIC	BIC
SN-CR	-18.2276	44.4553	53.3333
ST-CR	-1.7802	11.5605 (*1)	20.4385 (*1)
SSL-CR	-3.2474	14.4949	23.3729
SCN-CR	-4.4375	16.8750	25.7531

Next, we present the results of the diagnostic analysis for the ST-CR model, which best fits the data, and for the typical models in the context of SSMN distributions.

3.3.1 Detection of outliers

First, the censored values are imputed and then, the Mahalanobis distance $d_i = (y_i - \mu_i)/\sigma^2, i = 1,\ldots,68$, has been employed to detect extreme observations, using a complete dataset. See Section 4.3 in Guzman, Ferreira and Zeller (2020) for more details about the imputation of censored observations. According to Lange and Sinsheimer (1993) and Ferreira, Bolfarine and Lachos (2011), we have that in the SN case, the Mahalanobis distance $d_i \sim \chi^2_1$, thus one can use the quantile $\vartheta = \chi^2_1(\xi)$, where $0 < \xi < 1$, to identify outliers. For the ST distribution, we can use $d_i \sim F(\nu,1)$; for the SSL case $Pr(d_i \leq \vartheta) = Pr(\chi^2_1 \leq \vartheta) - \frac{2^\nu \Gamma(\nu+1/2)}{\vartheta^\nu \Gamma(1/2)} Pr(\chi^2_{2\nu+1} \leq \vartheta)$ and finally for the SCN model $Pr(d_i \leq \vartheta) = \nu Pr(\chi^2_1 \leq \gamma\vartheta) + (1-\nu)Pr(\chi^2_1 \leq \vartheta)$.

In the Figure 1, we report the index plot of the Mahalanobis distances for SSMN-CR models, where the cutoff lines correspond to the quantile ϑ, with $\xi = 0.95$. We can see from these figures that observations 13, 23, 29 and 49 appear as possible outliers for the ST, SSL and SCN models.

As we can see the analysis by the Q-function, given in the Equation (2.6), depends on the Mahalanobis distance d_i. Note that the values of weights $\widehat{u}_i$ in the Q-function depend on d_i and seem be inversely proportional to d_i. For example, in the ST model, $\widehat{u}_i = (\tau+1)/(\tau+\widehat{d}_i)$, with

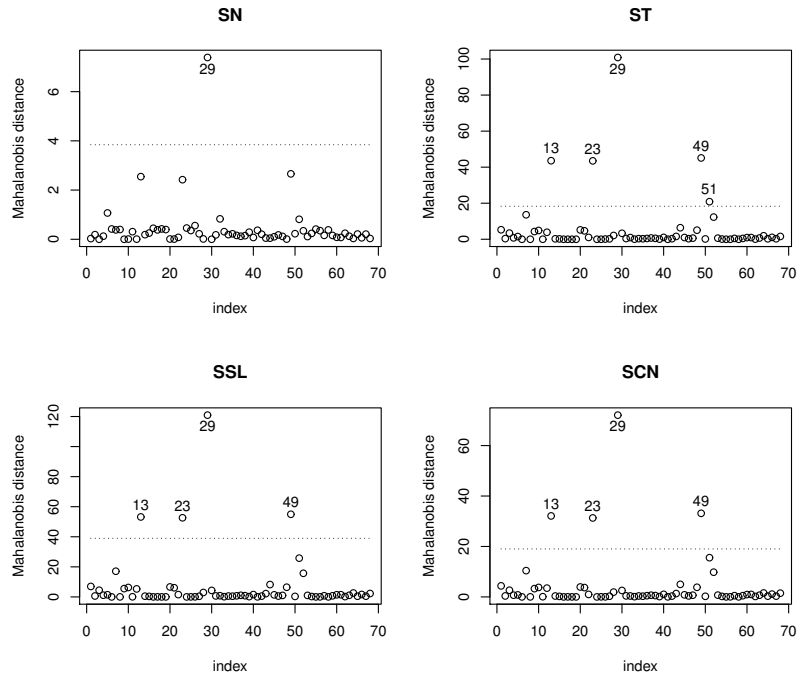


Figure 1 – Index plots of the $\widehat{d_i}$ for the SSMN-CR fitted models.

$\widehat{d_i} = (y_i - \widehat{\mu}_i)/\widehat{\sigma^2}$. In MCEM estimation procedure, the M-Step depends on $\widehat{u}_i$, which depends on the Mahanalobis distance d_i. Then, this distance can be used successfully to detect anomalous observations, candidates influential observations, in SSMN-CR models because for big values of d_i, we have small values of $\widehat{u}_i$.

The terms outliers and robustness play an important role in this work. One of the objectives underlying the estimation and diagnostic techniques considered in this work is the development of procedures under the class of SSMN distributions that are robust in the presence of outliers, that is, statistical methods that are less affected by extreme observations. In the Figure 2, the observations that stand out in Figure 1 show the lowest weights for the ST, SSL and SCN models. For the skew-normal case, $u_i = 1, \forall i$, and are shown with segmented lines in Figure 2. The results indicate that the SSMN-CR model, based on heavy-tailed distributions, can better accommodate atypical observations by assigning smaller weights during the estimation process.

A fact to be highlighted is that even robust parameter estimation models (skewed and heavy-tailed) can present unusual observations such as outliers or influential observations. Thus, diagnostics methods are still important tools for detecting anomalies in the fitted model.

3.3.2 Influence diagnostic analysis

3.3.2.1 Global influence

Here, we discuss the effects of influential observations on parameter estimates through individual exclusion, joint exclusion and conditional exclusion. First, Figure 3(a) displays the

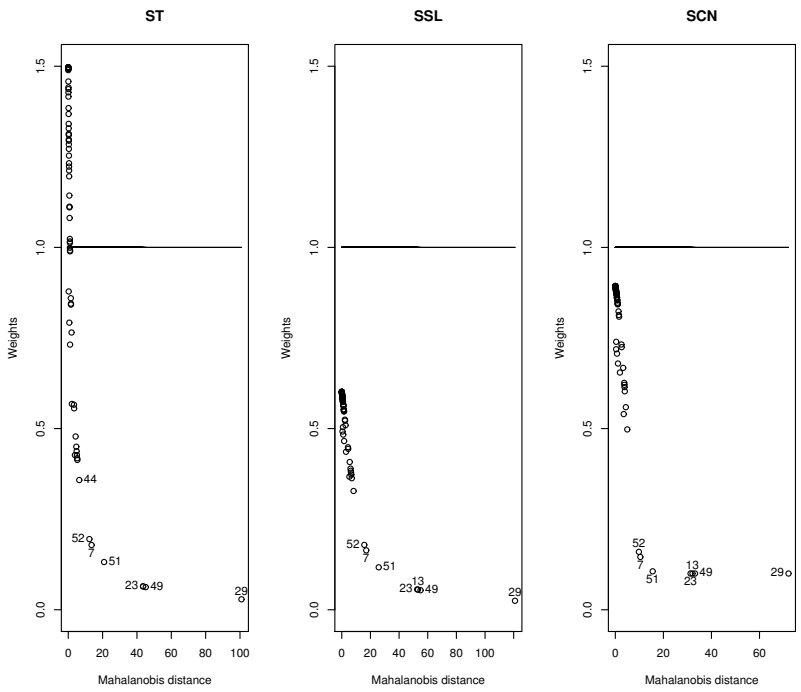
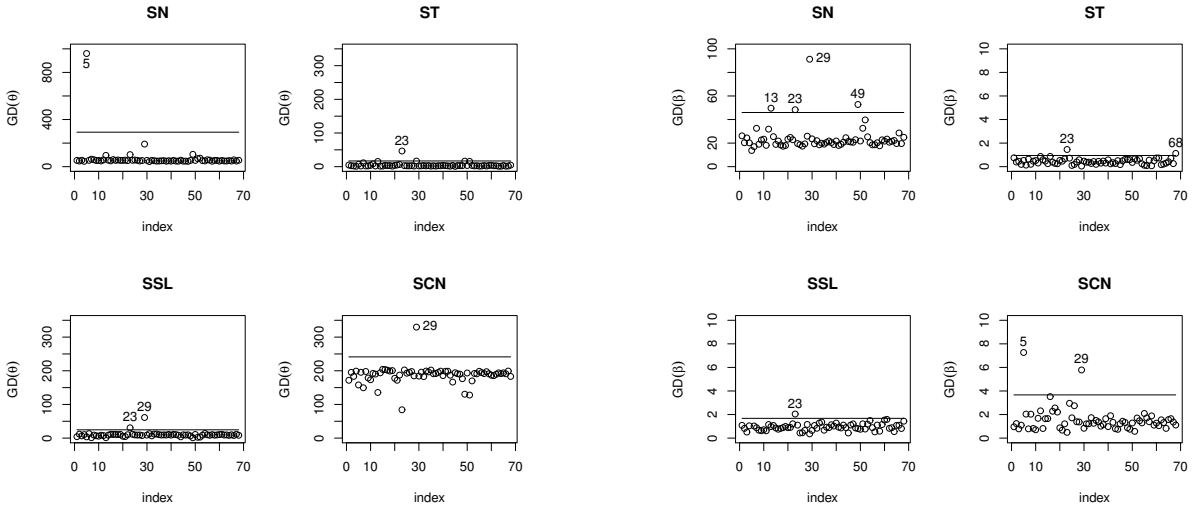


Figure 2 – Weights ($\widehat{u}_i$) for the SSMN-CR fitted models.

measures $GD_{[i]}(\theta), i = 1,\ldots,68$, for the SSMN-CR fitted models. In Figure 3(a), we can observe that observations 5 under the SN-CR model, 23 under the ST-CR model; 29 under the SCN-CR model and 23 and 29 under the SSL-CR model were identified as potentially influential on the parameter estimates. To assess the influence of these observations on the ML estimate of the components of θ, i.e., β, σ^2 and λ, we analyze the $GD_{[i]}(\beta), GD_{[i]}(\sigma^2)$ and $GD_{[i]}(\lambda)$ plots in Figures 3(b), 4(a) and 4(b), respectively. From Figures 4(a) and 4(b), we can see that observation 5 is influential regarding the parameters σ^2 and λ (in particular). We can still see that observation 5 stands out as influential under the SCN-CR model concerning the parameter β. In Figures 4(a) and 4(b), observation 29 is influential on the parameter estimates σ^2 and λ (in particular) under the SCN-CR model. This observation was also identified as potentially influential on parameter estimates λ under the SSL-CR model. Moreover, under the SSL-CR model, observation 23 is influential concerning the parameters β and σ^2. Finally, in Figures 3(b) and 4(b), note that observation 23 is seen to be influential under the ST-CR model regarding the parameters β and λ (in particular). In general, the results show that the exclusion of some observations mainly affected the estimates of λ for the four fitted models. The effects of such influential observations on parameter estimates is reduced when considering models with heavier tails than those of the SN-CR model, for instance, the ST-CR model.

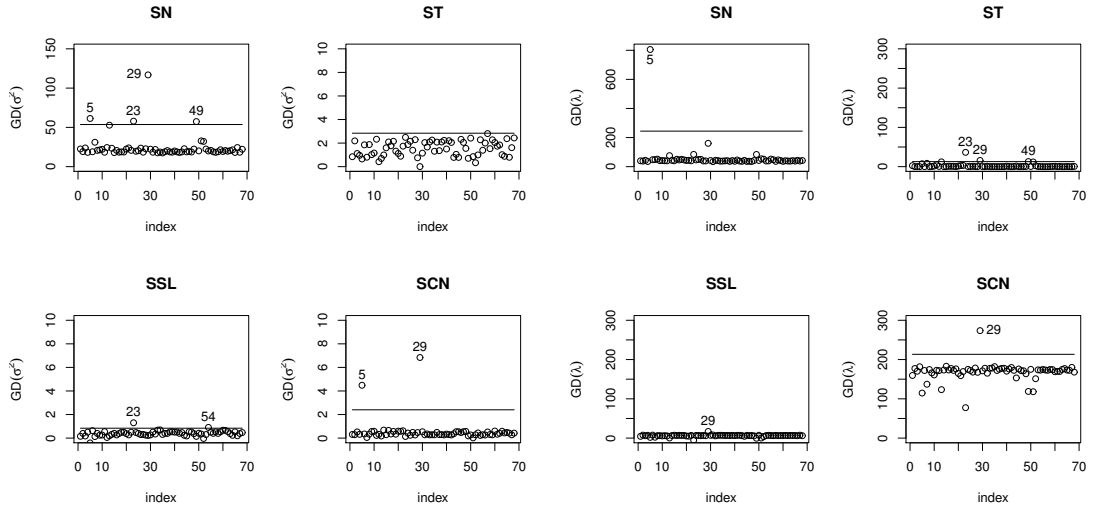
Since subjects $5, 23$ and 29 are the most influential, we will evaluate the joint influence and the conditional influence based on these observations; see Figures 5-7.

In Figure 5, note that for the SSMN-CR model, based on heavy-tailed distributions, the values $GD_{[5,i]}$ are closer to the values $GD_{[i]}, i \neq 5$ and $i = 1,\ldots,68$, indicating that the other



(a) Index plot of $GD_{[i]}(\theta)$. (b) Index plot of $GD_{[i]}(\beta)$.

Figure 3 – Index plots for the SSMN-CR fitted models.



(a) (b)

Figure 4 – Index plots of (a) $GD_{[i]}(\sigma^2)$ and (b) $GD_{[i]}(\lambda)$ for the SSMN-CR fitted models.

observations suffer less from the effect of observation 5, different from what occurs in the SN-CR model. In the SN-CR and ST-CR models, in general, the values $GD_{[5,i]} > GD_{[i]}, \forall i \neq 5$. It shows an enhancing effect by subject 5 with other subjects while both estimating θ. However, $GD_{[5,29]} < GD_{[29]}$ for the ST-CR, SSL-CR, and SCN-CR models, and only under the SCN-CR model, we have that $GD_{[5,23]} < GD_{[23]}$. Thus, these highlighted observations, in the heavy-tailed models, have their effect attenuated by observation 5 when estimating θ. In addition, in this scenario, we can see that cases 23 under the ST-CR model, 23 and 29 under the SSL-CR model, and 29 under the SCN-CR model have the highest joint Cook's distance. The conditional Cook's distances are smaller than the individual Cook's distances for most observations, except for

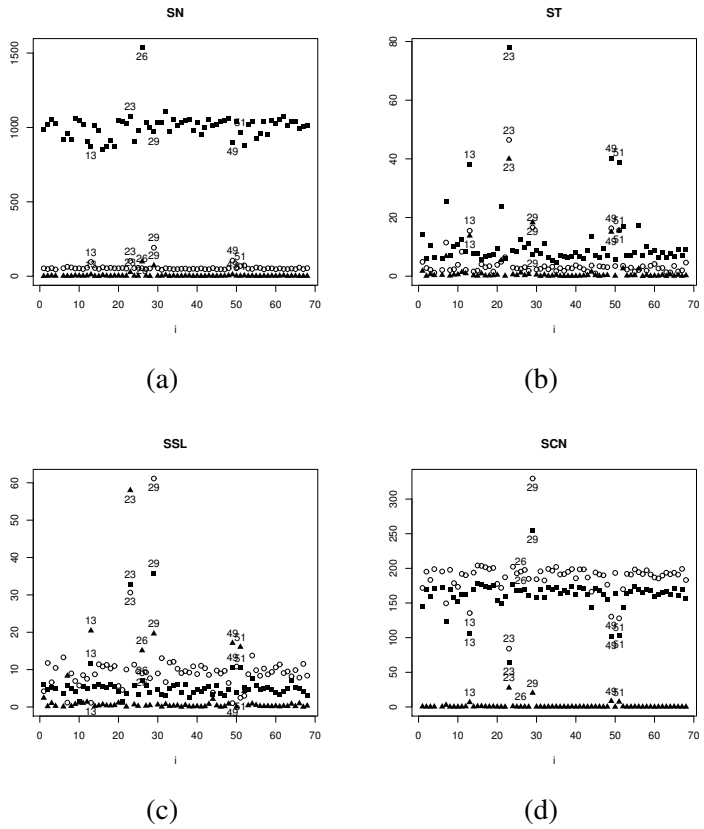
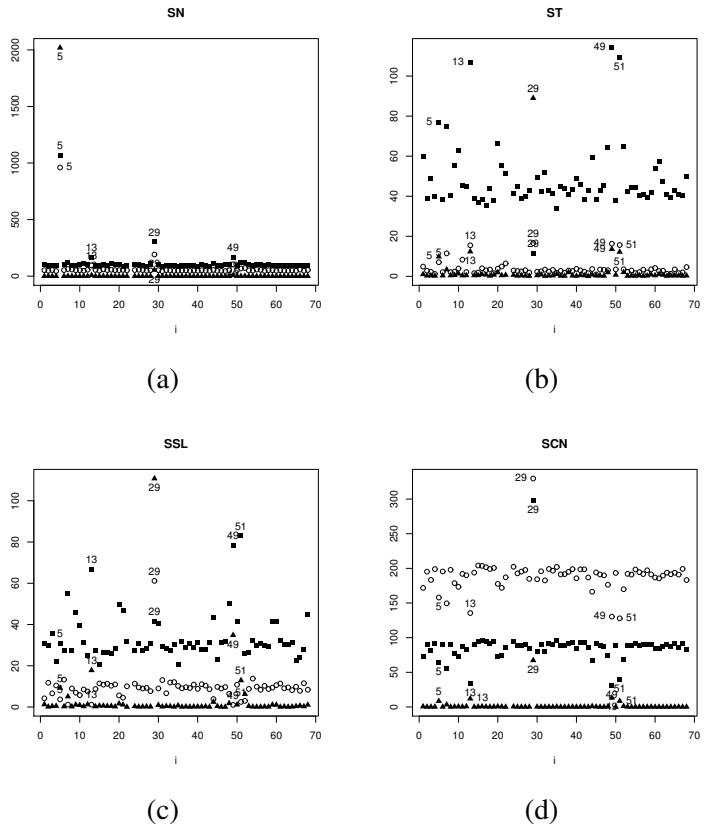


Figure 5 – Stellar abundance dataset - dots, squares and triangles denote $GD_{[i]}$, $GD_{[5,i]}$ and $GD_{[i|5]}$, respectively.

$i = 23$ and 29 in the SSL and ST models, respectively. Therefore, cases 23 and 29 are said to have a masking effect by observation 5 when θ is estimated in the respective models mentioned above.

From Figure 6, note that under the SN-CR, ST-CR and SSL-CR models, $GD_{[23,i]} > GD_{[i]}, \forall i \neq 23$, except for $i = 29$ in the ST and SSL models. Unlike what happens in the other models, it is observed that in the SCN-CR model $GD_{[23,i]} < GD_{[i]}, \forall i \neq 23$. Then, under the SCN-CR model, observation 23 is said to have a reducing effect relative to observation i when θ is estimated. Furthermore, in this context, we can see that cases 29 and 5 have the highest joint Cook distance under the SCN-CR and SN-CR models, respectively. The Cook's distances show that only $GD_{[5|23]} > GD_{[5]}$ under the SN-CR model, $GD_{[i|23]} > GD_{[i]}$ for $i = 5$ and 29 under the ST and SSL models. Then, observations 5 and 29 are said to have a masking effect by observation 23 when θ is estimated in the respective models mentioned above.

From Figure 7, the values that $GD_{[i|29]} < GD_{[i]}$, $\forall i \neq 29$, show that the influence of subjects i decreases after the deletion of subject 29. It suggests that subject i has been boosted by subject 29 while estimating the θ, except for $i = 5$ in the SN, ST and SSL models. Note that for all fitted models, $GD_{[29,i]} > GD_{[i]}$, $\forall i \neq 29$, except for $i = 23$ in the ST-CR model. Thus, observation 23 in the ST-CR model has its effect attenuated by observation 29 when estimating θ.



(a) (b)

(c) (d)

Figure 6 – Stellar abundance dataset - dots, squares and triangles denote $GD_{[i]}$, $GD_{[23,i]}$ and $GD_{[i|23]}$, respectively.

In addition, in this scenario, we can see that case 5 has the highest joint and conditional Cook's distances under the SN-CR model.

These findings have important implications for further inference; see Table 7.

3.3.2.2 Local influence

We now identify influential cases for the dataset using $M(0)$. Figures 8-11 display the index graphs of $M(0)$ for the proposed perturbation schemes. Table 6, summarizes the observations detected as influential under different perturbation schemes. This table lists the influential points identified according to various criteria in the local analysis. Given the large number of influential points, we focus on a subset known as the "consistent set of influential points." This subset includes points consistently identified as influential across all analysis methods. In this case, observations $5, 23$, and 29 are highlighted as consistently influential. Below are some comments on the perturbation schemes considered in this work.

*C*ase-weight perturbation and *S*cale perturbation: Figures 8 and 11(a) show that under the SSMN-CR models, observation 5 is identified as influential under the case-weights perturbation and scale perturbation. Furthermore, under the SCN-CR model, observation 29 (also detected as an outlier) is identified as influential under both perturbations.

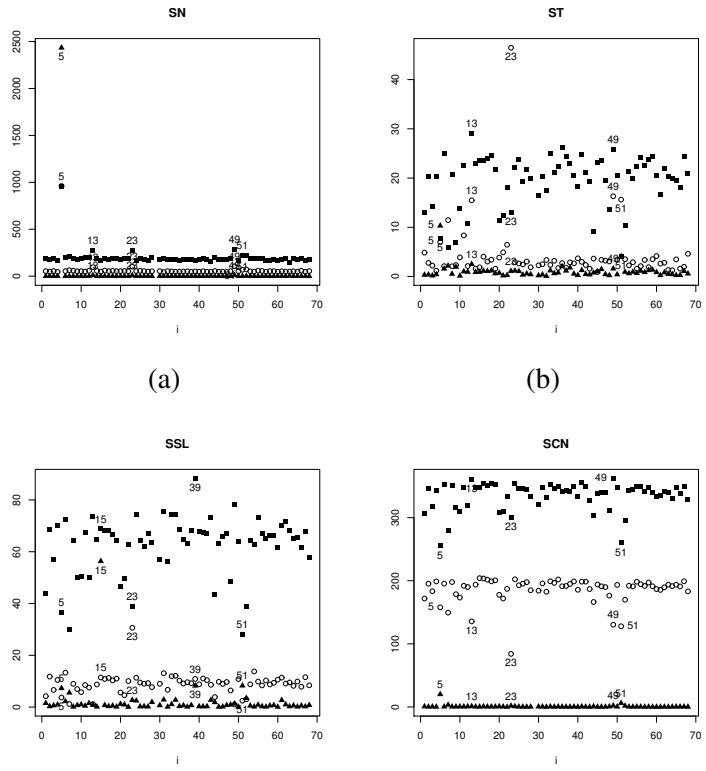


Figure 7 – Stellar abundance dataset - dots, squares and triangles denote $GD_{[i]}$, $GD_{[29,i]}$ and $GD_{[i|29]}$, respectively.

Table 6 – Influential observations under SSMN-CR models.

Perturbation schemes	**Fitted models**			
	SN-CR	ST-CR	SSL-CR	SCN-CR
Case weight perturbation	5-11-62	5	5	5-29
Response perturbation (“a")	13-23-29-49	29	29	29
Response perturbation (“m")	16	16-24	16-24	16-24
Explanatory perturbation (“a")	5-23-29	5-29	5-23-29	5-23-29
Explanatory perturbation (“m")	5-23-29	5-23-29	5-23-29	5-23-29
Scale perturbation	5-11	5	5	5-29
Perturbation of the skewness parameter	29	23-29	23-29	29

*R*esponse perturbation: We now examine the effects of perturbing the response variable by an additive perturbation scheme. Figure 9(a) indicates some influence when the response of item 29 is perturbed under the SSMN-CR models. Furthermore, under the SN-CR model, observation 23 (also detected as an outlier) is identified as influential. These are censored observations and stand out for having small values in the response variable, that is, $y_{23} = 0.25$ (below the first quartile) and $y_{29} = -0.4$ (minimum value of the respective variable). Using this perturbation, we can examine the influence on the response variable under the multiplicative perturbation scheme.

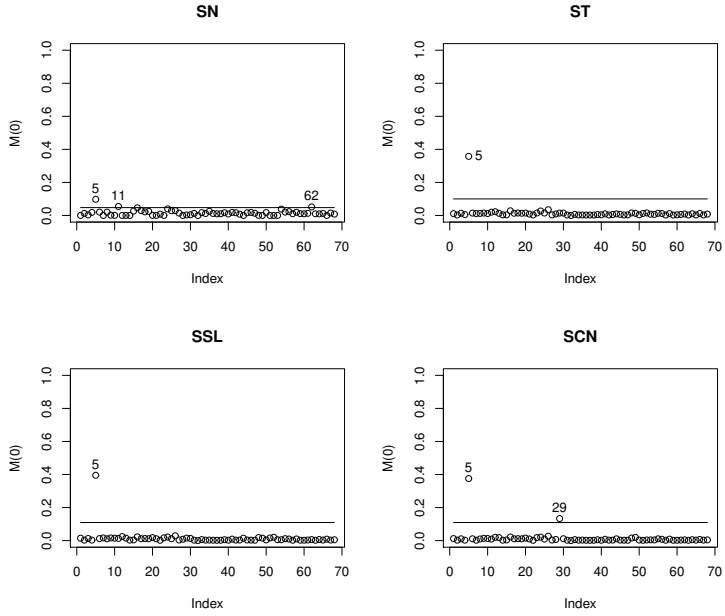
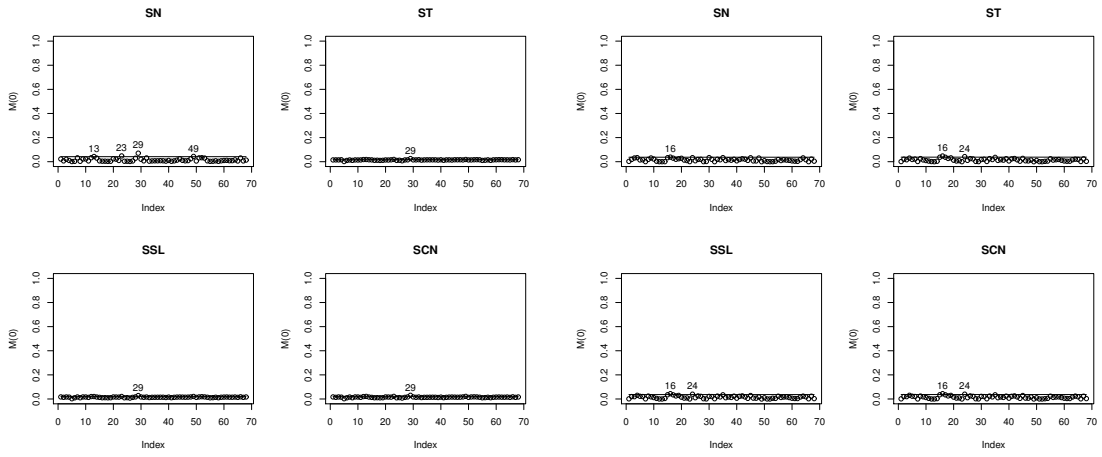
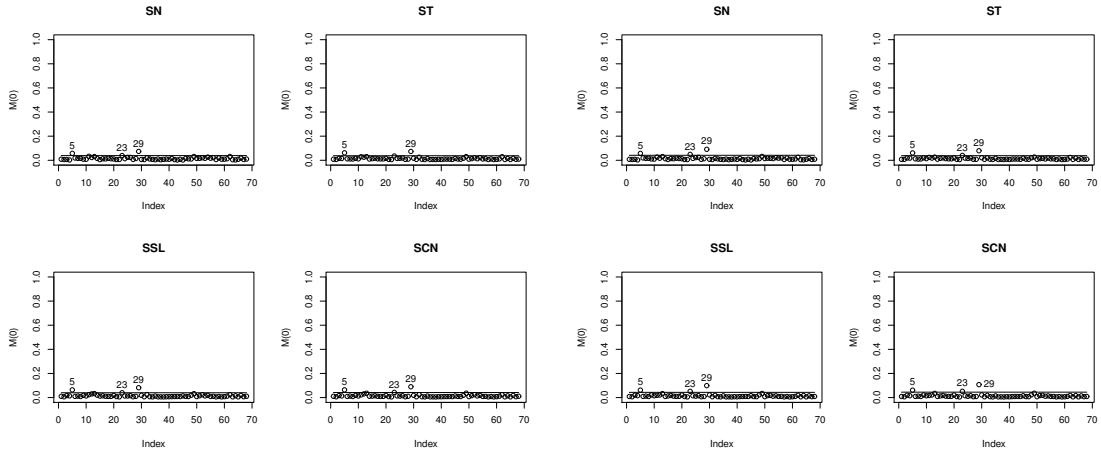


Figure 8 – Index plots of $M(0)$ under case weight perturbation for the SSMN-CR fitted models.



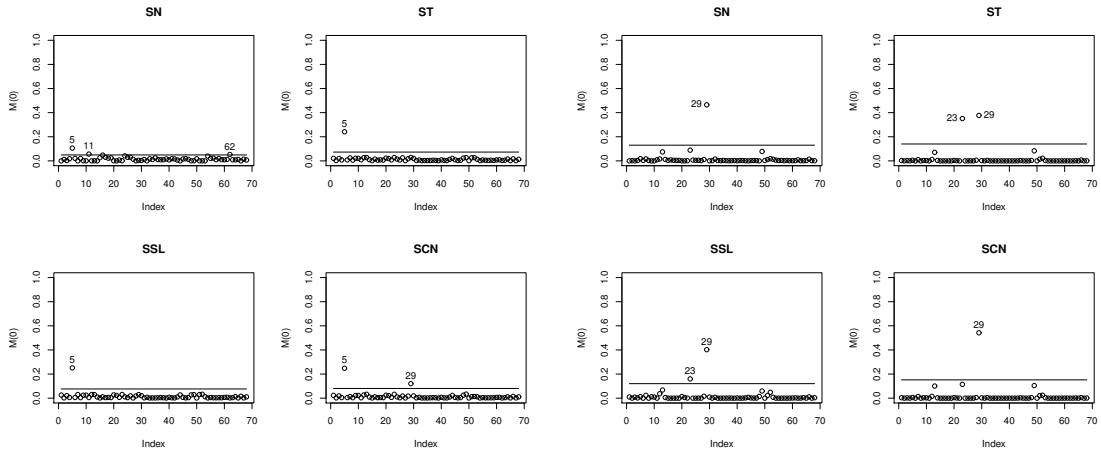
(a) Additive perturbation (b) Multiplicative perturbation

Figure 9 – Index plots of $M(0)$ under response perturbation for the SSMN-CR fitted models.



(a) Additive perturbation (b) Multiplicative perturbation

Figure 10 – Index plots of $M(0)$ under explanatory perturbation for the SSMN-CR fitted models.



(a) Scale perturbation (b) Perturbation of the skewness parameter

Figure 11 – Index plots of $M(0)$ under scale and skewness perturbations for the SSMN-CR fitted models.

In Figure 9(b), we see some influences when the responses of items 16 and 24 are perturbed under the ST-CR, the SSL-CR and SCN-CR models, while only observation 16 is identified as influential under the SN-CR model. Such observations are notable for having large values in the response variable, that is, $y_{16} = 1.36$ (maximum value of the respective variable) and $y_{24} = 1.33$ (value above the third quartile).

*E*xplanatory perturbation: Figure 10 shows that in the SSMN-CR model, cases $5, 23$ and 29 are identified as influential under the explanatory perturbation in additive and multiplicative perturbation schemes. These observations stand out because they have the following values in x: $x_5 = 5.641, x_{23} = 6.339$ and $x_{29} = 6.229$. We observe that the value of x_{23} corresponds to the maximum temperature, and the values x_{29} and x_5 are, respectively, above the third quartile and

Table 7 – Comparison of the RC% in the $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\sigma^2}$ and $\widehat{\lambda}$ for the SSMN-CR fitted models.

	SN-CR				ST-CR			
M	RC($\widehat{\beta}_0$)	RC($\widehat{\beta}_1$)	RC($\widehat{\sigma^2}$)	RC($\widehat{\lambda}$)	RC($\widehat{\beta}_0$)	RC($\widehat{\beta}_1$)	RC($\widehat{\sigma^2}$)	RC($\widehat{\lambda}$)
[#5]	1.071%	0.284%	27.833%	61.9308%	1.958%	0.953%	12.133%	7.313%
[#29]	2.111%	4.345%	38.459%	26.290%	1.359%	0.8702%	7.724%	11.005%
[#23]	0.302%	2.253%	26.589%	19.166%	1.058%	0.198%	22.721%	15.307%
[#5,#29]	7.469%	7.998%	47.795%	58.662%	3.424%	1.695%	22.222%	2.765%
[#23,#29]	11.078%	11.517%	47.399%	31.694%	1.326%	0.910%	11.982%	8.794%
[#5,#23,#29]	13.158%	12.947%	56.706%	57.445%	5.096%	2.820%	23.341%	0.804%
	SNC-CR				SSL-CR			
M	RC($\widehat{\beta}_0$)	RC($\widehat{\beta}_1$)	RC($\widehat{\sigma^2}$)	RC($\widehat{\lambda}$)	RC($\widehat{\beta}_0$)	RC($\widehat{\beta}_1$)	RC($\widehat{\sigma^2}$)	RC($\widehat{\lambda}$)
[#5]	2.510%	0.628%	26.248%	23.522%	2.274%	1.016%	15.478%	0.996%
[#29]	2.196%	0.5401%	30.390%	36.472%	0.2304%	0.012%	6.778%	16.885%
[#23]	2.477%	1.270%	4.148%	19.262%	0.253%	0.679%	40.151%	5.691%
[#5,#29]	11.530%	5.846%	52.743%	28.986%	3.582%	1.754%	18.072%	10.703%
[#23,#29]	6.093%	2.887%	36.537%	34.137%	4.0206%	2.551%	0.403%	14.125%
[#5,#23,#29]	17.243%	9.331%	60.836%	25.875%	7.107%	3.945%	20.582%	6.285%

between the first quartile and the median of the variable of interest.

*P*erturbation of the skewness parameter: In Figure 11(b) it can be seen that observation 29 is identified as influential only under the SN-CR and SCN-CR models when compared with the ST-CR and SSL-CR models. Note that under the ST-CR and SSL-CR models, observation 23 is identified as influential too.

Furthermore, note that the ML estimates are quite stable as they relate to the perturbations of the response and explanatory variables in the four models considered, as shown in Figures 9 and 10. The table 6 displays the observations detected through the different perturbation schemes of the applied local influence analysis. Note that observations $5, 23$ and 29 are detected as potentially influential. To quantifier the impact of these observations on the ML estimates, we readjusted the model, discarding each one. Consequently, in Table 7 we present a comparison of the RC% (relative changes in %) in the $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\sigma^2}$ and $\widehat{\lambda}$ for the SSMN-CR fitted model. Let $RC(\widehat{\alpha}) = \left|\frac{(\widehat{\alpha} - \widehat{\alpha}_{[M]})}{\widehat{\alpha}}\right|$, where $\alpha = \beta_0, \beta_1, \sigma^2$ or λ and $\widehat{\alpha}_{[M]}$ are the estimates of α obtained using the data without the observations in M. We found that the observations $5, 23$ and 29, detected as influential in both the global and local influence diagnotics, cause a significant change in the parameters σ^2 and λ in the four fitted models. Furthermore, removing these observations together affects the estimates of all parameters of the models, in particular for the parameters σ^2 and/or λ in the four fitted models. Note that the biggest changes occur in the SN-CR model. Thus, our main conclusion for this data set is that the maximum-likelihood estimates from the ST-CR model (best fit) seem to be more robust than the estimates from the SN-CR model.

Additionally, we carry out a series of simulations based on actual data to explore the effectiveness of the methods given in the chapter.

3.3.3 Effectivenes of the proposed diagnostic measures

This simulation study considers the generation of 100 resamples with replacement of $n = 68$ of the dataset under investigation (non-parametric bootstrap). Global and local diagnostic measures were calculated in each sample in order to evaluate the effectiveness of the proposed measures in the context of the SSMN-CR model. It is worth mentioning that, due to the characteristics of the data set, the ST model emerged as the most appropriate. Consequently, this simulation study was conducted for the SN-CR and ST-CR models to provide a comprehensive comparison. After resampling, we observed that the average censoring proportion is 18.22% and the median censoring proportion is 19.12%, indicating that, in our study, we maintained the censoring level around the censoring proportion of the original dataset.

Following the approach described in Section 3.1, we counted the number of times the observations highlighted in Figure 1 and Table 6 were identified as influential and the average number of influential observations identified, for all samples. The results are presented in Tables 8-9. In general, the capability of our proposal for detecting influential points seems to be reasonable. Note that observations 5, 23 and 29 are detected as potentially influential, as expected.

Table 8 – Summary of number of detected influential observations for all Bootstrap samples for each type of perturbation, including Mahalanobis distance and GD measure, under the SN-CR and ST-CR models. SD - Standard Deviation.

Measures	**SN-CR**		**ST-CR**	
	Mean	SD	Mean	SD
Mahalanobis distance	2.03	1.09	4.92	1.81
Case weight perturbation	3.21	1.34	2.28	1.21
Scale perturbation	3.20	1.34	1.99	1.09
Perturbation of the skewness parameter	2.32	1.42	2.38	1.44
Explanatory perturbation (“a”)	3.03	1.25	3.39	1.20
Explanatory perturbation (“m”)	3.10	1.18	3.29	1.18
Response perturbation (“a”)	2.75	1.13	1.33	0.99
Response perturbation (“m”)	0.74	1.07	2.64	1.21
GD(θ)	1.71	1.09	3.53	1.21

3.3.4 Influence of a single outlier

Finally, aspects of the robustness of the SSMN-CR model can be illustrated by disturbing an observation in the data. Specifically, we compared the ST-CR model (the one that best fits the data for this application) with the SN-CR model. Changes in ML estimates of θ can be evaluated replacing v_k, from the data, by $v_k(\delta) = v_k + \delta$, for δ between 0 and 5 in increments of 0.5. In other words, we create an outlier, first, contaminating the typical observation 4 ($y_4 = 1.19$). After, we contaminate observation 5, detected as influential for δ between 0 and 15 in increments of 1. The influence of outliers on estimates can be assessed based on the mean magnitude of relative

Table 9 – Frequency (in parentheses) of influential observations for all Bootstrap samples for each type of perturbation, including Mahalanobis distance and GD measure, under the SN-CR and ST-CR models.

Measures	SN-CR	ST-CR
Mahalanobis distance	29(121)	29(125)-13(102)-23(95)-49(101)-51(57)
Case weight perturbation	5(82)-11(11)-62(16)	5(82)
Response perturbation (“a”)	13(49)-23(43)-29(125)-49(53)	29(85)
Response perturbation (“m”)	16(42)	16(78)-24(81)
Explanatory perturbation (“a”)	5(82)-23(20)-29(68)	5(80)-29(109)
Explanatory perturbation (“m”)	5(82)-23(35)-29(86)	5(80)-23(34)-29(121)
Scale perturbation	5(82)-11(11)	5(82)
Perturbation of the skewness parameter	29(125)	23(33)-29(119)
GD(θ)	5(63)-29(39)	23(71)-29(58)

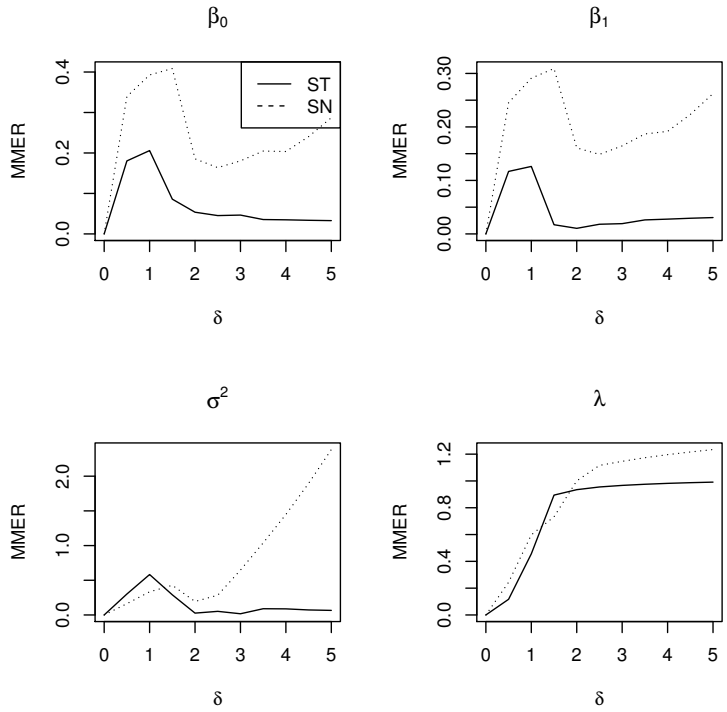


Figure 12 – Stellar abundances dataset - contamination of observation 4. Mean magnitude of relative error (MMER) of EM estimates for $\beta_0, \beta_1, \sigma^2$ and λ.

error (MMER) defined by Guzman, Ferreira and Zeller (2020) (see Section 4.4). In Figures 12 and 13, we present the results of the MMERs for different contaminations δ. As expected, note that, under the SN-CR model, the outlying observations have much more impact on the ML estimates of θ. This suggests that the ST-CR model provided an appropriate way for achieving robust statistical inference.

3.4 Conclusions

The SSMN-CR models provide a satisfactory fit to the data illustrated in Section 3.3 of this research. When a proposed model is acceptable and available, an influential diagnostic study

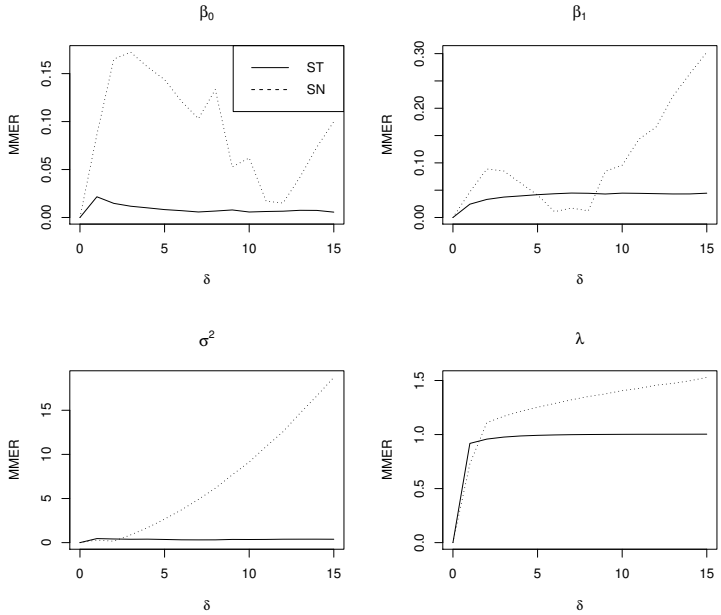


Figure 13 – Stellar Abundances dataset - contamination of observation 5. Mean magnitude of relative error (MMER) of the EM estimates for $\beta_0, \beta_1, \sigma^2$ and λ.

is recommended as the next step in the modeling process. In the present thesis, we deduced some diagnostic tools suitable for use in this model with applications in several areas. In particular, we propose the case exclusion technique and the local influence approach, under some perturbation schemes, for the SSMN-CR model, based on the Q function, inspired by the results of Zhu and Lee. Aspects of the robustness of the ML estimators under the SSMN-CR models were observed through influence analysis. In other words, in the context of asymmetry, once outliers are detected, models with heavier tails are robust alternatives to the SN-CR model. These techniques provide the professional with valuable tools that allow the identification of potentially influential elements and how to evaluate the real effects of disturbances on parameter estimates.

CHAPTER

4

CONCLUDING REMARKS

In this thesis, we document advances in research related to diagnostic analysis for SSMN-CR models. As mentioned previously, in statistical modeling, researchers typically begin by examining the inferential aspect of the model: how it is fitted, how its parameters can be estimated, which techniques are suitable, and the quality of the estimates, among other considerations. Once the inferential part is established, it is customary to evaluate the sensitivity of the model: Is it robust in its predictions, estimates, and adjustments? How does it respond to different intensities of disturbances? How does it react to influential observations or groups of observations?.

In this context, our objective was to address common problems such as lack of fit or inclusion of asymmetry in models based on normal distribution and increased robustness in the presence of outliers in the estimates. We achieved this by adopting the SSMN distribution family, which includes skewness control and heavier tails, allowing for less sensitive results even in the presence of outliers. Furthermore, this study focused on cases where observations are partially observed, such as when the response variable Y is censored. Therefore, the SSMN-CR model was studied and tested under different local and global diagnostic schemes.

In Chapter 3, within the context of global influence, we evaluate the impact of observations on the estimation of parameters when they are excluded from the data. In local influence analysis, we examine the detection of influential observations when some model assumption is disturbed. Finally, we evaluate this impact by removing these observations and calculating the RC measure, as presented in Table 7.

The research conducted demonstrated that the SSMN-CR models, yields a satisfactory fit to the data from the Stellar Abundances dataset - astrodatR, as evidenced in Section 3.3 of this study. Once a proposed model is deemed acceptable and available, conducting an influence diagnostics study becomes a recommended next step in the modeling process.

We have derived diagnostic tools suitable for application in this model, with relevance across various domains. Thus, it can be argued that the findings presented herein complement

those of Guzman, Ferreira and Zeller (2020). Specifically, we introduce the case-deletion technique and the local influence approach, incorporating various perturbation schemes, for the SSMN-CR model, utilizing the Q-function as inspired by Zhu and Lee. Through the analysis of influence, we observed aspects of the robustness of maximum likelihood estimators under the ST-CR, SSL-CR and SCN-CR models. Notably, in scenarios involving skewness, models exhibiting heavier tails emerge as robust alternatives to the SN-CR model once outliers are detected.

These techniques furnish practitioners with valuable tools for identifying potentially influential elements and assessing the actual effects of perturbations on parameter estimates.

The objective of this work is to identify influential observations within a censoring context. To achieve this, it is essential to include in the simulation an assessment of this capability in the context of censoring, varying sample sizes and parameter values. The magnitude of the imputed influence must also be analyzed and complemented, as the simulation was performed using standard deviation.

A concern arises when censoring mechanisms are implemented in the following manner: if a cutoff line for censoring is identified at a quantile of the sampled data, this implies that the cutoff line originates from a random process related to the distribution of the order statistics of the data. Consequently, the censorship distribution would be closely linked to the response distribution, and the likelihood analysis method would need to be restructured, as the likelihood approach assumes independent censoring.

If censoring is always based on the quantiles of the sample, the censoring distribution will be closely related to the distribution of the order statistics of that quantile. This raises implications for simulation processes. Specifically, if you are researching influential points and handling outliers on the right, caution is needed to avoid inadvertently removing genuine outliers using the cutoff line. Removing such outliers could introduce confusion and distort the simulation results, as it might make it appear as though there is always a single outlier present, leading to consistently skewed model outcomes.

Creating a simulated graphical scheme to expose patterns and understand how disturbances affect practice would be highly valuable. This approach could offer insights into different scenarios and enhance our understanding of disturbances.

Introducing disturbances in explanatory variables is straightforward for continuous variables, but less common for categorical variables. Exploring how to apply disturbances in categorical variables could be an interesting area of research.

Anomalously, the ST model, which performed poorly in simulations, actually performed best in practice for model adjustment. This discrepancy suggests the need for simulations that can better distinguish the identification potential of models and identify any simulation errors. More extensive exploration and testing of simulations are needed to draw reliable conclusions.

In practice, influential observations typically indicate outliers in usual regression contexts. However, in the models researched in this thesis, these observations do not behave as traditional outliers, given their similarity to generalized linear models.

When dealing with influential points, the final model is usually presented both with and without these points. This allows users to investigate how influential points impact parameter estimates and make informed decisions.

All methods developed in this work were implemented in the R Core Team (2019) software, and codes are available upon request.

The research described in this thesis is summarized in two publications that have been submitted to international journals.

4.1 Future Work Perspectives

A promising area for future research involves exploring the Mean Shift Outliers Models (MSOM) and its associated tests, as illustrated in the Subsection 4.1.1. MSOM is an extension of the SSMN-CR model, adding an additional parameter, denoted as ϕ, which is used to assess whether a specific observation can be considered an outlier. Therefore, by proposing the MSOM Model, we are also introducing outlier tests aimed at identifying and assessing the presence of outliers in the data modeled by SSMN-CR. These tests have the potential to significantly enhance our ability to detect and handle outliers in statistical analyses, thus contributing to a more comprehensive and accurate understanding of the observed data.

In concluding this thesis, it is advantageous to reflect on some insights gleaned from years of research and development in the aforementioned studies. During the literature review, it becomes evident that some authors emphasize a direct correlation between outlier tests and methods grounded in case deletion see for example Li, Xu and Zhu (2009). Furthermore, several authors propose that outlier tests can be effectively addressed through ridge and lasso-like techniques, which involve penalization in regression models. An intriguing recent study delves into the application of the EM algorithm to address regularization challenges in high-dimensional mixed-effects linear models in Oliveira, Schumacher and Lachos (2023). Further exploration may unveil the substantial potential of this work, as it encompasses the utilization of EM with lasso regularization, potentially facilitating a more comprehensive exploration of outlier tests.

In the articles by Zhang, Liu and Wu (2016) and Pan, Liu and Song (2021), there are studies developed on MSOU models and Outliers Tests, and they propose a very interesting idea, which is under the assumption of sparsity in the parameter ϕ incorporated in the regression to indicate whether the i-th observation is an outlier or not, having in the model more parameters than observations. The problem of detecting outliers can be solved through a penalized Variable Selection method, that is, they reformulate the task of detecting outliers into a high-dimensional

variable selection structure so that we can employ currently well-developed tools such as the method of regularization regression (Ridge and Lasso regression). These results can be extended to the SSMN-CR models.

The study of SSMN-CR models can be broadly extended, for example, to a Bayesian approach with reference to works such as Massuia *et al.* (2017).

4.1.1 The MSOM for SSMN-CR Models

In this Subsection, we briefly present the extension of SSMN-CR models to the proposed MSOM models as a potential direction for continuing the research presented in this thesis. First, we define the MSOM models and discuss the outlier test they enable. Then, we outline the estimation process using the EM algorithm, which allows for the construction of the Q function.

The specification of the SSMN-CR model is presented in Section 2.1.3. The MSOM was introduced in the work of Xie and Wei (2007), and its version for regression models under the SSMN family of distributions was developed by Ferreira, Mattos and Balakrishnan (2016). In this section, we present the theoretical development of the MSOM for the SSMN-CR model used in our research.

In general, the MSOM for the SSMN-CR models is defined as:

$$Y_i = \mathbf{x}_i^\top \beta + \xi_i, \quad i = 1,\ldots,n, \quad j \neq i \tag{4.1}$$

$$Y_i = \mathbf{x}_i^\top \beta + \phi + \xi_i, \quad i = 1,\ldots,n, \quad j = i \tag{4.2}$$

$$\xi_i \overset{iid}{\sim} SSMN(0,\sigma^2,\lambda,H),\ i = 1,\ldots,n, \tag{4.3}$$

where ϕ is an extra parameter that, if non-zero, indicates that the j-th case (observation) is a candidate for an outlier according to Cook, Holschuh and Weisberg (1982). Y_i is an observed continuous response variable for individual i and ξ_i is a random error. Associated with individual i, it is assumed a known $p \times 1$ covariate vector $\mathbf{x}_i$, as defined in Section 2.1.3.

Considering the assumption that the response variable is not fully observed for all subjects. For the i-th subject and assuming left-censoring, Y_i is a latent variable and the observed data take the form (V_i, ρ_i), according to Section 2.1.3.

Additionally, note that we can formulate an Outlier Test considering

$$H_0 : \phi = 0 \quad \text{vs} \quad H_1 : \phi \neq 0.$$

where, if H_0 is rejected, the j-th observation can be selected as a possible Outlier. To evaluate these hypotheses we can use asymptotic tests such as the likelihood ratio test, wald, score and gradient. We will see in the Subsection 4.1.2 further details about the asymptotic tests.

Under the established conditions the log-likelihood function of the $\theta = (\beta^\top, \phi, \sigma^2, \lambda, \tau^\top)^\top$

observed data is given by

$$\ell(\theta|\mathbf{v},\rho) = \ell_{[i]}(\theta)+\ell_j(\theta) \tag{4.4}$$

$$= \sum_{i=1,\ i\neq j}^{n} \ell_{[i]}(\theta)+\ell_j(\theta), \tag{4.5}$$

where

$$\ell_i(\theta) = \rho_i \log\left[F\left(\frac{v_i-\mu_i}{\sigma}\right)\right]+(1-\rho_i)\log[f_{SSMN}(v_i|\theta,H)], \tag{4.6}$$

with

$$\begin{cases} \mu_i = \mathbf{x}_i^\top\beta, & i=1,\ldots,n \quad \text{e} \quad i\neq j, \\ \mu_i = \mathbf{x}_i^\top\beta+\phi, & i=j, \end{cases} \tag{4.7}$$

due to the integrals present in equation (4.4), finding the MLE's (Maximum Likelihood Estimators) for θ by direct maximization of $\ell(\theta)$ becomes a difficult task. We will implement the EM algorithm. Thus, the MSOM can be described as follows:

$$\begin{aligned} Y_i|U_i=u_i,T_i=t_i &\overset{ind}{\sim} N\left(\mu_i+\frac{\sigma\lambda}{(u_i(u_i+\lambda^2))^{1/2}}t_i,\frac{\sigma^2}{u_i+\lambda^2}\right) \\ U_i &\overset{iid}{\sim} H(\tau) \\ T_i &\overset{iid}{\sim} TN(0,1;(0,+\infty)),\ i=1,\ldots,n, \end{aligned} \tag{4.8}$$

all independent, where $TN(r,s;(a,b))$ denotes the univariate normal distribution $(N(r,s))$, truncated on the interval (a,b). Defining the vectors $\mathbf{y}=(y_1,\ldots,y_n)^\top$, $\mathbf{t}=(t_1,\ldots,t_n)^\top$ and $\mathbf{u}=(u_1,\ldots,u_n)^\top$ we have that the complete data log-likelihood associated with $\mathbf{y}_c=(\mathbf{v}^\top,\rho^\top,\mathbf{y}^\top,\mathbf{t}^\top,\mathbf{u}^\top)$. Thus,

$$\begin{aligned} \ell_c(\theta|\mathbf{y}_c) = {} & c-n\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[u_iy_i^2-2\mu_iu_iy_i+\mu_i^2u_i+t_i^2 \\ & -2\lambda t_iy_i+2\lambda\mu_it_i+\lambda^2(y_i^2-2\mu_iy_i+\mu_i^2)], \end{aligned}$$

where c is a constant that does not depend on θ. So, it follows that, after algebraic manipulations, we have to

$$Q(\theta|\widehat{\theta}) = \sum_{i=1}^{n} Q_i(\theta|\widehat{\theta}) = \sum_{i=1,i\neq j}^{n} Q_i(\theta|\widehat{\theta})+Q_j(\theta|\widehat{\theta}) \tag{4.9}$$

$$= Q_{[j]}(\theta|\widehat{\theta})+Q_j(\theta|\widehat{\theta}), \tag{4.10}$$

where

$$\begin{aligned} Q(\theta|\widehat{\theta}) \propto {} & -n\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[\widehat{uy^2}_i-2\mu_i\widehat{uy}_i+\mu_i^2\widehat{u}_i+\widehat{t^2}_i-2\lambda\widehat{ty}_i \\ & +2\lambda\mu_i\widehat{t}_i+\lambda^2(\widehat{y^2}_i-2\mu_i\widehat{y}_i+\mu_i^2)]+\sum_{i=1}^{n}E[\log h(u_i|\tau)]. \end{aligned} \tag{4.11}$$

4.1.1.1 EM algorithm for MSOM

The specifications of the EM algorithm for MSOM are:

Step E: Giving $\theta = \widehat{\theta}^{(k)}$ in the k-th iteration. we need to calculate $\widehat{uy^2}_i, \widehat{uy}_i, \widehat{u}_i, \widehat{t^2}_i, \widehat{ty}_i, \widehat{t}_i, \widehat{y}_i$ and $\widehat{y}_i^2$, $i = 1, \ldots, n$, with

$$\begin{cases} \mu_i = \mathbf{x}_i^\top \beta, & i = 1, \ldots, n \quad \text{and} \quad i \neq j, \\ \mu_i = \mathbf{x}_i^\top \beta + \phi, & i = j, \end{cases} \tag{4.12}$$

Step M: Update $\widehat{\theta}^{(k+1)}$ by maximizing $Q(\theta|\widehat{\theta}^{(k)})$ over θ, which leads to the following closed form expressions

$$\begin{aligned}
\widehat{\beta}^{(k+1)} &= [\mathbf{X}^\top D(\widehat{\mathbf{u}}^{(k)} + \widehat{\lambda}^{2(k)} \mathbb{I}_n)\mathbf{X}]^{-1}\mathbf{X}^\top[\widehat{\mathbf{uy}}^{(k)} - \widehat{\lambda}^{(k)}\widehat{\mathbf{t}}^{(k)} + \widehat{\lambda}^{2(k)}\widehat{\mathbf{y}}^{(k)}] - \widehat{\phi}^{(k)}(1 + \widehat{u}_j^{(k)})\mathbf{X}_j, \\
\widehat{\phi}^{(k+1)} &= \frac{\widehat{\mathbf{uy}}_i^{(k)} - \widehat{\lambda}^{(k)}\widehat{\mathbf{t}}_j^{(k)} + \widehat{\lambda}^{2(k)}\widehat{\mathbf{y}}_j^{(k)}}{1 + \widehat{\mathbf{u}}_j} - \mathbf{x}_j^\top \widehat{\beta}^{(k)} \\
\widehat{\lambda}^{(k+1)} &= \frac{\sum_{i=1}^n [\widehat{ty}_i^{(k)} - \widehat{\mu}_i^{(k+1)}\widehat{t}_i^{(k)}]}{\sum_{i=1}^n [\widehat{y^2}_i^{(k)} - 2\widehat{\mu}_i^{(k+1)}\widehat{y}_i^{(k)} + \widehat{\mu}_i^{2(k+1)}]} \quad \text{and} \\
\widehat{\sigma^2}^{(k+1)} &= \frac{1}{2n}\sum_{i=1}^n [\widehat{uy}_i^{(k)} - 2\widehat{\mu}_i^{(k+1)}\widehat{uy}_i^{(k)} + \widehat{\mu}_i^{2(k+1)}\widehat{u}_i^{(k)} + \widehat{t^2}_i^{(k)} - 2\widehat{\lambda}^{(k+1)}\widehat{ty}_i^{(k)} \\
&\quad + 2\widehat{\lambda}^{(k+1)}\widehat{\mu}_i^{(k+1)}\widehat{t}_i^{(k)} + \widehat{\lambda}^{2(k+1)}(\widehat{y^2}_i^{(k)} - 2\widehat{\mu}_i^{(k+1)}\widehat{y}_i^{(k)} + \widehat{\mu}_i^{2(k+1)})].
\end{aligned}$$

the expressions for $\widehat{\lambda}^{(k+1)}$ and $\widehat{\sigma^2}^{(k+1)}$ are the same as in the article Guzman, Ferreira and Zeller (2020), where the parameter τ associated with the mixture variable U is known; see, for instance, Osorio, Paula and Galea (2007) and Zeller, Lachos and Vilca-Labra (2011). In this case, the profile likelihood and the Schwarz information criterion can be used for determining the optimum value of τ.

The iterations are repeated until a suitable convergence rule is satisfied. For example, the following criterion can be used: $||\widehat{\theta}^{(k+1)} - \widehat{\theta}^{(k)}|| < 10^{-4}$.

4.1.2 Asymptotic Tests

The hypotheses will be evaluated using the *Likelihood Ratio Test* and *The Gradient Test*.

4.1.2.1 Likelihood Ratio Test (LRT)

The statistic for this test defined in Neyman and Pearson (1928) is given by

$$MS_{LR} = 2[\ell(\widehat{\theta}) - \ell(\widetilde{\theta})],$$

where the log-likelihood function $\ell(\theta)$ is defined in 4.4, $\widehat{\theta}$ and $\widetilde{\theta}$ are the MLE's (Maximum Likelihood Estimators) on the restricted (Under H_0) and unrestricted (Under H_1) models, respectively.

4.1.2.2 Gradient Test (GT)

The statistic for this test defined in Terrell (2002), is given by

$$MS_G = S^\top(\widetilde{\theta})(\widehat{\theta} - \widetilde{\theta}),$$

where $S(\theta)$ is the score function.

4.1.3 Considerations

The MSOM in this section are being implemented and tested through simulations. However, this research is still ongoing and has not been completed due to the time constraints of finishing the doctorate and the extensive computational time required for these types of studies, which necessitate extensive simulations and experiments. The theoretical development was included in this work to illustrate that our research remains current and active.

For the Asymptotic Test we selected this both statistics (LRT and GT) for the ease of computing they offer. Furthermore, we have closed expressions for the Maximum Likelihood Estimators from the EM algorithm under both the restricted and unrestricted models for most parameters of interest. So, calculating the likelihood region statistic is not complicated when compared to the Wald and Score statistics which depend on Fisher's information matrix available in the work of Guzman, Ferreira and Zeller (2020).

Each of these statistics possesses an asymptotic distribution, specifically the chi-square (χ^2) distribution, under the null hypothesis (H_0). In practical terms, this means that as the sample size increases indefinitely, the distribution of these statistics converges to the chi-square distribution. Consequently, we can assess the significance of our findings by comparing the observed values of these statistics to critical values from the chi-square distribution.

To formally test the null hypothesis (H_0), we reject it at a significance level denoted by α if the computed values of the statistics fall below the corresponding critical values from the chi-square distribution. These critical values are determined by the degree of freedom associated with each statistic and the chosen significance level α. Specifically, if the computed statistic value is smaller than the corresponding quantile of the chi-square distribution at a level of $1-\alpha$, we reject the null hypothesis, indicating that there is sufficient evidence to conclude that the observed data deviates significantly from what would be expected under the null hypothesis.

BIBLIOGRAPHY

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: **Selected papers of hirotugu akaike**. [S.l.]: Springer, 1998. p. 199–213. Citation on page 55.

ANDREWS, D. F.; MALLOWS, C. L. Scale mixtures of normal distributions. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 36, n. 1, p. 99–102, 1974. Citation on page 40.

ARELLANO-VALLE, R. B.; CASTRO, L. M.; GONZÁLEZ-FARÍAS, G.; MUÑOZ-GAJARDO, K. A. Student-t censored regression model: properties and inference. **Statistical Methods & Applications**, v. 21, n. 4, p. 453–473, 2012. Citations on pages 31 and 37.

BRANCO, M. D.; DEY, D. K. A general class of multivariate skew-elliptical distributions. **Journal of Multivariate Analysis**, v. 79, n. 1, p. 99–113, 2001. Citation on page 35.

CASTRO, M. D.; GALEA-ROJAS, M.; BOLFARINE, H. Local influence assessment in heteroscedastic measurement error models. **Computational Statistics & Data Analysis**, v. 52, n. 2, p. 1132–1142, 2007. Citation on page 50.

COOK, R. D. Detection of influential observation in linear regression. **Technometrics**, Taylor & Francis, v. 19, n. 1, p. 15–18, 1977. Citations on pages 47 and 48.

COOK, R. D. Assessment of local influence. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley, v. 48, n. 2, p. 133–155, 1986. Citations on pages 39 and 47.

COOK, R. D.; HOLSCHUH, N.; WEISBERG, S. A note on an alternative outlier model. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 44, n. 3, p. 370–376, 1982. Citation on page 72.

COOK, R. D.; WEISBERG, S. **Residuals and influence in regression**. [S.l.]: New York: Chapman and Hall, 1982. Citations on pages 47 and 48.

DEMPSTER, A. Maximum likelihood estimation from incomplete data via the em algorithm. **Journal of the Royal Statistical Society**, v. 39, p. 1–38, 1977. Citation on page 35.

FERREIRA, C. d. S.; PAULA, G. A.; LANA, G. C. Estimation and diagnostic for partially linear models with first-order autoregressive skew-normal errors. **Computational Statistics**, Springer, v. 37, n. 1, p. 445–468, 2022. Citation on page 31.

FERREIRA, C. da S.; BOLFARINE, H.; LACHOS, V. H. Skew scale mixtures of normal distributions: Properties and estimation. **Statistical Methodology**, Elsevier, v. 8, n. 2, p. 154–171, 2011. Citations on pages 29, 35, 36, and 55.

FERREIRA, C. da S.; LACHOS, V. H.; GARAY, A. M. Inference and diagnostics for heteroscedastic nonlinear regression models under skew scale mixtures of normal distributions. **Journal of Applied Statistics**, Taylor & Francis, v. 47, n. 9, p. 1690–1719, 2020. Citation on page 31.

FERREIRA, C. S.; LACHOS, V. H.; BOLFARINE, H. Inference and diagnostics in skew scale mixtures of normal regression models. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 85, n. 3, p. 517–537, 2015. Citation on page 30.

FERREIRA, C. S.; MATTOS, T. B.; BALAKRISHNAN, N. Mean-shift outliers model in skew scale-mixtures of normal distributions. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 86, n. 12, p. 2346–2361, 2016. Citation on page 72.

FERREIRA, C. S.; ZELLER, C. B.; GARCIA, R. R. de O. Heteroscedastic partially linear model under skew-normal distribution with application in ragweed pollen concentration. **Journal of Applied Statistics**, Taylor & Francis, p. 1–28, 2022. Citation on page 31.

FUNG, W.-K.; ZHU, Z.-Y.; WEI, B.-C.; HE, X. Influence diagnostics and outlier tests for semiparametric mixed models. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 64, n. 3, p. 565–579, 2002. Citation on page 30.

GARAY, A. M.; LACHOS, V. H.; BOLFARINE, H.; CABRAL, C. R. B. Linear censored regression models with scale mixtures of normal distributions. **Statistical Papers**, v. 58, n. 1, p. 247–278, 2017. Citations on pages 31 and 37.

GELMAN, A.; HWANG, J.; VEHTARI, A. Understanding predictive information criteria for bayesian models. **Statistics and computing**, Springer, v. 24, n. 6, p. 997–1016, 2014. Citation on page 55.

GUZMAN, D. C. F.; FERREIRA, C. S.; ZELLER, C. B. Linear censored regression models with skew scale mixtures of normal distributions. **Journal Applied Statistics**, https://doi.org/10.1080/02664763.2020.1795814, p. 1–26, 2020. Citations on pages 13, 15, 29, 33, 36, 54, 55, 66, 70, 74, and 75.

LACHOS, V. H.; CABRAL, C. R. B. Diagnostic tools for identifying outliers in multivariate skew-t regression models. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, n. 12, p. 2320–2337, 2017. Citation on page 43.

LANGE, K.; SINSHEIMER, J. S. Normal/independent distributions and their applications in robust regression. **Journal of Computational and Graphical Statistics**, Taylor & Francis Group, v. 2, n. 2, p. 175–198, 1993. Citations on pages 36 and 55.

LAWRANCE, A. Deletion influence and masking in regression. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 57, n. 1, p. 181–189, 1995. Citations on pages 38 and 39.

LAWRANCE, A. J. Deletion influence and masking in regression. **Journal of the Royal Statistical Society: Series B**, v. 57, n. 1, p. 181–189, 1995. Citation on page 48.

LEE, S.-Y.; XU, L. Influence analyses of nonlinear mixed-effects models. **Computational Statistics & Data Analysis**, Elsevier, v. 45, n. 2, p. 321–341, 2004. Citation on page 49.

LEIVA, V.; BARROS, M.; PAULA, G. A.; GALEA, M. Influence diagnostics in log-birnbaum–saunders regression models with censored data. **Computational Statistics & Data Analysis**, v. 51, p. 5694–5707, 2007. Citation on page 50.

LI, A.-P.; CHEN, Z.-X.; XIE, F.-C. Diagnostic analysis for heterogeneous log-birnbaum–saunders regression models. **Statistics & Probability Letters**, Elsevier, v. 82, n. 9, p. 1690–1698, 2012. Citation on page 31.

LI, Z.; XU, W.; ZHU, L. Influence diagnostics and outlier tests for varying coefficient mixed models. **Journal of Multivariate Analysis**, Elsevier, v. 100, n. 9, p. 2002–2017, 2009. Citations on pages 38, 39, 48, and 71.

LOUREDO, G. M.; ZELLER, C. B.; FERREIRA, C. S. Estimation and influence diagnostics for the multivariate linear regression models with skew scale mixtures of normal distributions. **Sankhya B**, Springer, p. 1–39, 2021. Citations on pages 30, 38, and 48.

MASSUIA, M. B.; CABRAL, C. R. B.; MATOS, L. A.; LACHOS, V. H. Influence diagnostics for student-t censored linear regression models. **Statistics**, v. 49, n. 5, p. 1074–1094, 2015. Citation on page 30.

MASSUIA, M. B.; GARAY, A. M.; CABRAL, C. R.; LACHOS, V. Bayesian analysis of censored linear regression models with scale mixtures of skew-normal distributions. **Statistics and its Interface**, International Press of Boston, v. 10, n. 3, p. 425–439, 2017. Citation on page 72.

MATOS, L. A.; LACHOS, V. H.; LIN, T. I.; CASTRO, L. M. Heavy-tailed longitudinal regression models for censored data: a robust parametric approach. **Test**, v. 28, p. 844–878, 2019. Citation on page 30.

MATTOS, T. d. B.; GARAY, A. M.; LACHOS, V. H. Likelihood-based inference for censored linear regression models with scale mixtures of skew-normal distributions. **Journal of Applied Statistics**, Taylor & Francis, v. 45, n. 11, p. 2039–2066, 2018. Citation on page 54.

MIRFARAH, E.; NADERI, M.; CHEN, D.-G. Mixture of linear experts model for censored data: A novel approach with scale-mixture of normal distributions. **Computational Statistics & Data Analysis**, Elsevier, v. 158, p. 107182, 2021. Citation on page 42.

NEYMAN, J.; PEARSON, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part ii. **Biometrika**, JSTOR, v. 20, n. 3/4, p. 263–294, 1928. Citation on page 74.

OLIVEIRA, D. C.; SCHUMACHER, F. L.; LACHOS, V. H. The use of the em algorithm for regularization problems in high-dimensional linear mixed-effects models. **arXiv preprint arXiv:2308.01518**, 2023. Citation on page 71.

OSORIO, F.; PAULA, G. A.; GALEA, M. Assessment of local influence in elliptical linear models with longitudinal structure. **Computational Statistics & Data Analysis**, Elsevier, v. 51, n. 9, p. 4354–4368, 2007. Citation on page 74.

PAN, Y.; LIU, Z.; SONG, G. Outlier detection under a covariate-adjusted exponential regression model with censored data. **Computational Statistics**, Springer, v. 36, n. 2, p. 961–976, 2021. Citation on page 71.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019. Available: <https://www.R-project.org/>. Citation on page 71.

RAMOS, P. L.; GUZMAN, D. C.; MOTA, A. L.; RODRIGUES, F. A.; LOUZADA, F. Sampling with censored data: a practical guide. **arXiv preprint arXiv:2011.08417**, 2020. Citations on pages 33 and 40.

SANCHEZ, L. B.; FERREIRA, C. da S. **ssmn: Skew Scale Mixtures of Normal Distributions**. [S.l.], 2016. R package version 1.1. Available: <https://cran.r-project.org/package=ssmn>. Citation on page 42.

SANTOS, D. M.; LACHOS, V. H. Simulation-based diagnostics for heavy-tailed and skewed regression models. **Statistics and Computing**, Springer, v. 31, n. 1, p. 289–305, 2021. Citation on page 43.

SANTOS, N.; LÓPEZ, R. G.; ISRAELIAN, G.; MAYOR, M.; REBOLO, R.; TAORO, M. R. Perez de; RANDICH, S. Beryllium abundances in stars hosting giant planets. **Astronomy & Astrophysics**, v. 386, n. 3, p. 1028–1038, 2002. Citation on page 54.

SCHUMACHER, F. L.; LACHOS, V. H.; VILCA-LABRA, F. E.; CASTRO, L. M. Influence diagnostics for censored regression models with autoregressive errors. **Australian & New Zealand Journal of Statistics**, Wiley Online Library, v. 60, n. 2, p. 209–229, 2018. Citation on page 53.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, JSTOR, p. 461–464, 1978. Citation on page 55.

TERRELL, G. R. The gradient statistic. **Computing Science and Statistics**, v. 34, n. 34, p. 206–215, 2002. Citation on page 75.

WU, L. **Mixed Effects Models for Complex Data**. [S.l.]: Boca Raton, FL: Chapman & Hall/CRC, 2010. Citation on page 31.

XIE, F.-C.; WEI, B.-C. Diagnostics analysis for log-birnbaum–saunders regression models. **Computational statistics & data analysis**, Elsevier, v. 51, n. 9, p. 4692–4706, 2007. Citation on page 72.

XU, L.; LEE, S.-Y.; POON, W.-Y. Deletion measures for generalized linear mixed effects models. **Computational Statistics & Data Analysis**, Elsevier, v. 51, n. 2, p. 1131–1146, 2006. Citation on page 48.

ZELLER, C. B.; CABRAL, C. R. B.; LACHOS, V. H.; BENITES, L. Finite mixture of regression models for censored data based on scale mixtures of normal distributions. **Advances in Data Analysis and Classification**, Springer, v. 13, n. 1, p. 89–116, 2019. Citation on page 30.

ZELLER, C. B.; LABRA, F. V.; LACHOS, V. H.; BALAKRISHNAN, N. Influence analyses of skew-normal/independent linear mixed models. **Computational Statistics & Data Analysis**, Elsevier, v. 54, n. 5, p. 1266–1280, 2010. Citations on pages 30 and 31.

ZELLER, C. B.; LACHOS, V. H.; VILCA-LABRA, F. E. Local influence analysis for regression models with scale mixtures of skew-normal distributions. **Journal of Applied Statistics**, v. 38, n. 2, p. 343–368, 2011. Citations on pages 30 and 74.

ZHANG, J.; LIU, Y.; WU, Y. Exponential regression for censored data with outliers. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 86, n. 3, p. 431–442, 2016. Citation on page 71.

ZHU, H.; LEE, S. Local influence for incomplete data models. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 63, n. 1, p. 111–126, 2001. Citations on pages 13, 15, 29, 36, 39, 47, 48, 49, and 54.

ZHU, H.; LEE, S.; WEI, B.; ZHOU, J. Case-deletion measures for models with incomplete data. **Biometrika**, JSTOR, p. 727–737, 2001. Citations on pages 13, 15, 36, 38, 47, and 48.

ZHU, Z.-Y.; HE, X.; FUNG, W.-K. Local influence analysis for penalized gaussian likelihood estimators in partially linear models. **Scandinavian Journal of Statistics**, v. 30, n. 4, p. 767–780, 2003. Citation on page 30.

APPENDIX

A

ALGORITHMS AND CODES

Algorithm 4 – Data Generation Algorithm

1: **procedure** GENERATESAMPLES(*n*, *location*, *scale*, *shape*, *nu*, *gamma*)
2: Initialize parameters (location, scale, shape, nu, gamma) if provided; otherwise, use default values.
3: Generate *n* random numbers *u* from a uniform distribution.
4: Perform transformations on *u* to obtain *ku*, *shape*1, and *scale*1 according to the distribution's characteristics.
5: Generate random samples *y* using the transformed parameters and additional random numbers generated from a standard normal distribution.
6: **return** the generated random samples *y*.
7: **end procedure**

Algorithm 5 – RSSN: Algorithm for Generating Skew Normal Distributed Random Variables

procedure RSSN($n, location, scale, shape, dp$)
  Input: n - number of random variables to generate
    $location$ - location parameter of the Skew-Normal distribution
    $scale$ - scale parameter of the Skew-Normal distribution
    $shape$ - shape parameter of the Skew-Normal distribution
    dp - optional parameter vector specifying $location$, $scale$, and $shape$
  Output: Vector of n random variables from the Skew-Normal distribution
  if dp is not **NULL then**
    if $shape$ is not missing **then**
      Stop: You cannot set both component parameters and dp
    end if
    $location \leftarrow dp[1]$
    $scale \leftarrow dp[2]$
    $shape \leftarrow dp[3]$
  end if
  if $scale \leq 0$ **then**
    Stop: Parameter $scale$ must be positive
  end if
  $\delta \leftarrow \frac{shape}{\sqrt{1+shape^2}}$
  for $i = 1$ to n **do**
    Generate standard normal random variables $u1$ and $u2$
    $y_i \leftarrow location + \sqrt{scale} \times (\delta \times |u1| + \sqrt{1-\delta^2} \times u2)$
  end for
  return Vector y containing the generated random variables
end procedure

Algorithm 6 – RSTN: Algorithm for Generating Skew Student-t Distributed Random Variables

procedure RSTN($n, location, scale, shape, nu, dp$) ▷ Generate n skew Student-t distributed random variables
  if !is.null(dp) **then**
    if !missing($shape$) **then**
      stop("You cannot set both component parameters and dp")
    end if
    $location \leftarrow dp[1]$
    $scale \leftarrow dp[2]$
    $shape \leftarrow dp[3]$
    $nu \leftarrow dp[4]$
  end if
  if $nu \leq 0$ **then**
    stop("Parameter nu must be positive")
  end if
  if $scale \leq 0$ **then**
    stop("Parameter scale must be positive")
  end if
  if $nu < 1$ **then**
    warning("Nu < 1 can generate values tending to infinite", call. = FALSE)
  end if
  $u \leftarrow \text{rgamma}(n, \frac{nu}{2}, \frac{nu}{2})$
  $ku \leftarrow \frac{1}{u}$
  $shape1 \leftarrow shape \times \sqrt{ku}$
  $scale1 \leftarrow scale \times ku$
  $\delta \leftarrow \frac{shape1}{\sqrt{1+shape1^2}}$
  $u1 \leftarrow \text{rnorm}(n)$
  $u2 \leftarrow \text{rnorm}(n)$
  $y \leftarrow location + \sqrt{scale1} \times (\delta \times |u1| + \sqrt{1-\delta^2} \times u2)$
  return(y)
end procedure

Algorithm 7 – RSSL: Algorithm for Generating Skew Slash Distributed Random Variables

procedure RSSL($n, location, scale, shape, nu, dp$) ▷ Generate n scaled stable distributed random variables
  if !is.null(dp) **then**
    if !missing($shape$) **then**
      stop("You cannot set both component parameters and dp")
    end if
    $location \leftarrow dp[1]$
    $scale \leftarrow dp[2]$
    $shape \leftarrow dp[3]$
    $nu \leftarrow dp[4]$
  end if
  if $nu \leq 0$ **then**
    stop("Parameter nu must be positive")
  end if
  if $scale < 0$ **then**
    stop("Parameter scale must be positive")
  end if
  $v \leftarrow \text{runif}(n, 0, 1)$
  $u \leftarrow v^{1/nu}$
  $ku \leftarrow \frac{1}{u}$
  $shape1 \leftarrow shape \times \sqrt{ku}$
  $scale1 \leftarrow scale \times ku$
  $\delta \leftarrow \frac{shape1}{\sqrt{1+shape1^2}}$
  $u1 \leftarrow \text{rnorm}(n)$
  $u2 \leftarrow \text{rnorm}(n)$
  $y \leftarrow location + \sqrt{scale1} \times (\delta \times |u1| + \sqrt{1-\delta^2} \times u2)$
  return(y)
end procedure

Algorithm 8 – RSCN: Algorithm for Generating Skew Contaminated Normal Distributed Random Variables

procedure RSCN($n, location, scale, shape, nu, gama, dp$) ▷ Generate n scaled compound normal distributed random variables
  if !is.null(dp) **then**
    if !missing($shape$) **then**
      stop("You cannot set both component parameters and dp")
    end if
    $location \leftarrow dp[1]$
    $scale \leftarrow dp[2]$
    $shape \leftarrow dp[3]$
    $nu \leftarrow dp[4]$
    $gama \leftarrow dp[5]$
  end if
  if $nu \leq 0$ **or** $nu \geq 1$ **then**
    stop("Parameter nu must be between 0 and 1.0")
  end if
  if $gama \leq 0$ **or** $gama \geq 1$ **then**
    stop("Parameter gama must be between 0 and 1.0")
  end if
  if $scale \leq 0$ **then**
    stop("Parameter scale must be positive")
  end if
  $uu \leftarrow \text{rbinom}(n, 1, nu)$
  $u \leftarrow gama \times uu + (1 - uu)$
  $ku \leftarrow \frac{1}{u}$
  $shape1 \leftarrow shape \times \sqrt{ku}$
  $scale1 \leftarrow scale \times ku$
  $\delta \leftarrow \frac{shape1}{\sqrt{1+shape1^2}}$
  $u1 \leftarrow \text{rnorm}(n)$
  $u2 \leftarrow \text{rnorm}(n)$
  $y \leftarrow location + \sqrt{scale1} \times (\delta \times |u1| + \sqrt{1-\delta^2} \times u2)$
  return(y)
end procedure

Source code 1 – R Code-Left Censorship Generator

```
ni <- n # number of observations
ci <- perc # censoring percentage

# Introducing left censoring
nc = floor(ni*ci) # number of censored observations
ind_censored = sort(sample(1:ni, nc, replace = FALSE)) #
    indices of censored observations
u = runif(nc) # random numbers
```

```
 8: c1 = mapply(function(ic, i) max(c(y[ic] - u[i], y[ic]+u[i]-1)),
       ind_censored, 1:length(ind_censored)) # thresholds for left
       censoring
 9: for(i in 1:length(ind_censored)){
10:   y[ind_censored[i]] = c1[i] # replacing censored observations
      with thresholds
11: }
12:
13: phi <- as.numeric(1:length(y) %in% ind_censored) # creating
      indicator variable for censored observations
```

Source code 2 – R Code-Right Censorship Generator

```
 1: beta0=matrix(c(1,-1,2,-2)) # Define the coefficients for the
      regression model
 2:
 3: sigma2=1 # Set the variance of the error term
 4:
 5: # Set parameters for the distribution of the error term
 6: lambda=-3
 7: nu=5
 8:
 9: n=100 # Define the sample size
10:
11: # Generate random values for the predictor variables
12: x1=rnorm(n,0,1)
13: x2=rnorm(n,0,1)
14: x3=rnorm(n,0,1)
15:
16: # Create the design matrix X
17: X=matrix(1,n,4)
18: X[,2]=x1
19: X[,3]=x2
20: X[,4]=x3
21:
22: mu=X %*% beta0 # Calculate the mean values using the regression
       coefficients
23:
24: # Generate the error term following a skew-t normal
      distribution
25: erro=rstn(n, location=0, scale=sigma2, shape=lambda, nu=nu)
26:
```

```
# Generate the response variable y as the sum of the mean
    values and error term
y=mu+erro

# Define the level of censorship
perc <- 0.2

# Determine the number of censored observations
ni<-n
ci<-perc
nc = floor(ni*ci)

# Sample the indices of censored observations
ind_censored = sort(sample(1:ni, nc, replace = F))

# Generate random values for censoring
u = runif(nc)

# Apply left censoring mechanism
c1=-Inf
c2 = mapply(function(ic, i) min(c(y[ic] + u[i], y[ic]-u[i]+1)),
     ind_censored, 1:length(ind_censored))
for(i in 1:length(ind_censored)){
  y[ind_censored[i]]=c2[i]
}

cutof=c2 # Store the censoring cutoff values

# Create an indicator variable for censored observations
aux1=rep(0,n)
aux1[ind_censored]=1
cc=aux1

y # Display the resulting response variable y
```

Source code 3 – R Code-Interval Censorship Generator

```
Setting the seed for reproducibility
set.seed(2024)

Parameters
beta0 <- matrix(c(1, -1, 2, -2))
```

```
sigma2 <- 1
lambda <- -3
nu <- 5
n <- 10

Generating covariates
x1 <- rnorm(n, 0, 1)
x2 <- rnorm(n, 0, 1)
x3 <- rnorm(n, 0, 1)
X <- matrix(1, n, 4)
X[, 2] <- x1
X[, 3] <- x2
X[, 4] <- x3

Calculating the mean
mu <- X %*% beta0

Generating errors following a skew t-normal distribution
erro <- rstn(n, location = 0, scale = sigma2, shape = lambda,
    nu = nu)

Generating the response variable
y <- mu + erro

Applying interval censoring
perc <- 0.2 # Censoring level
ni <- n
ci <- perc

Determining the number of censored observations
nc <- floor(ni * ci)

Selecting random indices for censoring
ind_censored <- sort(sample(1:ni, nc, replace = FALSE))

Generating random thresholds for censoring
u <- runif(nc)
c1 <- mapply(function(ic, i) max(c(y[ic] - u[i], y[ic] + u[i] -
    1)), ind_censored, 1:length(ind_censored))
c2 <- mapply(function(ic, i) min(c(y[ic] + u[i], y[ic] - u[i] +
    1)), ind_censored, 1:length(ind_censored))

```

```
45: Applying censoring
46: for (i in 1:length(ind_censored)) {
47: y[ind_censored[i]] <- (c1[i] + c2[i]) / 2
48: }
49:
50: Displaying the censored data
51: y
```


ICMC
USP
SÃO CARLOS