University of Brasília
Institute of Exact Sciences
Department of Statistics

Master's Dissertation

Symmetric generalized Heckman models

by

Shayane dos Santos Cordeiro

Brasília, April 2022

Symmetric generalized Heckman models

by

Shayane dos Santos Cordeiro

A dissertation submitted to the Departament of Statistics at the University of Brasília in partial fulfilment of the requirements for the degree Master in Statistics.

Supervisor: Prof. Dr. Helton Saulo Bezerra dos Santos.

Brasília, April 2022

A dissertation submitted to the Departament of Statistics at the University of Brasília in partial fulfilment of the requirements for the degree Master in Statistics.

Approved by:

Prof. Dr. Helton Saulo Bezerra dos Santos

Supervisor, EST/UnB

Prof. Dr. Roberto Vila Gabriel

EST/UnB

Prof. Dr. Jeremias da Silva Leão

(DE/UFAM)

Prof. Dr. José Augusto Fiorucci

EST/UnB

There is no branch of mathematics, however abstract,
which may not some day be applied to phenomena of the real world.
(Nikolai Lobachevsky)

To my family and friends.

Acknowledgments

Thanks to the faculty of PPGEST/UnB and my supervisor Helton Saulo Bezerra dos Santos.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Resumo

MODELOS DE HECKMAN GENERALIZADOS SIMÉTRICOS

O problema de viés de seleção amostral surge quando uma variável de interesse está correlacionada com uma variável latente, e envolve situações em que a variável de interesse tem parte das suas observações censuradas. A censura é uma espécie de limitação na amostra em que determinadas observações da variável resposta não são verificadas, não por sua ausência, mas por vezes porque o objeto de estudo não sofreu o evento de interesse, porém outras informações que ajudam a explicar o evento foram obtidas. Esse problema ocorre, em diversas áreas da Economia, Ciências Políticas, Estatística, Sociologia entre outras.

Para evitar problemas de seleção amostral o recomendado é utilizar toda a amostra de dados, uma vez que as variáveis explicativas foram observadas e a variável resposta censurada pode transmitir informação sobre todo o conjunto de dados. Uma forma de verificar se a variável de interesse censurada transmite informação é utilizar uma covariável que capture o viés ao se considerar uma amostra, em que apenas as variáveis dependentes foram observadas. Caso esse viés seja significativo, deve-se trabalhar com a base de dados completa

O matemático e economista James Joseph Heckman foi o primeiro a estudar dados com de viés de seleção amostral e em 1976 propôs um modelo de seleção amostral baseado na distribuição normal bivariada que considera tanto a variável de interesse quanto a variável la-

tente, apesar do seu empenho o método utilizava a estimação por máxima verossimilhança e foi bastante criticado devido a dificuldade de sua implementação e suposições do modelo, o que o levou a propor um modelo alternativo mais simples denominado método dos dois passos, também conhecido como modelo Tobit tipo 2, na literatura econométrica. Estudos propostos, tais como Nelson (1984), Paarsch (1984), Manning, Duan, and Rogers (1987), Stolzenberg and Relles (1990) and Leung and Yu (1996) sugerem que o modelo pode reduzir ou eliminar o viés de seleção quando seus pressupostos são atendidos. Contudo o desvio de normalidade pode ocasionar uma distorção nos resultados ou mesmo inviabilizar o ajuste.

A suposição de normalidade tem sido relaxada por modelos mais flexíveis, ao sugerir o uso de outras distribuições bivariadas em substituição a distribuição normal tais como a Student-t aplicada por Marchenko and Genton (2012) and Lachos, Prates, and Dey (2021) que apresenta caudas mais pesadas e permite ajustes mais robustos, Skew-normal (Ogundimu and Hutton, 2016), abordagem Bayesiana Ding (2014) e baseada em cópulas Lee (1983). Abordagens semi-paramétricas (Ahn and Powell, 1993) e não-paramétricas (M. Das and Vella, 2003) também foram consideradas, contudo as abordagens paramétricas permitem identificar o intercepto do modelo o que pode ser útil em análises com predições.

No modelo de Heckman Clássico os erros são normalmente distribuídos, com parâmetros de dispersão e correlação constantes, a generalização do modelo clássico consiste em introduzir covariáveis aos parâmetros de dispersão e correlação, a fim de modelar dados reais que frequentemente apresentam dispersão variável, possibilitando a identificação de covariáveis responsáveis pela variabilidade dos dados e o viés de seleção. Nesse sentido este trabalho tem como objetivo propor modelos de seleção amostral Heckman generalizados baseados nas distribuições simétricas (Fang, Kotz, and Ng, 1990). Trata-se de uma nova classe de modelo de seleção amostral em que são acrescidas covariáveis aos parâmetros de dispersão e de correlação, que possibilitam explicar a heterocedasticidade e o viés de seleção amostral respectivamente.

Neste estudo, na seção 1.2 introduzimos o modelo de Heckman generalizado simétrico, obtendo sua função densidade de probabilidade, que apresenta dois componentes um discreto e

outro contínuo, que é utilizada para a estimação dos parâmetros do modelo através da função de log-verossimilhança. Na seção 1.3 derivamos o modelo de Heckman-Student-t generalizado que é um caso especial do modelo de Heckman generalizado simétrico, obtendo a função densidade de probabilidade e estimando os parâmetros do modelo.

Na seção 1.4, um estudo de simulação de Monte Carlo realizado para avaliar o comportamento do método de estimação de parâmetros dos modelos de Heckman-normal generalizado e Heckman-Student-t utilizando o viés e o Erro Quadrático Médio (EQM), considerando quatro cenários mostrou bons resultados, na presença de altas/baixas taxas de censura e correlação.

Dois conjuntos de dados reais, gastos ambulatoriais da base *Medical Expenditure Panel Survey (MEPS)* de 2001, também utilizados por Cameron and Trivedi (2009), Marchenko and Genton (2012), M. Zhelonkin and Ronchetti (2016) e Bastos and Barreto-Souza (2020), disponível no *software* R via pacote *ssmrob* de M. Zhelonkin et al. (2016) e as bases públicas dos governos dos Estados de São Paulo e Minas Gerais, com covariáveis que explicam o Investimento em Educação (IE) no ano de 2018, são analisados, na seção 1.5, para ilustrar a abordagem proposta e revelaram o bom ajuste do modelo de Heckman-t generalizado comparado com o modelo normal generalizado, além estimação dos parâmetros, também foram obtidos os resíduos do tipo-martingale (MT) e o ajuste dos respectivos quantis favoreceram o modelo proposto no estudo que se ajusta melhor a dados com valores extremos.

Palavras-chave: Modelos de Heckman generalizados, distribuições simétricas, dispersão variável, correlação variável.

Abstract

The sample selection bias problem arises when a variable of interest is correlated with a latent variable, and involves situations in which the response variable had part of its observations censored. Heckman (1976) proposed a sample selection model based on the bivariate normal distribution that fits both the variable of interest and the latent variable. Recently, this assumption of normality has been relaxed by more flexible models such as the Student-t distribution (Marchenko and Genton, 2012; Lachos, Prates, and Dey, 2021). The aim of this work is to propose generalized Heckman sample selection models based on symmetric distributions (Fang, Kotz, and Ng, 1990). This is a new class of sample selection models, in which variables are added to the dispersion and correlation parameters. A Monte Carlo simulation study is performed to assess the behavior of the parameter estimation method. Two real data sets are analyzed to illustrate the proposed approach.

Keywords: Generalized Heckman models, symmetric distributions, variable dispersion, variable correlation.

Contents

Chapter 1

Symmetric generalized Heckman models

1.1 Introduction

It is common in the areas of economics, statistics, sociology, among others, that in the sampling process there is a relationship between a variable of interest and a latent variable, in which the former is observable only in a subset of the population under study. This problem is called sample selection bias and was studied by Heckman (1976). The author proposed a sample selection model by joint modeling the variable of interest and the latent variable. The classical Heckman sample selection (classical Heckman-normal model) model received several criticisms, due to the need to assume bivariate normality and the difficulty in estimating the parameters using the maximum likelihood (ML) method, which led to the introduction of an alternative estimation method known as the two-step method; see Heckman (1979). Some studies on Heckman models have been done by Nelson (1984), Paarsch (1984), Manning, Duan, and Rogers (1987), Stolzenberg and Relles (1990) and Leung and Yu (1996). These works suggested that the Heckman sample selection model can reduce or eliminate selection bias when the assumptions hold, but deviation from normality assumption may distort the results.

The normality assumption of the classical Heckman-normal model (Heckman, 1976) has been relaxed by more flexible models such as the Student-t distribution (Marchenko and Gen-

ton, 2012; Ding, 2014; Lachos, Prates, and Dey, 2021) and the skew-normal distribution (Ogundimu and Hutton, 2016). Moreover, the classical Heckman-normal model assumes that the dispersion and correlation (sample selection bias parameter) are constant, which may not be adequate. In this context, the present work aims to propose generalized Heckman sample selection models based on symmetric distributions (Fang, Kotz, and Ng, 1990). In the proposed model, covariates are added to the dispersion and correlation parameters, then we have covariates explaining possible heteroscedasticity and sample selection bias, respectively. Our proposed methodology can be seen as a generalization of the generalized Heckman-normal model with varying sample selection bias and dispersion parameters by Bastos, Barreto-Souza, and Genton (2021), which is based on the bivariate normal distribution as the classical Heckman-normal model.

The rest of this work proceeds as follows. In Section 1.2, we briefly describe the bivariate symmetric distributions. We then introduce the symmetric generalized Heckman models. In this section, we also describe the maximum likelihood (ML) estimation of the model parameters. In Section 1.3, we derive the generalized Heckman-Student-t model, which is a special case of the symmetric generalized Heckman models. In Section 1.4, we carry out a Monte Carlo simulation study for evaluating the performance of the estimators. In Section 1.5, we apply the generalized Heckman-Student-t to two real data sets to demonstrate the usefulness of the proposed model, and finally in Section 1.6, we provide some concluding remarks.

1.2 Symmetric generalized Heckman models

Let $\boldsymbol{Y} = (Y_1, Y_2)^\top$ be a random vector following a bivariate symmetric (BSY) distribution (Fang, Kotz, and Ng, 1990) with location (mean) vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$, covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

and density generator g_c, with $\mu_i \in \mathbb{R}$, $\sigma_i > 0$, for $i = 1, 2$. We use the notation $\boldsymbol{Y} \sim \mathrm{BSY}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g_c)$. Then, the probability density function (PDF) of $\boldsymbol{Y} \sim \mathrm{BSY}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g_c)$ is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, g_c) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} g_c\left((\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \quad \boldsymbol{y} \in \mathbb{R}^2. \tag{1.1}$$

The density generator g_c in (1.1) leads to different bivariate symmetric distributions, which may contain an extra parameter (or extra parameter vector).

We propose a generalization of the classical Heckman-normal model (Heckman, 1976) by considering independent errors terms following a BSY distribution with regression structures for the sample selection bias ($0 < \rho < 1$) and dispersion ($\sigma > 0$) parameters:

$$\begin{pmatrix} Y_i^* \\ U_i^* \end{pmatrix} \sim \mathrm{BSY}\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_i^2 & \rho_i\sigma_i \\ \rho_i\sigma_i & 1 \end{pmatrix}, g_c\right), \quad i = 1, \ldots, n. \tag{1.2}$$

In the above equation $\mu_{1i}, \mu_{2i}, \sigma_i$ and ρ_i are are the mean, dispersion and correlation parameters, respectively, with the following regression structure $g_1(\mu_{1i}) = \boldsymbol{x}_i^\top\boldsymbol{\beta}$, $g_2(\mu_{2i}) = \boldsymbol{w}_i^\top\boldsymbol{\gamma}$, $h_1(\sigma_i) = \boldsymbol{z}_i^\top\boldsymbol{\lambda}$ and $h_2(\rho_i) = \boldsymbol{v}_i^\top\boldsymbol{\kappa}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_l)^\top \in \mathbb{R}^l$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top \in \mathbb{R}^p$ and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_q)^\top \in \mathbb{R}^q$ are vectors of regression coefficients, $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ik})^\top$, $\boldsymbol{w_i} = (w_{i1}, \ldots, w_{il})^\top$, $\boldsymbol{z_i} = (z_{i1}, \ldots, z_{ip})^\top$ and $\boldsymbol{v_i} = (v_{i1}, \ldots, v_{iq})^\top$ are the values of k, l, p and q covariates, and $k + l + p + q < n$. The links $g_1(\cdot), g_2(\cdot), h_1(\cdot)$ and $h_2(\cdot)$ are strictly monotone and twice differentiable. The link functions $g_1 : \mathbb{R} \to \mathbb{R}$, $g_2 : \mathbb{R} \to \mathbb{R}$, $h_1 : \mathbb{R}^+ \to \mathbb{R}$ and $h_2 : [-1, 1] \to \mathbb{R}$ must be strictly monotone, and at least twice differentiable, with $g_1^{-1}(\cdot)$, $g_2^{-1}(\cdot)$, $h_1^{-1}(\cdot)$, and $h_2^{-1}(\cdot)$ being the inverse functions of $g_1(\cdot)$, $g_2(\cdot)$, $h_1(\cdot)$, and $h_2(\cdot)$, respectively. For $g_1(\cdot)$ and $g_2(\cdot)$ the most common choice is the identity link, whereas for $h_1(\cdot)$ and $h_2(\cdot)$ the most common choices are logarithm and arctanh (inverse hyperbolic tangent) links, respectively.

We can agglutinate the information from U_i^* in the following indicator function $U_i = \mathbb{1}_{\{U_i^* > 0\}}$. Let $Y_i = Y_i^* U_i$ be the observed outcome, for $i = 1, \ldots, n$. Only n_1 out of n ob-

servations Y_i^* for which $U_i^* > 0$ are observed. This model is known as "Type 2 tobit model" in the econometrics literature. Notice that $U_i \sim \text{Bernoulli}(\mathbb{P}(U_i^* > 0))$. By using law of total probability, for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \sigma, \rho)^\top$, the random variable Y_i has distribution function

$$
\begin{aligned}
F_{Y_i}(y_i;\boldsymbol{\theta}) &= \mathbb{P}(Y_i \leq y_i | U_i^* > 0)\mathbb{P}(U_i^* > 0) + \mathbb{P}(Y_i \leq y_i | U_i^* > 0)\mathbb{P}(U_i^* \leq 0) \\
&= \begin{cases} \mathbb{P}(Y_i^* \leq y_i | U_i^* > 0)\mathbb{P}(U_i^* > 0), & \text{if } y_i < 0, \\ \mathbb{P}(Y_i^* \leq y_i | U_i^* > 0)\mathbb{P}(U_i^* > 0) + \mathbb{P}(U_i^* \leq 0), & \text{if } y_i \geq 0. \end{cases}
\end{aligned}
$$

The function F_{Y_i} has only one jump, at $y_i = 0$, and $\mathbb{P}(Y_i = 0) = \mathbb{P}(U_i^* \leq 0)$. Therefore, Y_i is a random variable that is neither discrete nor absolutely continuous, but a mixture of the two types. In other words,

$$
F_{Y_i}(y_i;\boldsymbol{\theta}) = \mathbb{P}(U_i^* \leq 0)F_{\text{d}}(y_i) + \mathbb{P}(U_i^* > 0)F_{\text{ac}}(y_i),
$$

where $F_{\text{d}}(y_i) = \mathbb{1}_{[0,+\infty)}(y_i)$ and $F_{\text{ac}}(y_i) = \mathbb{P}(Y_i^* \leq y_i | U_i^* > 0)$. Hence, the PDF of Y_i is given by

$$
\begin{aligned}
f_{Y_i}(y_i;\boldsymbol{\theta}) &= \mathbb{P}(U_i^* \leq 0)\delta_0(y_i) + \mathbb{P}(U_i^* > 0) f_{Y_i^*|U_i^*>0}(y_i;\boldsymbol{\theta}) \\
&= (\mathbb{P}(U_i^* \leq 0))^{1-u_i}(\mathbb{P}(U_i^* > 0))^{u_i}(f_{Y_i^*|U_i^*>0}(y_i;\boldsymbol{\theta}))^{u_i} \qquad (1.3) \\
&= \mathbb{P}(U_i = u_i)(f_{Y_i^*|U_i^*>0}(y_i;\boldsymbol{\theta}))^{u_i}, \quad u_i = 0, 1,
\end{aligned}
$$

wherein $\mathbb{P}(U_i = 0) = 1 - \mathbb{P}(U_i = 1) = \mathbb{P}(U_i^* \leq 0)$ for $i = 1, \ldots, n$, and δ_0 is the Dirac delta function. That is, the density of Y_i is composed of a discrete component described by the probit model $\mathbb{P}(U_i = u_i) = (\mathbb{P}(U_i^* \leq 0))^{1-u_i}(\mathbb{P}(U_i^* > 0))^{u_i}$, for $u_i = 0, 1$, and a continuous part given by the conditional PDF $f_{Y_i^*|U_i^*>0}(y_i;\boldsymbol{\theta})$.

Based on Arellano Valle, Branco, and Genton (2006), we know that if $(Y_i^*, U_i^*)^\top \sim \text{BSY}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g_c)$,

then the PDF of $Y_i^*|U_i^* > 0$ is given by

$$f_{Y_i^*|U_i^*>0}(y_i;\boldsymbol{\theta}) = \frac{1}{\sigma_i}\, f_{Z_{1i}}\left(\frac{y_i-\mu_{1i}}{\sigma_i}\right) \frac{G_i\Big(\frac{1}{\sqrt{1-\rho_i^2}}\,\mu_{2i} + \frac{\rho_i}{\sqrt{1-\rho_i^2}}\,\big(\frac{y_i-\mu_{1i}}{\sigma_i}\big)\Big)}{H_i(\mu_{2i})}, \tag{1.4}$$

where

$$G_i(x) = \int_{-x}^{\infty} f_{Z_{2i}|Z_{1i}}\Big(w_i \,\Big|\, \frac{y_i-\mu_{1i}}{\sigma_i}\Big)\,\mathrm{d}w_i, \quad H_i(x) = \int_{-x}^{\infty} f_{\rho_i Z_{1i}+\sqrt{1-\rho_i^2}\,Z_{2i}}(u_i)\,\mathrm{d}u_i,$$

with $Z_{1i} = RDV_{1i}$ and $Z_{2i} = R\sqrt{1-D^2}V_{2i}$. Here, $f_{Z_{1i},Z_{2i}}$ is the joint PDF of Z_{1i} and Z_{2i}, and f_X denotes the PDF corresponding to a random variable X. Moreover, the random variables V_{1i}, V_{2i}, R, and D are mutually independent and $\mathbb{P}(V_{ki} = -1) = \mathbb{P}(V_{ki} = 1) = 1/2$, $k = 1, 2$. The random variable D is positive and has PDF $f_D(d) = \frac{2}{\pi\sqrt{1-d^2}}$, $d \in (0, 1)$. On the other hand, the random variable R is positive and is called the generator of the random vector $(Y_i^*, U_i^*)^\top$. Particularly, R has PDF given by $f_R(r) = \frac{2rg_c(r^2)}{\int_0^\infty g_c(u)\,\mathrm{d}u}$, $r > 0$, where g_c is the density generator in (1.1).

By combining Equations (1.3) and (1.4), the following formula for the PDF of Y_i is valid:

$$f_{Y_i}(y_i;\boldsymbol{\theta}) = (1-H_i(\mu_{2i}))^{1-u_i}(H_i(\mu_{2i}))^{u_i}\left[\frac{1}{\sigma_i}\, f_{Z_{1i}}\left(\frac{y_i-\mu_{1i}}{\sigma_i}\right)\frac{G_i\big(\tau_i+\alpha_i\big(\frac{y_i-\mu_{1i}}{\sigma_i}\big)\big)}{H_i(\mu_{2i})}\right]^{u_i},$$

where $\alpha_i = \rho_i/\sqrt{1-\rho_i^2}$, $\tau_i = \mu_{2i}/\sqrt{1-\rho_i^2}$, $u_i = 1$ if $u_i^* > 0$ and $u_i = 0$ otherwise, $g_1(\mu_{1i}) = \boldsymbol{x}_i^\top\boldsymbol{\beta}$, $g_2(\mu_{2i}) = \boldsymbol{w}_i^\top\boldsymbol{\gamma}$, $h_1(\sigma_i) = \boldsymbol{z}_i^\top\boldsymbol{\lambda}$ and $h_2(\rho_i) = \boldsymbol{v}_i^\top\boldsymbol{\kappa}$.

The log-likelihood of the symmetric generalized Heckman model for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\lambda}^\top, \boldsymbol{\kappa}^\top)^\top$

is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log f_{Y_i}(y_i;\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} u_i\left[-\log(\sigma_i) + \log f_{Z_{1i}}\left(\frac{y_i-\mu_{1i}}{\sigma_i}\right) + \log G_i\left(\tau_i + \alpha_i\left(\frac{y_i-\mu_{1i}}{\sigma_i}\right)\right)\right] \\ &+ \sum_{i=1}^{n}(1-u_i)\log(1-H_i(\mu_{2i})). \end{aligned} \tag{1.5}$$

To obtain the ML estimate of $\boldsymbol{\theta}$, we maximize the log-likelihood function (1.5) by equating the score vector $\dot{\ell}(\boldsymbol{\theta})$ to zero, providing the likelihood equations. They are solved by means of an iterative procedure for non-linear optimization, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method.

The likelihood equations are given by

$$0 = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{u_i}{\sigma_i}\left[-\frac{f'_{Z_{1i}}(\frac{y_i-\mu_{1i}}{\sigma_i})}{f_{Z_{1i}}(\frac{y_i-\mu_{1i}}{\sigma_i})} - \frac{\alpha_i f_{Z_{2i}|Z_{1i}}\big(\alpha_i\,(\frac{y_i-\mu_{1i}}{\sigma_i}) + \tau_i \,\big|\, \frac{y_i-\mu_{1i}}{\sigma_i}\big)}{G_i\big(\alpha_i\,(\frac{y_i-\mu_{1i}}{\sigma_i}) + \tau_i\big)}\right]\frac{\partial \mu_{1i}}{\partial \beta_j};$$

$$0 = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \gamma_r} = \sum_{i=1}^{n}\left[\frac{u_i}{\sqrt{1-\rho_i^2}}\,\frac{f_{Z_{2i}|Z_{1i}}\big(\alpha_i\,(\frac{y_i-\mu_{1i}}{\sigma_i}) + \tau_i \,\big|\, \frac{y_i-\mu_{1i}}{\sigma_i}\big)}{G_i\big(\alpha_i\,(\frac{y_i-\mu_{1i}}{\sigma_i}) + \tau_i\big)} + \frac{f_{\rho_i Z_{1i}+\sqrt{1-\rho_i^2}\,Z_{2i}}(-\mu_{2i})}{1-H_i(\mu_{2i})}\right]\frac{\partial \mu_{2i}}{\partial \gamma_r};$$

$$0 = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \lambda_s} = \sum_{i=1}^{n} \frac{u_i}{\sigma_i}\left[-1 - \frac{(y_i-\mu_{1i})}{\sigma_i}\,\frac{f'_{Z_{1i}}(\frac{y_i-\mu_{1i}}{\sigma_i})}{f_{Z_{1i}}(\frac{y_i-\mu_{1i}}{\sigma_i})} - \alpha_i\,\frac{(y_i-\mu_{1i})}{\sigma_i}\,\frac{f_{Z_{2i}|Z_{1i}}\big(\alpha_i\,(\frac{y_i-\mu_{1i}}{\sigma_i}) + \tau_i \,\big|\, \frac{y_i-\mu_{1i}}{\sigma_i}\big)}{G_i\big(\alpha_i\,(\frac{y_i-\mu_{1i}}{\sigma_i}) + \tau_i\big)}\right]\frac{\partial \sigma_i}{\partial \lambda_s};$$

$$0 = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \kappa_m} = \sum_{i=1}^{n} \frac{u_i}{\sqrt{1-\rho_i^2}}\left[\frac{(\frac{y_i-\mu_{1i}}{\sigma_i})}{1-\rho_i^2} - \mu_{2i}\rho_i\right]\frac{f_{Z_{2i}|Z_{1i}}\big(\alpha_i\,(\frac{y_i-\mu_{1i}}{\sigma_i}) + \tau_i \,\big|\, \frac{y_i-\mu_{1i}}{\sigma_i}\big)}{G_i\big(\alpha_i\,(\frac{y_i-\mu_{1i}}{\sigma_i}) + \tau_i\big)}\,\frac{\partial \rho_i}{\partial \kappa_m};$$

where

$$\frac{\partial \mu_{1i}}{\partial \beta_j} = \frac{x_{ij}}{g_1'(\mu_{1i})}, \quad j = 1, \ldots, k;\ i = 1, \ldots, n;$$
$$\frac{\partial \mu_{2i}}{\partial \gamma_r} = \frac{w_{ir}}{g_2'(\mu_{2i})}, \quad r = 1, \ldots, l;\ i = 1, \ldots, n;$$
$$\frac{\partial \sigma_i}{\partial \lambda_s} = \frac{z_{is}}{h_1'(\sigma_i)}, \quad s = 1, \ldots, p;\ i = 1, \ldots, n;$$
$$\frac{\partial \rho_i}{\partial \kappa_m} = \frac{v_{im}}{h_2'(\rho_i)}, \quad m = 1, \ldots, q;\ i = 1, \ldots, n.$$

1.3 Generalized Heckman-Student-t model

The generalized Heckman-normal model proposed by Bastos, Barreto-Souza, and Genton (2021) is a special case of (1.2) when the underlying distribution is bivariate normal. In this work, we focus on the generalized Heckman-t model, which is based on the bivariate Student-t (Bt) distribution. This distribution is a good alternative in the symmetric family of distributions because it possesses has heavier tails than the bivariate normal distribution. From (1.1), if $\boldsymbol{Y} = (Y_1, Y_2)^\top$ follows a Bt distribution, then the associated PDF is given by

$$f(\boldsymbol{y}; \boldsymbol{\mu}; \boldsymbol{\Sigma}, \nu) = |\boldsymbol{\Sigma}|^{-1/2} \left\{ 1 + \frac{(\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma} (\boldsymbol{y} - \boldsymbol{\mu})}{\nu} \right\}^{-(\nu+2)/2}, \tag{1.6}$$

where ν is the number of degrees of freedom. Here, the density generator of the Bt distribution is given by $g_c(x) = (1 + x/\nu)^{-(\nu+2)/2}$. Therefore, if (Y_i^*, U_i^*) follow a Bt distribution, then, by Equation (1.4), the PDF of $Y_i^* | U_i^* > 0$ is written as

$$f_{Y_i^*|U_i^*>0}(y_i; \mu_{1i}, \sigma_i^2, \alpha_i, \tau_i, \nu) = \frac{1}{\sigma_i} f_\nu\left(\frac{y_i - \mu_{1i}}{\sigma_i}\right) \frac{F_{\nu+1}\left(\sqrt{\frac{(\nu+1)}{\nu + (\frac{y_i - \mu_{1i}}{\sigma_i})^2}} \left(\tau_i + \alpha_i (\frac{y_i - \mu_{1i}}{\sigma_i})\right)\right)}{F_\nu\left(\tau_i / \sqrt{1 + \alpha_i^2}\right)},$$

where f_ν and F_ν are the PDF and CDF, respectively, of a univariate Student-t distribution with ν degrees of freedom, $\alpha_i = \rho_i / \sqrt{1 - \rho_i^2}$ and $\tau_i = \mu_{2i} / \sqrt{1 - \rho_i^2}$. The log-likelihood for $\boldsymbol{\theta} =$

$(\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\lambda}^\top, \boldsymbol{\kappa}^\top, \nu)^\top$ is given by

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log f_{Y_i}(y_i; \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log \left\{ (F_\nu(-\mu_{2i}))^{1-u_i} (F_\nu(\mu_{2i}))^{u_i} (f_{Y_i^*|U_i^*>0}(y_i; \mu_{1i}, \sigma_i^2, \alpha_i, \tau_i, \nu))^{u_i} \right\} \\
&= \sum_{i=1}^{n} u_i \left[-\log(\sigma_i) + \log f_\nu \left(\frac{y_i - \mu_{1i}}{\sigma_i} \right) + \log F_{\nu+1} \left(\sqrt{\frac{(\nu+1)}{\nu + (\frac{y_i-\mu_{1i}}{\sigma_i})^2}} \left(\tau_i + \alpha_i \left(\frac{y_i - \mu_{1i}}{\sigma_i} \right) \right) \right) \right] \\
&+ \sum_{i=1}^{n} (1 - u_i) \log F_\nu(-\mu_{2i}), \qquad (1.7)
\end{aligned}$$

where $u_i = 1$ if $u_i^* > 0$ and $u_i = 0$ otherwise, μ_{1i}, μ_{2i}, σ_i and ρ_i are as in (1.2). The ML estimate of $\boldsymbol{\theta}$ is obtained by maximizing the log-likelihood function (1.7), that is, by equating the score vector $\dot{\ell}(\boldsymbol{\theta})$ (given in Section 1.2) to zero, providing the likelihood equations. They are solved using an iterative procedure for non-linear optimization, such as the BFGS quasi-Newton method.

1.4 Monte Carlo simulation

In this section, we carry out Monte Carlo simulation studies to evaluate the performance of the ML estimators under the symmetric generalized Heckman model. We focus on the generalized Heckman-t model and consider three different set of true parameter value, which leads to scenarios covering moderate to high censoring percentages. The studies consider simulated data generated from each scenario according to

$$\mu_{1i} = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} \qquad (1.8)$$

$$\mu_{2i} = \gamma_1 + \gamma_2 x_{1i} + \gamma_3 x_{2i} + \gamma_4 x_{3i} \qquad (1.9)$$

$$\log \sigma_i = \lambda_1 + \lambda_2 x_{1i} \tag{1.10}$$

$$\text{arctanh}\, \rho_i = \kappa_1 + \kappa_2 x_{1i} \tag{1.11}$$

for $i = 1, \ldots, n$, x_{1i}, x_{2i} and x_{3i} are covariates obtained from a normal distribution in the interval (0,1). Moreover, the simulation scenarios consider sample size $n \in \{500, 1000, 2000\}$ and $\nu = 4$, with NREP $= 1000$ Monte Carlo replicates for each sample size. In the structure presented in (1.8) - (1.11), μ_{1i} is the primary interest equation, while μ_{2i} represents the selection equation. The R software has been used to do all numerical calculations; see R Core Team (2022).

The performance of the ML estimators are evaluated through the bias and mean squared error (MSE), computed from the Monte Carlo replicas as

$$\widehat{\text{Bias}}(\widehat{\theta}) = \frac{1}{\text{NREP}} \sum_{i=1}^{\text{NREP}} \widehat{\theta}^{(i)} - \theta \quad \text{and} \quad \widehat{\text{MSE}}(\widehat{\theta}) = \frac{1}{\text{NREP}} \sum_{i=1}^{\text{NREP}} (\widehat{\theta}^{(i)} - \theta)^2, \tag{1.12}$$

where θ and $\widehat{\theta}^{(i)}$ are the true parameter value and its respective i-th ML estimate, and NREP is the number of Monte Carlo replicas.

We consider the following sets of true parameter values for the regression structure in (1.8)-(1.11):

- Scenario 1) $\boldsymbol{\beta} = (1.1, 0.7, 0.1)^\top$, $\boldsymbol{\gamma} = (0.9, 0.5, 1.1, 0.6)^\top$, and $\boldsymbol{\lambda} = (-0.4, 0.7)^\top$ and $\boldsymbol{\kappa} = (0.3, 0.5)^\top$.
- Scenario 2) $\boldsymbol{\beta} = (1.0, 0.7, 1.1)^\top$, $\boldsymbol{\gamma} = (0.9, 0.5, 1.1, 0.6)^\top$, $\boldsymbol{\lambda} = (-0.2, 1.2)^\top$, and $\boldsymbol{\kappa} = (0.7, 0.3)^\top$ or $\boldsymbol{\kappa} = (-0.7, 0.3)^\top$.
- Scenario 3) $\boldsymbol{\beta} = (1.1, 0.7, 0.1)^\top$, $\boldsymbol{\gamma} = (0, 0.5, 1.1, 0.6)^\top$, $\boldsymbol{\lambda} = (-0.4, 1.2)^\top$, and $\boldsymbol{\kappa} = (-0.3, -0.3)^\top$ (moderate correlation) or $\boldsymbol{\kappa} = (-0.7, -0.7)^\top$ (strong correlation).

To preserve the parameters and maintain a censoring rate around 50%, in Scenario 1 a threshold greater than zero was used, so $U_i^* > a$, according to Bastos, Barreto-Souza, and

Genton (2021) in general the value of a is zero, as any other value would be absorbed by the intercept, so considering another value does not cause problems for the model. In Scenario 2 the dispersion and correlation parameters were changed and the censoring rate was maintained around 30%. In Scenario 3, the censoring rate around 50% was obtained by changing the parameter the selection equation μ_{2i}.

The ML estimation results for the Scenarios 1), 2) and 3) are presented in Tables 1.1-1.3, respectively, wherein the bias and MSE are all reported. As the ML estimators are consistent and asymptotically normally distributed, we expect the bias and MSE to approach zero as n grows. Moreover, we expect that the performances of the estimates deteriorate as the censoring proportion (%) grows. A look at the results in Tables 1.1-1.3 allows us to conclude that, as the sample size increases, the bias and MSE both decrease, as expected. In addition, the performances of the estimates decrease when the censoring proportion increases.

Table 1.1: Bias and MSE for the indicated ML estimates of the generalized Heckman-t model parameters (Scenario 1).

			Generalized Heckman-t			Generalized Heckman-t		
Parameters		n	Censoring	Bias	MSE	Censoring	Bias	MSE
		500	30.8878	0.0034	0.0053	54.0284	0.0082	0.0148
β_1	1.1	1000	30.9352	0.0019	0.0025	53.421	0.0016	0.0060
		2000	31.0035	0.0021	0.0011	52.8706	0.0058	0.0030
		500	30.8878	0.0037	0.0014	54.0284	0.0060	0.0034
β_2	0.7	1000	30.9352	0.0020	0.0007	53.421	0.0033	0.0013
		2000	31.0035	0.0016	0.0003	52.8706	0.0020	0.0005
		500	30.8878	0.0009	0.002	54.0284	-0.0014	0.0045
β_3	0.1	1000	30.9352	-0.0003	0.0008	53.421	0.0003	0.0017
		2000	31.0035	0.0008	0.0004	52.8706	-0.0022	0.0008
		500	30.8878	0.0163	0.0137	54.0284	-1.0167	1.0443
γ_1	0.9	1000	30.9352	0.0055	0.0065	53.421	-1.0095	1.0246
		2000	31.0035	-0.0008	0.0032	52.8706	-0.9981	0.9980
		500	30.8878	0.0066	0.0091	54.0284	0.0071	0.0077
γ_2	0.5	1000	30.9352	0.0064	0.0044	53.421	0.0026	0.0039
		2000	31.0035	0.0025	0.0021	52.8706	0.0021	0.0019
		500	30.8878	0.0177	0.0194	54.0284	0.0131	0.0170
γ_3	1.1	1000	30.9352	0.0081	0.0085	53.421	0.0090	0.0074
		2000	31.0035	0.0032	0.0041	52.8706	0.0044	0.0038
		500	30.8878	0.0091	0.0110	54.0284	0.0076	0.0089
γ_4	0.6	1000	30.9352	0.0081	0.0085	53.421	0.0043	0.0045
		2000	31.0035	0.0028	0.0026	52.8706	0.0020	0.0020
		500	30.8878	0.0139	0.0564	54.0284	0.0034	0.0467
κ_1	0.3	1000	30.9352	0.0048	0.0048	53.421	0.0080	0.0207
		2000	31.0035	-0.0005	0.0102	52.8706	-0.0029	0.0097
		500	30.8878	0.0589	0.0554	54.0284	0.0416	0.0383
κ_2	0.5	1000	30.9352	0.0054	0.0225	53.421	0.0202	0.0157
		2000	31.0035	0.0120	0.0083	52.8706	0.0136	0.0064
		500	30.8878	0.0011	0.0049	54.0284	-0.0029	0.0075
λ_1	-0.4	1000	30.9352	0.0210	0.018	53.421	0.0009	0.0035
		2000	31.0035	0.0018	0.0011	52.8706	-0.0014	0.0018
		500	30.8878	0.0017	0.0031	54.0284	0.0037	0.0046
λ_2	0.7	1000	30.9352	0.0006	0.0023	53.421	0.0026	0.0021
		2000	31.0035	0.0007	0.0007	52.8706	0.0006	0.0010
		500	30.8878	0.3633	2.3387	54.0284	0.6463	8.4442
ν	4	1000	30.9352	0.1506	0.5565	53.421	0.2221	0.8141
		2000	31.0035	0.0833	0.2377	52.8706	0.0820	0.3078

Table 1.2: Bias and MSE for the indicated ML estimates of the generalized Heckman-t model parameters (Scenario 2).

			Generalized Heckman-t				Generalized Heckman-t	
n	Parameters	Censoring	Bias	MSE	Parameters	Censoring	Bias	MSE
500		31.017	0.0048	0.0045		30.9268	0.0052	0.0039
1000	β_1 1.0	30.9395	0.0031	0.0022	β_1 1.0	30.8644	-0.0002	0.0016
2000		30.9098	0.0013	0.0009		30.9734	0.0005	0.0009
500		31.017	0.0013	0.0009		30.9268	0.0045	0.0008
1000	β_2 0.7	30.9395	-0.0002	0.0046	β_2 0.7	30.8644	0.0010	0.0003
2000		30.9098	0.0002	0.0001		30.9734	0.0006	0.0001
500		31.017	-0.001	0.0011		30.9268	0.0001	0.0007
1000	β_3 1.1	30.9395	0.0015	0.0008	β_3 1.1	30.8644	0.0009	0.0003
2000		30.9098	-0.0007	0.0002		30.9734	0.0003	0.0001
500		31.017	0.0071	0.0141		30.9268	0.0108	0.0153
1000	γ_1 0.9	30.9395	0.0165	0.0814	γ_1 0.9	30.8644	0.0205	0.0555
2000		30.9098	0.0036	0.0031		30.9734	0.0027	0.0034
500		31.017	0.0072	0.0090		30.9268	0.0111	0.0096
1000	γ_2 0.5	30.9395	0.0076	0.0101	γ_2 0.5	30.8644	0.0115	0.0270
2000		30.9098	-0.0005	0.0020		30.9734	0.0016	0.0035
500		31.017	0.013 0	0.0180		30.9268	0.0143	0.0175
1000	γ_3 1.1	30.9395	0.0151	0.0254	γ_3 1.1	30.8644	0.0170	0.0450
2000		30.9098	0.0020	0.0041		30.9734	0.0028	0.0051
500		31.017	0.0067	0.0102		30.9268	-0.0003	0.0094
1000	γ_4 0.6	30.9395	0.0064	0.0195	γ_4 0.6	30.8644	0.0077	0.0139
2000		30.9098	0.001	0.0023		30.9734	0.0023	0.0024
500		31.017	0.0299	0.0560		30.9268	-0.0449	0.0662
1000	κ_1 0.7	30.9395	0.0197	0.0628	κ_1 -0.7	30.8644	-0.0198	0.033
2000		30.9098	0.0031	0.0089		30.9734	-0.0058	0.0218
500		31.017	0.052	0.0702		30.9268	0.0363	0.067
1000	κ_2 0.3	30.9395	0.0197	0.0576	κ_2 0.3	30.8644	0.0141	0.0632
2000		30.9098	0.0097	0.0097		30.9734	0.0108	0.0345
500		31.017	0.0004	0.0053		30.9268	0.0043	0.0054
1000	λ_1 -0.2	30.9395	-0.0018	0.0041	λ_1 -0.2	30.8644	-0.0031	0.0039
2000		30.9098	0.0019	0.0012		30.9734	-0.0004	0.0013
500		31.017	0.0032	0.0038		30.9268	0.0096	0.0035
1000	λ_2 1.2	30.9395	0.0018	0.0020	λ_2 1.2	30.8644	0.0058	0.0016
2000		30.9098	-0.0001	0.0007		30.9734	0.0024	0.0010
500		31.017	0.4404	2.3837		30.9268	0.5096	17.9132
1000	ν 4	30.9395	0.1418	0.6321	ν 4	30.8644	0.1619	0.6922
2000		30.9098	0.1234	0.2899		30.9734	0.0789	0.2731

Table 1.3: Bias and MSE for the indicated ML estimates of the generalized Heckman-t model parameters (Scenario 3).

			Generalized Heckman-t					Generalized Heckman-t	
n	Parameters	Censoring	Bias	MSE	Parameters	Censoring	Bias	MSE	
500		49.8604	-0.0037	0.0094		49.9824	-0.0109	0.0072	
1000	β_1 1.1	49.9469	-0.0008	0.0035	β_1 1.1	49.9075	-0.0049	0.003	
2000		50.0265	-0.0012	0.0017		49.925	-0.0046	0.0011	
500		49.8604	-0.0025	0.0016		49.9824	-0.0057	0.0014	
1000	β_2 0.7	49.9469	-0.0006	0.0005	β_2 0.7	49.9075	-0.0026	0.0005	
2000		50.0265	-0.0009	0.0002		49.925	-0.0025	0.0002	
500		49.8604	0.0001	0.0014		49.9824	0.0017	0.0011	
1000	β_3 0.1	49.9469	<0.0001	0.0005	β_3 0.1	49.9075	0.0004	0.0004	
2000		50.0265	0.0005	0.0002		49.925	-0.0002	0.0002	
500		49.8604	0.0025	0.0065		49.9824	0.0012	0.0063	
1000	γ_1 0	49.9469	0.0036	0.003	γ_1 0	49.9075	0.0029	0.004	
2000		50.0265	-0.0023	0.0015		49.925	0.0022	0.0014	
500		49.8604	0.0067	0.0078		49.9824	0.0051	0.0071	
1000	γ_2 0.5	49.9469	0.0061	0.0035	γ_2 0.5	49.9075	0.0036	0.0051	
2000		50.0265	0.0009	0.0019		49.925	0.0003	0.0015	
500		49.8604	0.009	0.0161		49.9824	0.0095	0.0155	
1000	γ_3 1.1	49.9469	0.0047	0.008	γ_3 1.1	49.9075	0.0116	0.0185	
2000		50.0265	0.0024	0.0037		49.925	0.0046	0.0035	
500		49.8604	0.0010	0.0091		49.9824	0.0048	0.0083	
1000	γ_4 0.6	49.9469	0.0014	0.0042	γ_4 0.6	49.9075	0.0056	0.0086	
2000		50.0265	-0.001	0.0023		49.925	0.0038	0.002	
500		49.8604	-0.0105	0.04		49.9824	-0.0166	0.0437	
1000	κ_1 -0.3	49.9469	-0.0107	0.0173	κ_1 -0.7	49.9075	-0.0061	0.0183	
2000		50.0265	-0.0015	0.0085		49.925	-0.0034	0.0078	
500		49.8604	-0.0302	0.0394		49.9824	-0.0831	0.0609	
1000	κ_2 -0.3	49.9469	-0.0189	0.0163	κ_2 -0.7	49.9075	-0.0317	0.0195	
2000		50.0265	-0.0053	0.007		49.925	-0.0214	0.0086	
500		49.8604	0.0032	0.0067		49.9824	-0.0041	0.0072	
1000	λ_1 -0.4	49.9469	-0.0003	0.0035	λ_1 -0.4	49.9075	-0.006	0.0034	
2000		50.0265	0.0005	0.0016		49.925	-0.0033	0.0018	
500		49.8604	0.0062	0.004		49.9824	0.0039	0.0038	
1000	λ_2 1.2	49.9469	0.0057	0.0019	λ_2 1.2	49.9075	0.0029	0.0018	
2000		50.0265	0.0026	0.0009		49.925	0.0016	0.0008	
500		49.8604	0.5755	3.9653		49.9824	0.546	8.083	
1000	ν 4	49.9469	0.2316	0.8686	ν 4	49.9075	0.1677	0.793	
2000		50.0265	0.1139	0.3197		49.925	0.0841	0.4107	

1.5 Application to real data

In this section, two real data sets, corresponding to outpatient expense and investments in education, are analyzed. The outpatient expense data data set has already been analyzed in the literature by Heckman models Marchenko and Genton (2012), whereas the education investment data in education is new and is analyzed for the first time here.

1.5.1 Outpatient expense

In this subsection, a real data set corresponding to outpatient expense from the Medical Expenditure Panel Survey 2001 (MEPS) database is used to illustrate the proposed methodology. This data set has information about the cost and provision of outpatient services, and is the most complete coverage about health insurance in the United States, according to the Agency for Healthcare Research and Quality (AHRQ).

The MEPS data set contains information collected from 3328 individuals between 21 and 64 years. The variable of interest is the expenditure on medical services in the logarithm scale ($Y_i^* = lnambx$), while the latent variable ($U_i^* = dambexp$) is the willingness of the individual to spend; $U_i = \mathcal{I}_{\{U_i^*>0\}i}$ corresponds the decision of the individual to spend. It was verified that 526 (15.8%) of the outpatient costs are identified as zero (censored). The covariates considered in the are: age is the age measured in tens of years; fem is a dummy variable that assumed value 1 for women and 0 for men; $educ$ is the years of education; $blhisp$ is a dummy variable for ethnicity (1 for black or Hispanic and 0 if non-black and non-Hispanic); $totcr$ is the total number of chronic diseases; ins is the insurance status; and $income$ denotes the individual income.

Table 1.4 reports descriptive statistics of the observed medical expenditures, including the minimum, mean, median, maximum, standard deviation (SD), coefficient of variation (CV), coefficient of skewness (CS) and coefficient of (excess) kurtosis (CK) values. From this table, we note the following: the mean is almost equal to the median; a very small negative skewness

value; and a very small kurtosis value. The symmetric nature of the data is confirmed by the histogram shown in Figure Figure 1.1(a). The boxplot shown in Figure 1.1(b) indicates some potential outliers. Therefore, we observe that a symmetric distribution is a reasonable assumption, more specifically a Student-t model, since since we have to accommodate outliers.

Table 1.4: Summary statistics for the medical expenditure data.

minimum	Mean	Median	maximum	SD	CV	CS	CK	n
0.693	6.557	6.659	10.819	1.406	21.434%	-0.315	0.006	2801

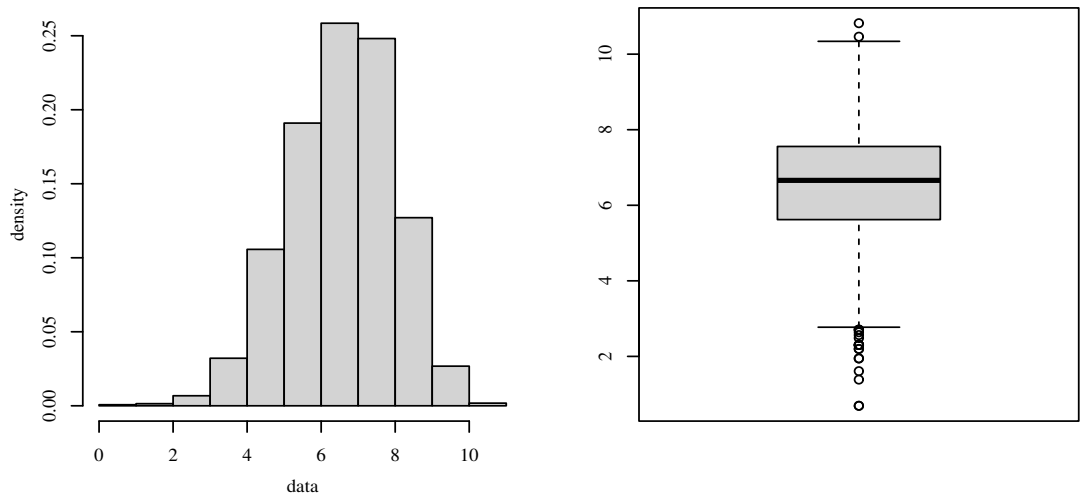

Figure 1.1: Histogram (a) and boxplots (b) for the medical expenditure data.

We then analyze the medical expenditure data using the generalized Heckman-t model, expressed as

$$lnambx = \beta_0 + \beta_1\, age_i + \beta_2\, fem_i + \beta_3\, educ_i + \beta_4\, blhisp_i + \beta_5\, totchr_i + \beta_6\, ins_i, \quad (1.13)$$

$$dambexp = \gamma_0 + \gamma_1\, age_i + \gamma_2\, fem_i + \gamma_3\, educ_i + \gamma_4\, blhisp_i + \gamma_5\, totchr_i + \gamma_6\, ins_i + \gamma_7\, income_i, \quad (1.14)$$

$$\log \sigma_i = \lambda_0 + \lambda_1\, age_i + \lambda_2\, totchr_i + \lambda_3\, ins_i, \quad (1.15)$$

$$\text{arctanh}\, \rho_i = \kappa_0 + \kappa_1\, fem_i + \kappa_2\, totchr_i. \quad (1.16)$$

We initially compare the adjustments of the generalized Heckman-t (GHt) model, in terms of Akaike (AIC) and Bayesian information (BIC), with the adjustments of the classical Heckman-normal (CHN) (Heckman, 1976) and generalized Heckman-normal (GHN) (Bastos, Barreto-Souza, and Genton, 2021) models; see Table 1.5. The AIC and BIC values reveal that the GHt model provides the best adjustment, followed by the GHN model.

Table 1.5: AIC and BIC of the indicated Heckman models.

	CHN	GHN	GHt
AIC	11706.44	11660.29	11635.05
BIC	11810.31	11794.71	11775.59

Table 1.6 presents the estimation results of the GHN and GHt models. From this table, we observe the following results for both models: the parameters associated with the covariates $totchr$ and ins that model the dispersion are significant, in both models, indicating the presence of heteroscedasticity in the data, which justifies biased estimates in the estimators.

Statistical evidence displayed from Table 1.6 is sufficient to reject the absence of selection bias, when applying the test of Wald to evaluate $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$, we observed sufficient evidence to reject H_0 at the 5% significance level. Correlation parameter are significant to covariates fem and $totchr$.

In Table 1.6, the positive estimated parameters indicate an increase in the response variable, the interpretation of parameters is related to the log of outpatient expenses. In the outcome equation $age, fem, totchr$ are significant at any level and $blhisp$ is significant in the level at 5%, $educ$ is not significant and ins is not significant in the level at 5%. We can, for example, affirm, based on the result of this model, that keeping the other parameters fixed, changing a unit in age represents an increase of $\exp(0.1838) = 1.2018$ and $\exp(0.1895) = 1.2086$, i.e 20.18% and 20.86% of increase in ambulatory expenses for the models GHN and GHt respectively. In a similar way, the other parameters of the model can be interpreted, so that keeping the other parameters fixed, changing a unit in $totchr$ represents an increase of $\exp(0.4306) = 1.5383$ and $\exp(0.4464) = 1.5627$, i.e 53.83% and 56.27% of increase in ambulatory expenses for the

models GHN and GHt respectively.

To the selection equation, in the Table 1.6, the covariates $age, fem, educ, blhisp, totchr, ins$ are significant at any level and $income$ is significant in the level at 5%. The interpretation which representing willingness to spend with health, is made by the odds ratio. The chances of investing in medical expenses increase by $\exp(0.0931) = 1.0976$, i.e, 9% when the age increase one unit, and considering the model GHt. For the other parameters the analyzes are similar.

Table 1.6: Estimation results of the GHN and GHt models.

Probit selection equation								
Variables	Estimates		Std. Error		t Value		p-value	
	GHN	GHt	GHN	GHt	GHN	GHt	GHN	GHt
$(Intercept)$	-0.5903	-0.6406	0.1867	0.2014	-3.1620	-3.1800	<0.001	<0.001
age	0.0863	0.0930	0.0264	0.0288	3.2610	3.2230	<0.001	<0.001
fem	0.6299	0.7087	0.0597	0.0681	1.0543	1.0395	<0.001	<0.001
$educ$	0.0569	0.0590	0.0114	0.0122	4.9830	4.8110	<0.001	<0.001
$blhisp$	-0.3368	-0.3726	0.0596	0.0647	-5.6430	-5.7590	<0.001	<0.001
$totchr$	0.7585	0.8728	0.0686	0.0858	1.1043	1.0168	<0.001	<0.001
ins	0.1727	0.1863	0.0611	0.0665	2.8240	2.7990	<0.001	<0.001
$income$	0.0022	0.0025	0.0012	0.0013	1.8380	1.8750	0.0661.	0.061.
Outcome equation								
Variables	Estimates		Std. Error		t Value		p-value	
	GHN	GHt	GHN	GHt	GHN	GHt	GHN	GHt
$(Intercept)$	5.7041	5.6078	0.1930	0.1912	2.9553	2.9316	<0.001	<0.001
age	0.1838	0.1895	0.0234	0.0230	7.8460	8.2350	<0.001	<0.001
fem	0.2498	0.2555	0.0587	0.0580	4.2530	4.4000	<0.001	<0.001
$educ$	0.0013	0.0062	0.0101	0.0100	0.1290	0.6240	0.8970	0.5329
$blhisp$	-0.1283	-0.1344	0.0577	0.0569	-2.2210	-2.3630	0.0264*	0.0182*
$totchr$	0.4306	0.4464	0.0305	0.0297	1.4115	1.5002	<0.001	<0.001
ins	-0.1027	-0.0976	0.0513	0.0501	-2.000	-1.9470	0.0456*	0.0516.
Dispersion								
Variables	Estimates		Std. Error		t Value		p-value	
	GHN	GHt	GHN	GHt	GHN	GHt	GHN	GHt
$Intercept)$	0.5081	0.4172	0.0573	0.0643	8.8550	6.4820	<0.001	<0.001
age	-0.0249	-0.0209	0.0125	0.0136	-1.9870	-1.5360	0.0469*	0.1246
$totchr$	-0.1046	-0.1118	0.0191	0.0208	-5.4760	-5.3720	<0.001	<0.001
ins	-0.1070	-0.1117	0.0277	0.0303	-3.8630	-3.6810	<0.001	<0.001
Correlation								
Variables	Estimates		Std. Error		t Value		p-value	
	GHN	GHt	GHN	GHt	GHN	GHt	GHN	GHt
$Intercept$	-0.6475	-0.6051	0.1143	0.1118	-5.666	-5.413	<0.001	<0.001
fem	-0.4040	-0.4220	0.1356	0.1489	-2.978	-2.835	<0.001	<0.001
$totchr$	-0.4379	-0.4999	0.1862	0.2102	-2.351	-2.378	0.0187*	0.0174*
df	-	12.3230	-	2.7570	-	4.4690	-	<0.001

Figure 1.2 displays the quantile versus quantile (QQ) plots of the martingale-type (MT) residuals for the GHN and GHt models. This residual is given by

$$r_i^{\mathrm{MT}} = \mathrm{sign}(r^{\mathrm{M}_i})\sqrt{-2\left(r^{\mathrm{M}_i} + u_i \log(u_i - r^{\mathrm{M}_i})\right)}, \quad i = 1, \ldots, n. \tag{1.17}$$

where $r^{\mathrm{M}_i} = u_i + \log(\widehat{S}(t_i))$, $\widehat{S}(t_i)$ is the fitted survival function, and $u_i = 0$ or 1 indicating that case i is censored or not, respectively; see Therneau, Grambsch, and Fleming (1990). The MT residual is asymptotically standard normal, if the model is correctly specified whatever the specification of the model is. From Figure 1.2, we see clearly that the GHt model provides better fit than GHN model.

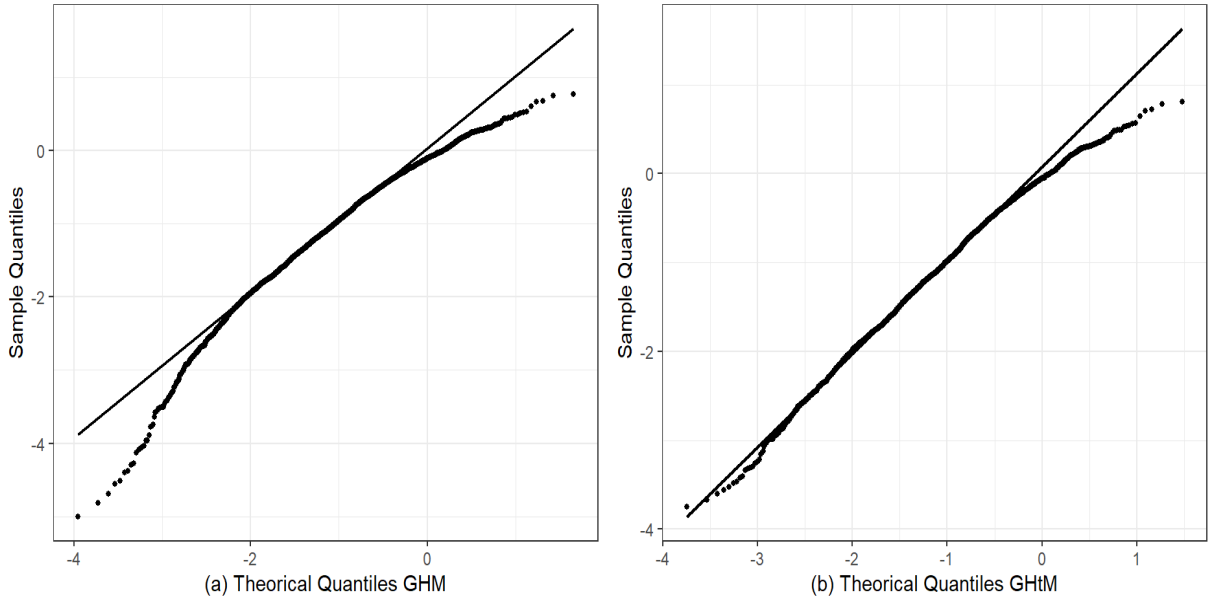

Figure 1.2: QQ plot for the MT residuals for the GHN and GHt models.

1.5.2 Investments in education

The Investments in Education database (IE 2018) made by municipalities goes back to the Fund for Maintenance and Development of Basic Education and Valuing Education Professionals (FUNDEB) . The resources correspond to the set redistributed of 27 federative units (26 states and 1 from the Federal District), which are redistributed according to the number of students who are enrolled in its basic education network. This rule is established to school census data from the previous year (example: 2018 resources were based on the number of students from 2017). This method helps to better distribute resources across the country as it takes into account account the size education networks.

IE database consists in data of two states Sao Paulo (SP) and Minas Gerais (MG) and your respective municipalities. The variable of interest, education investments [1] with 1503 observations, in which 102 (7%) values are not observed and are identified as zero for the models. The explanatory variables are: $income$ [3] represents per capita income collected by the municipality; gnp [3] is the Gross National Product ; $distribute$ [2] is an dummy variable represents Financial Compensation for the Exploration of Minerals (CFEM), this resource must be destined for investments in the areas of health, education and infrastructure for the community, 0 - distributed, 1 - not distributed; sp is an indicator variable for state (sp receives value 1); $enrollment$ [4] is the school census enrollment numbers.

As in the previous study, the response variable investment in education is in the logarithm scale $Y_i^* = lninvest$ represents the logarithm of the response variable, the latent variable ($U_i^* = dinvest$) denotes the willingness of the ith municipality to invested; $U_i = \mathcal{I}_{\{U_i^*>0\}i}$ corresponds the decision the ith municipality to invested.

[1] https://repositorio.shinyapps.io/plataforma_de_dados_municipais
[2] https://dados.gov.br/dataset/sistema-arrecadacao
[3] http://www.ipeadata.gov.br/Default.aspx
[3] https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=resultados
[4] https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar

The descriptive statistics for the investments in education are presents in Table 1.7. From this table, we note the following: the mean is almost equal to the median; a very small negative skewness value, and a high kurtosis value. The symmetric nature of the data is confirmed by the histogram shown in Figure Figure 1.3(a). The boxplot shown in Figure 1.3(b) indicates potential outliers. To the data education investments, as well in the expenditure on medical services, symmetric distribution is a reasonable assumption, more specifically a Student-t model, since since we have to accommodate outliers.

Table 1.7: Summary statistics for the education investment data.

minimum	Mean	Median	maximum	SD	CV	CS	CK	n
4.271	12.212	12.575	18.511	1.903	15.587%	-0.5864	3.70	1401

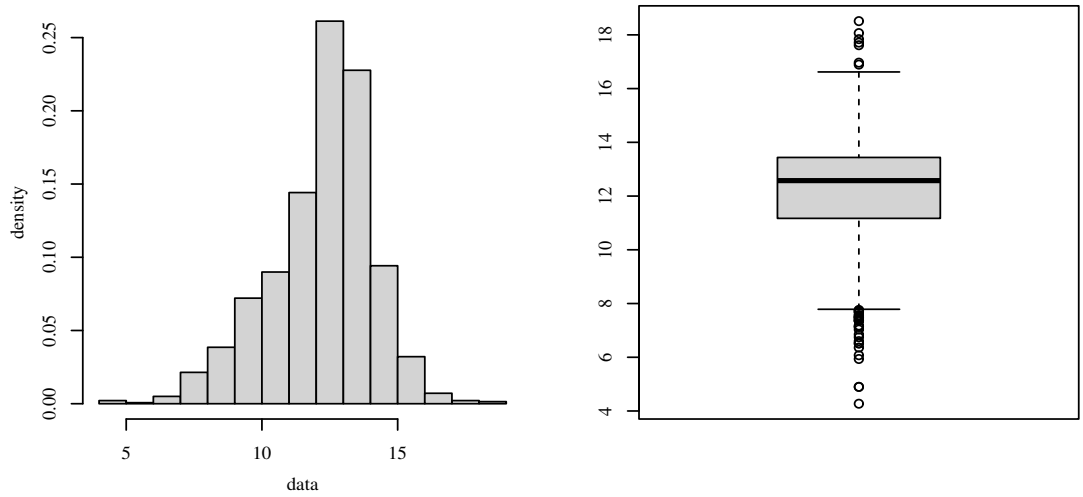

Figure 1.3: Histogram (a) and boxplots (b) for the education investments data.

The equation in model about de variables in study

$$dinvest = \beta_0 + \beta_1\, income_i + \beta_2\, sp_i + \beta_3\, enrollment_i \tag{1.18}$$

$$lninvest = \gamma_0 + \gamma_1\, income_i + \gamma_2\, sp_i + \gamma_3\, enrollment_i + \gamma_4\, gnp_i \tag{1.19}$$

$$\log \sigma_i = \lambda_0 + \lambda_1\, income_i + \lambda_2\, sp_i \tag{1.20}$$

$$\text{arctanh}\,\rho_i = \kappa_0 + \kappa_1\, income_i + \kappa_2\, distribute_i \tag{1.21}$$

Was also compared adjustments of the GHt model, in terms of AIC and BIC, with the adjustments of the CHN (Heckman, 1976) and GHN (Bastos, Barreto-Souza, and Genton, 2021) models; see Table 1.8. The AIC and BIC values reveal that the GHt model provides the best adjustment, followed by the GHN model.

Table 1.8: AIC and BIC of the indicated Heckman models.

	CHN	GHN	GHt
AIC	8269.058	5873.579	5719.590
BIC	8306.147	5953.055	5804.365

In Table 1.9 the parameters associated with the covariates $income$ and sp shows that the dispersion is significant, in both models, indicating the presence of heteroscedasticity. Correlation parameter are significant in the level at 5% to covariate $income$ in the GHN and GHt models, while the sp covariate is significant to any level in the GHt model.

In the outcome equation $income, sp, enrollment$ and gnp are significant at any level to the GHN and GHt models. For the interpretation of parameters, keeping the other covariates fixed, change a unit in $income$ represents an reduce of $\exp(-0.2683) = 0.7647$ and $\exp(-0.2167) = 0.8052$, i.e 23.53% and 19.48% of reduction of investments in education for the GHN and GHt models, respectively. For the parameter gnp, change a unit represents an increase of $\exp(0.0202) = 1.020$ and $\exp(0.0151) = 1.0152$, i.e 2% and 1.5% of increase of investments in education expenses for the models GHN and GHt, respectively. The expressive increase in the variable investment in education occurs with the variable state, here sp has an average investment in education approximately $\exp(1.0063) = 2.7355$ ($\exp(1.0714) = 2.9195$) 3 (three) times higher than the state of mg for the model GHt (GHN).

In the selection equation, which representing willingness to spend with education, the GHt model presented all significant covariates, while the GHN model only $enrollment$ is significant. The interpretation is made by the odds ratio, for example the chances of investing in education increase by $\exp(0.3674) = 1.4439$, i.e, 44% when the state is sp, keeping other vari-

ables constant and considering the estimated parameter with significant p-value GHt. Note the negative sign of the $income$ covariate here, for each unit increase the chances of investing in education decrease by 5%, keeping the other covariates constants.

Table 1.9: Estimation results of the GHN and GHt models..

Probit selection equation								
Variables	Estimate		Std. Error		t Value		p-Value	
	GHN	GHt	GHN	GHt	GHN	GHt	GHN	GHt
$(intercept)$	1.0523	1.4222	0.1013	0.0955	10.3870	14.8880	<0.01	<0.001
$income$	-0.0058	-0.0530	0.0223	0.0164	-0.2630	-3.2330	0.7920	<0.001
sp	0.1065	0.3674	0.0990	0.1032	1.0750	3.5600	0.2820	<0.001
$enrollment$	0.0311	0.0537	0.0032	0.0026	9.6380	20.6150	<0.001	<0.001
Outcome equation								
Variables	Estimate		Std. Error		t Value		p-Value	
	GHN	GHt	GHN	GHt	GHN	GHt	GHN	GHt
$(intercept)$	12.4789	12.4058	0.1252	0.1208	99.6480	102.6960	<0.001	<0.001
$income$	-0.2683	-0.2167	0.0332	0.0274	-8.0640	-7.9030	<0.001	<0.001
sp	1.0714	1.0063	0.1003	0.0881	10.6810	11.4190	<0.001	<0.001
$enrollment$	0.0031	0.0226	0.0005	0.0025	5.6670	8.8410	<0.001	<0.001
gnp	0.0202	0.0151	0.0023	0.0005	8.4630	27.4250	<0.001	<0.001
Dispersion								
Variables	Estimate		Std. Error		t Value		p-Value	
	GHN	GHt	GHN	GHt	GHN	GHt	GHN	GHt
$(intercept)$	0.5921	0.2947	0.0503	0.0623	11.7560	4.7250	<0.001	<0.001
$income$	0.0382	0.0546	0.0130	0.0146	2.9330	3.7180	<0.001	<0.001
sp	-0.3142	-0.4745	0.0416	0.0534	-7.5470	-8.8800	<0.001	<0.001
Correlation								
Variables	Estimate		Std. Error		t Value		p-Value	
	GHN	GHt	GHN	GHt	GHN	GHt	GHN	GHt
$(intercept)$	-3.7263	-6.7669	0.5463	0.9682	-6.8210	-6.9890	<0.001	<0.001
$income$	0.1686	0.2274	0.0839	0.1161	2.0080	-6.9890	0.0448*	0.050.
$distribute$	0.8298	3.1118	0.4411	0.6867	1.8810	1.9580	0.0602.	<0.001
df	-	3.6240	-	0.4320	-	8.3900	-	<0.001

The Figure 1.4 displays the QQ plots of the MT residuals. This figure indicates that the MT residuals in the GHt model shows better agreement with the reference distribution.

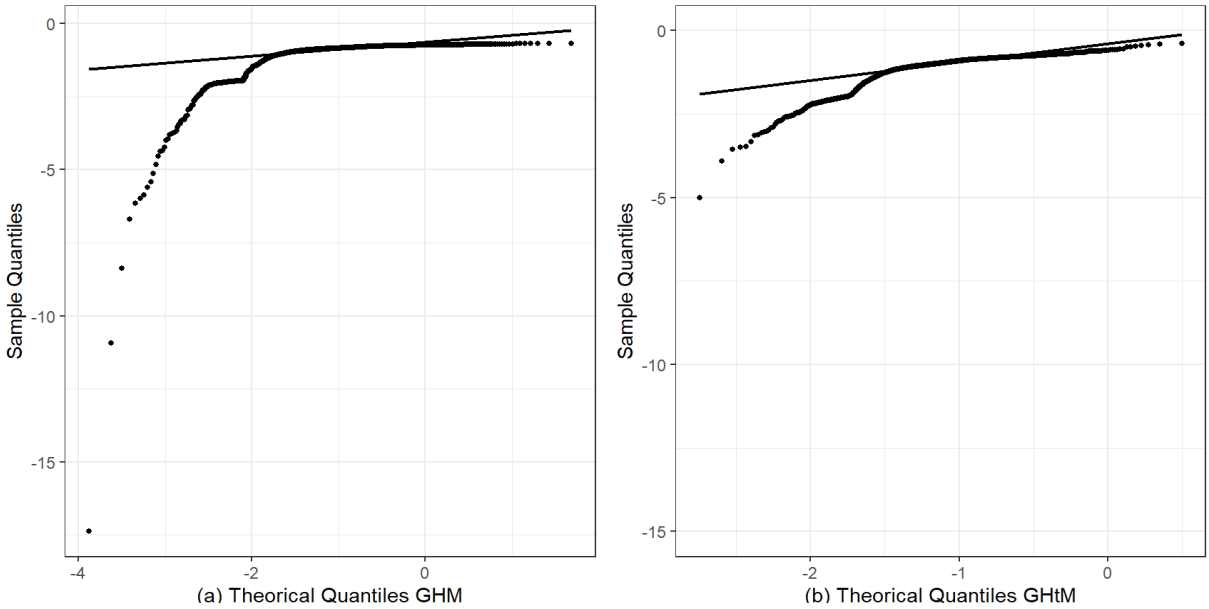

Figure 1.4: QQ plot for the MT residuals for the GHN and GHt models.

1.6 Concluding Remarks

In this paper, a class of Heckman sample selection models were proposed based symmetric distributions. In such models, covariates were added to the dispersion and correlation parameters, allowing the accommodation of heteroscedasticity and a varying sample selection bias, respectively. A Monte Carlo simulation study has showed good results of the parameter estimation method. We have considered high/low censoring rates and the presence of strong/weak correlation. We have applied the proposed model along with some other two existing models to two data sets corresponding to outpatient expense and investments in education. The applications favored the use of the proposed generalized Heckman-t model over the classical Heckman-normal and generalized Heckman-normal models. As part of future research, it will be of interest to propose sample selection models based on skew-symmetric distributions. Furthermore, the behavior of the Wald, score, likelihood ratio and gradient tests can be investigated. Work on these problems is currently in progress and we hope to report these findings in future.

References

Ahn, H. and Powell, J. L. (1993). "Semiparametric estimation of censored selection models with a nonparametric selection mechanism." *Journal of Econometrics*, 58:3–29.

Arellano Valle, R. B., Branco, M. D., and Genton, M. G. (2006). "A Unified View on Skewed Distributions Arising From Selections". *The Canadian Journal of Statistics* 34, pp. 581–601.

Bastos, F. S. and Barreto-Souza (2020). "Birnbaum-Saunders sample selection model". *Journal of Applied Statistics* 48, pp. 1896–1916.

Bastos, F. S., Barreto-Souza, W., and Genton, M. G (2021). "A generalized Heckman model with varying sample election bias and dispersion parameters". *Statistica Sinica*.

Cameron, C. A. and Trivedi, P. K. (2009). "Microeconometrics Using Stata". *TX: Stata press, College Station, revised edition,*

Ding, P. (2014). "Bayesian robust inference of sample selection using selection-t models". *Journal of Multivariate Analysis*, 124:451–464.

Fang, K. T., Kotz, S., and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. London, UK: Chapman and Hall.

Heckman, J. J. (1976). "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models". *Annals of Economic and Social Measurement* 5, pp. 475–492.

— (1979). "Sample selection bias as a specification error". *Econometrica* 47, pp. 153–161.

Lachos, Victor H., Prates, Marcos O., and Dey, Dipak K. (2021). "Heckman selection-t model: Parameter estimation via the EM-algorithm". *Journal of Multivariate Analysis* 184, p. 104737.

Lee, L.F. (1983). "Generalized econometric models with selectivity". *Econometrica*, 51:507–512.

Leung, S. F. and Yu, S. (1996). "On the choice between sample selection and twopart models". *Journal of Econometrics*, 72:197–229.

M. Das, W. K. Newey and Vella, F. (2003). "Nonparametric estimation of sample selection models". *The Review of Economic Studies*, 70:33–58,

M. Zhelonkin, M. G. Genton and Ronchetti, E. (2016). "Robust inference in sample selection models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:805–827.

Manning, W., Duan, N., and Rogers, W. (1987). "Monte carlo evidence on the choice between sample selection and two-part models". *Journal of Econometrics*, 35:59 –82.

Marchenko, Y. V. and Genton, M. G. (2012). "A heckman selection-t model." *Journal of the American Statistical Association* 107, pp. 304–317.

Nelson, F. D. (1984). "Eciency of the two-step estimator for models with endogenous sample selection." Ed. by Journal of Econometrics, 24:181 –196.

Ogundimu, E. O. and Hutton, J. L. (2016). "A sample selection model with skew-normal distribution". *Scandinavian Journal of Statistics*, 43:172–190.

Paarsch, H. J. (1984). "A monte carlo comparison of estimators for censored regression models". Ed. by Journal of Econometrics, 24:197–213.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Stolzenberg, R. M. and Relles, D. A. (1990). "heory testing in a world of constrained research design: The significance of heckman's censored sampling bias correction for nonexperimental research". *Sociological Methods & Research*, 18:395–415.

Therneau, T.M., Grambsch, P.M., and Fleming, T.R. (1990). "Martingale-based residuals for survival models". *Biometrika* 77, pp. 147–160.