Symmetric generalized Heckman models

Helton Saulo, Roberto Vila and Shayane S. Cordeiro
Department of Statistics, Universidade de Brasília, Brasília, Brazil

June 22, 2022

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 et al., 2021). The aim of this work is to propose generalized Heckman sample selection models based on symmetric distributions (Fang et al., 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.

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 et al. (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 Genton, 2012; Ding, 2014; Lachos et al., 2021), the skew-normal distribution (Ogundimu and Hutton, 2016) and the Birnbaum-Saunders distribution (Bastos and Barreto-Souza, 2021). 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 et al., 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 proposed by Bastos et al. (2021) and the Heckman-Student-$t$ model proposed by Marchenko and Genton (2012). We demonstrate that the proposed symmetric generalized Heckman model outperforms the generalized Heckman-normal and Heckman-Student-$t$ models in terms of model fitting, making it a practical and useful model for modelling data with sample selection bias.

The rest of this work proceeds as follows. In Section 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 3, we derive the generalized Heckman-Student-$t$ model, which is a special case of the symmetric generalized Heckman models. In Section 4, we carry out a Monte Carlo simulation study for evaluating the performance of the estimators. In Section 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 6, we provide some concluding remarks.

## 2 Symmetric generalized Heckman models

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

$$
\Sigma = \begin{pmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2
\end{pmatrix}, \tag{2.1}
$$

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

$$
f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}, \Sigma, g_c) = \frac{1}{|\Sigma|^{1/2} Z_{g_c}} g_c \left((\mathbf{y} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})\right), \quad \mathbf{y} \in \mathbb{R}^2, \tag{2.2}
$$

where $|\Sigma| = \sigma_1^2 \sigma_2^2 (1 - \rho^2)$ and $Z_{g_c}$ is a normalization constant so that $f_{\mathbf{Y}}$ is a PDF, that is,

$$
Z_{g_c} = \int_{\mathbb{R}^2} \frac{1}{|\Sigma|^{1/2}} g_c \left((\mathbf{y} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})\right) \mathrm{d}\mathbf{y} = \pi \int_{0}^{\infty} g_c(u) \mathrm{d}u.
$$
The density generator \( g_c \) in (2.2) 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:

\[
\begin{pmatrix} Y_i^* \\ U_i^* \end{pmatrix} \sim \text{BSY} \left( \mu = \begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_i^2 & \rho_i \sigma_i \\ \rho_i \sigma_i & 1 \end{pmatrix}, g_c \right), \quad i = 1, \ldots, n. \tag{2.3}
\]

In the above equation \( \mu_{1i}, \mu_{2i}, \sigma_i \) and \( \rho_i \) are the mean, dispersion and correlation parameters, respectively, with the following regression structure \( g_1(\mu_{1i}) = x_i^\top \beta; \ g_2(\mu_{2i}) = w_i^\top \gamma, \)

\( h_1(\sigma_i) = z_i^\top \lambda \) and \( h_2(\rho_i) = v_i^\top \kappa, \) where \( \beta = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k, \ \gamma = (\gamma_1, \ldots, \gamma_l)^\top \in \mathbb{R}^l, \ \lambda = (\lambda_1, \ldots, \lambda_p)^\top \in \mathbb{R}^p \) and \( \kappa = (\kappa_1, \ldots, \kappa_q)^\top \in \mathbb{R}^q \) are vectors of regression coefficients, \( x_i = (x_{i1}, \ldots, x_{ik})^\top, \ w_i = (w_{i1}, \ldots, w_{iq})^\top, \ z_i = (z_{i1}, \ldots, z_{ip})^\top \) and \( v_i = (v_{i1}, \ldots, v_{iq})^\top \) 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 \( g_i : \mathbb{R} \to \mathbb{R}, \ g_2 : \mathbb{R} \to \mathbb{R}, \ h_1 : \mathbb{R}^+ \to \mathbb{R} \) 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 \( U_i = 1_{\{Y_i^* > 0\}}. \)

Let \( Y_i = Y_i^* U_i \) be the observed outcome, for \( i = 1, \ldots, n. \) Only \( n_1 \) out of \( n \) observations \( Y_i^* \) for which \( U_i^* > 0 \) are observed. This model is known as “Type 2 tobit model” in the econometrics literature. Notice that \( U_i \sim \text{Bernoulli}(P(U_i^* > 0)). \) By using law of total probability, for \( \theta = (\beta^\top, \gamma^\top, \sigma, \rho)^\top, \) the random variable \( Y_i \) has distribution function

\[
F_{Y_i}(y_i; \theta) = P(Y_i \leq y_i | U_i^* > 0)P(U_i^* > 0) + P(Y_i \leq y_i | U_i^* > 0)P(U_i^* \leq 0)
\]

\[
= \begin{cases} P(Y_i^* \leq y_i | U_i^* > 0)P(U_i^* > 0), & \text{if } y_i < 0, \\ P(Y_i^* \leq y_i | U_i^* > 0)P(U_i^* > 0) + P(U_i^* \leq 0), & \text{if } y_i \geq 0. \end{cases}
\]

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

\[
F_{Y_i}(y_i; \theta) = P(U_i^* \leq 0)F_d(y_i) + P(U_i^* > 0)F_{ac}(y_i),
\]

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

\[
f_{Y_i}(y_i; \theta) = P(U_i^* \leq 0)\delta_0(y_i) + P(U_i^* > 0)f_{Y_i^*|U_i^*>0}(y_i; \theta)
\]

\[
= (P(U_i^* \leq 0))^{1-u_i}((P(U_i^* > 0))^{u_i}(f_{Y_i^*|U_i^*>0}(y_i; \theta)))^{u_i}
\]

\[
= P(U_i = u_i)(f_{Y_i^*|U_i^*>0}(y_i; \theta))^{u_i}, \quad u_i = 0, 1,
\]
wherein \( P(U_i = 0) = 1 - P(U_i = 1) = P(U_i^* \leq 0) \) for \( i = 1, \ldots, n \), and \( \delta_0 \) is the Dirac delta function. That is, the density of \( Y_i \) is composed of a discrete component described by the probit model \( P(U_i = u_i) = (P(U_i^* \leq 0))^{1-u_i}(P(U_i^* > 0))^{u_i} \), for \( u_i = 0, 1 \), and a continuous part given by the conditional PDF \( f_{Y_i^*|U_i^*>0}(y_i; \theta) \).

### 2.1 Finding the conditional density of \( Y_i^* \), given that \( U_i^* > 0 \)

In the context of sample selection models, the interest lies in finding the PDF of \( Y_i^*|U_i^*>0 \) given that \((Y_i^*, U_i^*)^\top \) follows the BSY distribution (2.3); see Theorem 2.5.

Before stating and proving the main result (Theorem 2.5) of this section, throughout the paper we will adopt the following notations:

\[
G_i(x) = \int_{-\infty}^{\infty} f_{Z_{2i}|Z_{1i}}(w_i \mid \frac{y_i - \mu_{1i}}{\sigma_i}) \, dw_i \tag{2.5}
\]

and

\[
H_i(x) = \int_{-\infty}^{\infty} f_{\rho_i Z_{1i} + \sqrt{1-\rho_i^2} Z_{2i}}(u_i) \, du_i, \tag{2.6}
\]

where

\[
Z_{1i} = RDV_{1i} \quad \text{and} \quad Z_{2i} = R \sqrt{1-D^2} V_{2i} \tag{2.7}
\]

have the joint PDF \( f_{Z_{1i}, Z_{2i}} \) and \( f_X \) denotes the PDF corresponding to a random variable \( X \).

Here, the random variables \( V_{1i}, V_{2i}, R, \) and \( D \) are mutually independent and \( P(V_{ki} = -1) = P(V_{ki} = 1) = 1/2, \ k = 1, 2 \). The random variable \( D \) in (2.7) is positive and has PDF

\[
f_D(d) = \frac{2}{\pi \sqrt{1-d^2}}, \quad d \in (0, 1). \tag{2.8}
\]

The random variable \( R \) in (2.7) is positive and is called the generator of the random vector \((Y_i^*, U_i^*)^\top \). Moreover, \( R \) has PDF given by

\[
f_R(r) = \frac{2rg_c(r^2)}{\int_0^\infty g_c(u) \, du}, \quad r > 0,
\]

where \( g_c \) is the density generator in (2.2).

**Proposition 2.1.** Let us denote \( S = RD \) and \( T = R \sqrt{1-D^2} \).

1. The PDFs of \( S \) and \( T \) are given by

\[
f_S(v) = f_T(v) = \frac{\int_v^\infty \frac{4g_c(w^2)}{\sqrt{1-w^2}} \, dw}{\pi \int_0^\infty g_c(u) \, du}, \quad v > 0. \tag{2.9}
\]
2. The PDFs of $Z_{1i}$ and $Z_{2i}$ are
\[ f_{Z_{1i}}(z) = f_{Z_{2i}}(z) = \frac{\int_z^{\infty} \frac{2g_c(w^2)}{\sqrt{1-w^2}} \, dw}{\pi \int_0^{\infty} g_c(u) \, du}, \quad -\infty < z < \infty. \]

**Proof.** Using the known formula for the PDF of product of two random variables $X$ and $Y$:
\[ f_{XY}(u) = \int_{-\infty}^{\infty} \frac{1}{|x|} f_{X,Y}(x, \frac{u}{x}) \, dx \]
the proof of (2.8) follows.

The proof of second item follows by combining the law of total probability with (2.8). \qed

**Proposition 2.2.** The random vector $(Z_{1i}, Z_{2i})$ is jointly symmetric about $(0, 0)$. That is, $f_{Z_{1i}, Z_{2i}}(x, y) = f_{Z_{1i}, Z_{2i}}(x, -y) = f_{Z_{1i}, Z_{2i}}(-x, y) = f_{Z_{1i}, Z_{2i}}(-x, -y)$.

**Proof.** Let $S = RD$ and $T = R \sqrt{T-D^2}$. Since $R$ and $D$ are mutually independent, by using change of variables (Jacobian Method), the joint density of $S$ and $T$ is
\[ f_{S,T}(s, t) = \frac{4g_c(s^2 + t^2)}{\pi \int_0^{\infty} g_c(u) \, du}, \quad s, t > 0. \] (2.9)

Moreover, since $V_1 \overset{d}{=} V_2 \sim \text{Bernoulli}(1/2)$ are mutually independent, by law of total probability we get
\[
F_{Z_{1i}, Z_{2i}}(x, y) = \frac{1}{4} \left\{ F_{S,T}(x, y) \mathbb{1}_{[0, \infty) \times [0, \infty)}(x, y) + \mathbb{P}(S \geq -x, T \geq -y) \mathbb{1}_{(-\infty, 0) \times (-\infty, 0)}(x, y) \right. \\
+ \mathbb{P}(S \leq x, T \geq -y) \mathbb{1}_{(0, \infty) \times (-\infty, 0)}(x, y) + \mathbb{P}(S \geq -x, T \leq y) \mathbb{1}_{(-\infty, 0) \times (0, \infty)}(x, y) \left\}, \tag{2.10}
\]
where $F_{X,Y}(\cdot, \cdot)$ denotes the (joint) distribution function of $(X, Y)$. Using the following well-known identity:
\[
\mathbb{P}(a_1 < X \leq b_1, a_2 < Y \leq b_2) = F_{X,Y}(b_1, b_2) - F_{X,Y}(b_1, a_2) - F_{X,Y}(a_1, b_2) + F_{X,Y}(a_1, a_2),
\]
we have
\[
\mathbb{P}(S \geq -x, T \geq -y) = F_{S,T}(\infty, \infty) - F_{S,T}(\infty, -y) - F_{S,T}(-x, \infty) + F_{S,T}(-x, -y), \quad x < 0, y < 0; \\
\mathbb{P}(S \leq x, T \geq -y) = F_{S,T}(x, \infty) - F_{S,T}(x, -y) - F_{S,T}(0, \infty) + F_{S,T}(0, -y), \quad x > 0, y < 0; \\
\mathbb{P}(S \geq -x, T \leq y) = F_{S,T}(\infty, y) - F_{S,T}(\infty, 0) - F_{S,T}(-x, y) + F_{S,T}(-x, 0), \quad x < 0, y > 0.
\]
By replacing the last three identities in (2.10) and then by differentiating \( F_{Z_{1i}, Z_{2i}}(x, y) \) with respect to \( x \) and \( y \), we obtain
\[
\begin{align*}
f_{Z_{1i}, Z_{2i}}(x, y) &= \frac{1}{4} \left\{ f_{S,T}(x, y)1_{(0,\infty)\times(0,\infty)}(x, y) + f_{S,T}(-x, -y)1_{(-\infty,0)\times(-\infty,0)}(x, y) \\
&\quad + f_{S,T}(x, -y)1_{(0,\infty)\times(-\infty,0)}(x, y) + f_{S,T}(-x, y)1_{(-\infty,0)\times(0,\infty)}(x, y) \right\} \\
&= \frac{1}{4} f_{S,T}(x, y)1_{\mathbb{R}\setminus\{0\}\times\mathbb{R}\setminus\{0\}}(x, y) \frac{g_c(x^2 + y^2)}{\pi \int_0^\infty g_c(u) \, du} 1_{\mathbb{R}\setminus\{0\}\times\mathbb{R}\setminus\{0\}}(x, y),
\end{align*}
\]
where in the last line we used the Equation (2.9). Note that the function \( f_{Z_{1i}, Z_{2i}} \) above is not defined on the abscise or ordinate axes (neither at the origin), but this is not relevant because these events have a null probability measure. Therefore, we can say that
\[
f_{Z_{1i}, Z_{2i}}(x, y) = \frac{g_c(x^2 + y^2)}{\pi \int_0^\infty g_c(u) \, du}, \quad -\infty < x, y < \infty. \tag{2.11}
\]

From (2.11) it is clear that the vector \( (Z_{1i}, Z_{2i})^\top \) is jointly symmetric about \((0, 0)\). \(\Box\)

**Proposition 2.3.** 1. The marginal PDFs of \( Z_{1i} \) and \( Z_{1i} \), denoted by \( f_{1i} \) and \( f_{2i} \), respectively, are given by
\[
\begin{align*}
f_{1i}(x) &= \frac{\int_{-\infty}^{\infty} g_c(x^2 + y^2) \, dy}{\pi \int_0^\infty g_c(u) \, du} = f_{Z_{1i}}(x), \quad -\infty < x < \infty, \\
f_{2i}(y) &= \frac{\int_{-\infty}^{\infty} g_c(x^2 + y^2) \, dx}{\pi \int_0^\infty g_c(u) \, du} = f_{Z_{2i}}(y), \quad -\infty < y < \infty,
\end{align*}
\]
where \( f_{Z_{1i}} \) and \( f_{Z_{2i}} \) are given in Proposition 2.2.

2. The random vector \( (Z_{1i}, Z_{2i})^\top \) is marginally symmetric about \((0, 0)\). That is, \( f_{1i}(x) = f_{1i}(-x) \) and \( f_{2i}(y) = f_{2i}(-y) \).

**Proof.** The proof of first item is immediate from the joint density of \((Z_{1i}, Z_{2i})\), given in (2.11), and by Proposition 2.1. The proof of the second item follows from the first one. \(\Box\)

**Proposition 2.4.** The function \( G_i \) in (2.2) satisfies the following identity:
\[
G_i(x) = 1 - G_i(-x) = \int_{-\infty}^{x} f_{Z_{2i}|Z_{1i}}(w_i \mid \frac{y_i - \mu_{1i}}{\sigma_i}) \, dw_i.
\]
That is, \( G_i \) is the conditional CDF of \( Z_{2i} \), given that \( Z_{1i} = (y_i - \mu_{1i})/\sigma_i \). Consequently, \( G_i \) is a distribution symmetric about 0.

**Proof.** The proof immediately follows by applying Proposition 2.2. \(\Box\)
We now proceed to establish the main result of this section.

**Theorem 2.5.** If \((Y^*_i, U^*_i)\) \(\sim\) BSY\((\mu, \Sigma, g_c)\) then the PDF of \(Y^*_i|U^*_i > 0\) is given by

\[
f_{Y^*_i|U^*_i > 0}(y_i; \theta) = \frac{1}{\sigma_i} f_{Z_{1i}} \left( \frac{y_i - \mu_{1i}}{\sigma_i} \right) \frac{G_i \left( \frac{1}{\sqrt{1-\rho_i^2}} \mu_{2i} + \frac{\rho_i}{\sqrt{1-\rho_i^2}} \left( \frac{y_i - \mu_{1i}}{\sigma_i} \right) \right)}{H_i(\mu_{2i})}, \tag{2.12}
\]

where \(G_i, H_i\) and \(Z_{1i}\) are given in Proposition 2.4, (2.6) and (2.7), respectively.

As a by-product of the proof we get that \(P(U^*_i > 0) = H_i(\mu_{2i})\).

**Proof.** Since the random vector \((Y^*_i, U^*_i)\) follows the BSY distribution \((2.3)\), this one admits the stochastic representation \cite{Abdous (2005)}

\[
Y^*_i = \sigma_i Z_{1i} + \mu_{1i},
\]

\[
U^*_i = \rho_i Z_{1i} + \sqrt{1-\rho_i^2} Z_{2i} + \mu_{2i}, \tag{2.13}
\]

where \(Z_{1i}\) and \(Z_{2i}\) are as in \((2.7)\). If \(Y^*_i = y_i\) then \(Z_{1i} = (y_i - \mu_{1i})/\sigma_i\). So, the conditional distribution of \(U^*_i\), given that \(Y^*_i = y_i\) is the same as the distribution of

\[
\rho_i \left( \frac{y_i - \mu_{1i}}{\sigma_i} \right) + \sqrt{1-\rho_i^2} Z_{2i} + \mu_{2i} \mid Y^*_i = y_i.
\]

Consequently, the PDF of \(U^*_i\) given that \(Y^*_i = y_i\) is given by

\[
f_{U^*_i|Y^*_i}(u_i|y_i) = \frac{f_{Z_{1i}, Z_{2i}} \left( \frac{y_i - \mu_{1i}}{\sigma_i}, \frac{1}{\sqrt{1-\rho_i^2}} (u_i - \mu_{2i}) = \frac{\rho_i}{\sqrt{1-\rho_i^2}} \left( \frac{y_i - \mu_{1i}}{\sigma_i} \right) \right)}{\sqrt{1-\rho_i^2} f_{Z_{1i}} \left( \frac{y_i - \mu_{1i}}{\sigma_i} \right)}. \tag{2.14}
\]

Further,

\[
f_{Y^*_i}(y_i) = \frac{1}{\sigma_i} f_{Z_{1i}} \left( \frac{y_i - \mu_{1i}}{\sigma_i} \right). \tag{2.15}
\]

By using the identity

\[
f_{Y^*_i|U^*_i > 0}(y_i; \theta) = f_{Y^*_i}(y_i) \frac{\int_{0}^{\infty} f_{U^*_i|Y^*_i}(u_i|y_i) \, du_i}{P(U^*_i > 0)} = f_{Y^*_i}(y_i) \frac{\int_{0}^{\infty} f_{U^*_i|Y^*_i}(u_i|y_i) \, du_i}{\int_{0}^{\infty} f_{U^*_i}(u_i) \, du_i},
\]

7
and by employing identities (2.14) and (2.15), we get

\[ f_{Y_i^*|U_i^*>0}(y_i; \theta) = \frac{1}{\sigma_i} f_{Z_{i1}} \left( \frac{y_i - \mu_{1i}}{\sigma_i} \right) \int_0^\infty f_{Z_{1i}, Z_{2i}} \left( \frac{y_i - \mu_{1i}}{\sigma_i}, \sqrt{1 - \rho_i^2} \left( u_i - \mu_{2i} \right) \right) \frac{1}{\sqrt{1 - \rho_i^2}} f_{Z_{1i}} \left( \frac{y_i - \mu_{1i}}{\sigma_i} \right) du_i \]

Finally, by using the notations of \( G_i \) and \( H_i \) given in (2.5) and (2.6), respectively, from the above identity and from Proposition 2.4 the proof follows.

**Corollary 2.6** (Gaussian density generator). If \((Y_i^*, U_i^*)^T \sim \text{BSY}(\mu, \Sigma, g_c)\), where \( g_c(x) = \exp(-x/2) \) is the density generator of the bivariate normal distribution, then the PDF of \( Y_i^*|U_i^*>0 \) is given by

\[ f_{Y_i^*|U_i^*>0}(y_i; \theta) = \frac{1}{\sigma_i} \phi \left( \frac{y_i - \mu_{1i}}{\sigma_i} \right) \Phi \left( \frac{1}{\sqrt{1 - \rho_i^2}} \mu_{2i} + \frac{\rho_i}{\sqrt{1 - \rho_i^2}} \frac{y_i - \mu_{1i}}{\sigma_i} \right) \frac{\Phi(\mu_{2i})}{\Phi(\mu_{2i})}, \]

wherein \( \phi \) and \( \Phi \) denote the PDF and CDF of the standard normal distribution, respectively.

**Proof.** If \((Y_i^*, U_i^*)^T \) follows the bivariate normal distribution then there exist independent standard normal random variables \( Z_{i1} \) and \( Z_{2i} \) such that a stochastic representation of type (2.13) is satisfied. Therefore, \( Z_{2i|Z_{1i}} = \) and \( \rho_i Z_{1i} + \sqrt{1 - \rho_i^2} Z_{2i} \) are distributed according to the standard normal distribution. Hence, \( G_i \) and \( H_i \), given in Proposition 2.4 and Item 2.6, respectively, are written as:

\[ G_i(x) = \int_{-\infty}^x f_{Z_{2i|Z_{1i}}} \left( w_i \left| \frac{y_i - \mu_{1i}}{\sigma_i} \right. \right) dw_i = \mathbb{P} \left( Z_{2i} \leq -x \mid Z_{1i} = \frac{y_i - \mu_{1i}}{\sigma_i} \right) = \Phi(x) \]

and

\[ H_i(x) = \int_{-x}^\infty f_{\rho_i Z_{1i} + \sqrt{1 - \rho_i^2} Z_{2i}}(u_i) du_i = \mathbb{P} \left( \rho_i Z_{1i} + \sqrt{1 - \rho_i^2} Z_{2i} > -x \right) = \Phi(x). \]

Applying Theorem 2.5 the proof concludes.

**Remark 2.7.** It is well-known that, if \((X_1, X_2)^T \) is distributed from a bivariate Student-\( t \) distribution, notation \((X_1, X_2)^T \sim t_\nu(\mu, \Sigma)\), where \( \mu = (\mu_1, \mu_2) \in \mathbb{R}^2 \) and \( \Sigma \) is as in (2.1), then
both the marginal and the conditional distributions of $X_2$ given $X_1$ are also univariate Student’s $t$ distributions: $X_1 \sim t_{\nu}(\mu_1, \sigma_1)$ and

$$X_2|X_1 = x_1 \sim t_{\nu+1}\left(\mu_2 + \rho\sigma_2 \left(\frac{x_1 - \mu_1}{\sigma_1}\right), \frac{\nu + (\frac{x_1 - \mu_1}{\sigma_1})^2}{\nu + 1}\sigma_2^2(1 - \rho^2)\right).$$

The above statement is equivalent to

$$\sqrt{\frac{\nu + 1}{\nu + r^2}(1 - \rho^2)} \left(\frac{X_2 - \mu_2}{\sigma_2} - \rho r\right) \left| \frac{X_1 - \mu_1}{\sigma_1} = r \sim t_{\nu+1}(0, 1) \equiv t_{\nu+1}.\right.$$

**Corollary 2.8** (Student-$t$ density generator). If $(Y_i^*, U_i^*)^\top \sim BSY(\mu, \Sigma, g_c)$, where $g_c(x) = (1 + x/\nu)^{-(\nu+2)/2}$ is the density generator of the bivariate Student-$t$ distribution with $\nu$ degrees of freedom, then the PDF of $Y_i^* | U_i^* > 0$ is given by

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

wherein $f_\nu$ and $F_\nu$ denote the PDF and CDF of a classic Student-$t$ distribution with $\nu$ degrees of freedom, respectively.

**Proof.** It is well-known that the vector $(Y_i^*, U_i^*)^\top$ following a bivariate Student-$t$ distribution has the stochastic representation; see Subsection 9.2.6, p. 354 of Balakrishnan and Lai (2009):

$$Y_i^* = \sigma_i Z_{1i} + \mu_{1i},$$
$$U_i^* = \rho_i Z_{1i} + \sqrt{1 - \rho_i^2} Z_{2i} + \mu_{2i},$$

where $Z_{1i} = \sqrt{\nu/\nu} Z_{1i}/\sqrt{Q}$ and $Z_{2i} = \sqrt{\nu/\nu} Z_{2i}/\sqrt{Q}$ are Student-$t$ random variables, wherein $Q \sim \chi^2_\nu$ (chi-square with $\nu$ degrees of freedom) is independent of $\tilde{Z}_{1i}$ and $\rho_i \tilde{Z}_{1i} + \sqrt{1 - \rho_i^2} \tilde{Z}_{2i}$, and $\tilde{Z}_{1i}, \tilde{Z}_{2i} \sim N(0, 1)$ are independent.

When $(Y_i^* - \mu_{1i})/\sigma_i = r$, from Remark 2.7 we have

$$\sqrt{\frac{\nu + 1}{\nu + r^2}(1 - \rho_i^2)} (U_i^* - \mu_{2i} - \rho_i r) \left| \frac{Y_i^* - \mu_{1i}}{\sigma_i} = r \sim t_{\nu+1}.\right.$$

Hence,

$$F_{\nu+1}(w) = P\left(\sqrt{\frac{\nu + 1}{\nu + r^2}(1 - \rho_i^2)} (U_i^* - \mu_{2i} - \rho_i r) \leq w \left| \frac{Y_i^* - \mu_{1i}}{\sigma_i} = r\right.\right), \quad -\infty < w < \infty.$$
By using the representation (2.16), a simple algebraic manipulation shows that the right-hand
of the above probability is

\[ P \left( Z_{2i} \leq \frac{w}{\nu + 1} \bigg| Z_{1i} = r \right). \]

Setting \( w = -\sqrt{(\nu + 1)/(\nu + r^2)} x \) and \( r = (y_i - \mu_{1i})/\sigma_i \) we obtain

\[ P \left( Z_{2i} \leq -x \bigg| Z_{1i} = \frac{y_i - \mu_{1i}}{\sigma_i} \right) = F_{\nu + 1} \left( \sqrt{\frac{\nu + 1}{\nu + (y_i - \mu_{1i})^2}} x \right). \]

Therefore, the function \( G_i \), given in Proposition 2.3 is

\[ G_i(x) = \int_{-\infty}^{\infty} f_{Z_{2i}|Z_{1i}}(w_i \bigg| \frac{y_i - \mu_{1i}}{\sigma_i}) \, dw_i = P (Z_{2i} \leq -x \bigg| Z_{1i} = \frac{y_i - \mu_{1i}}{\sigma_i}) = F_{\nu + 1} \left( \sqrt{\frac{\nu + 1}{\nu + (y_i - \mu_{1i})^2}} x \right). \]

On the other hand, note that \( \rho_i Z_{1i} + \sqrt{1 - \rho_i^2} Z_{2i} = \sqrt{\nu} (\rho_i \tilde{Z}_{1i} + \sqrt{1 - \rho_i^2} \tilde{Z}_{2i})/\sqrt{\nu} \) is distrib-
uted according to the a Student-\( t \) distribution with \( \nu \) degrees of freedom because \( \rho_i \tilde{Z}_{1i} + \sqrt{1 - \rho_i^2} \tilde{Z}_{2i} \sim N(0, 1) \). Then

\[ H_i(x) = \int_{-\infty}^{\infty} f_{\rho_i Z_{1i} + \sqrt{1 - \rho_i^2} Z_{2i}}(u_i) \, du_i = P (\rho_i Z_{1i} + \sqrt{1 - \rho_i^2} Z_{2i} > -x) = F_{\nu}(x). \]

Applying Theorem 2.5 the proof follows. \( \square \)

Given the density generator \( g_c \) we can directly determine the PDF of \( Y_i^* | U_i^* > 0 \) as follows.

**Corollary 2.9.** If \( (Y_i^*, U_i^*)^T \sim BSY(\mu, \Sigma, g_c) \) then the PDF of \( Y_i^* | U_i^* > 0 \) is given by

\[ f_{Y_i^* | U_i^* > 0}(y_i; \theta) = \frac{\rho_i \sqrt{1 - \rho_i^2}}{\sigma_i} \int_{-\infty}^{\mu_{2i}/\sqrt{1 - \rho_i^2} + \rho_i (y_i - \mu_{1i})/\sqrt{1 - \rho_i^2}} \left\{ \int_{-\infty}^{x} g_c \left( \frac{(y_i - \mu_{1i})^2 + w_i^2}{\sigma_i^2} \right) \, dw_i \right\} \, dx, \]

where \( z_i = \mu_{2i}/\sqrt{1 - \rho_i^2} + \rho_i (y_i - \mu_{1i})/\sqrt{1 - \rho_i^2} \).

**Proof.** By using the known formula for the PDF of sum of two random variables \( X \) and \( Y \):

\[ f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x) \, dx, \]

10
we have that $H_i$, defined in (2.6), can be written as

$$H_i(x) = \frac{1}{\rho_i \sqrt{1 - \rho_i^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_e((\frac{\zeta}{\rho_i})^2 + (\frac{\mu_1}{\rho_i})^2) \frac{d\zeta}{\sqrt{1 - \rho_i^2}}. $$

Proposition 2.3 provides

$$f_{Z_i}(\frac{y_i - \mu_1}{\sigma_1}) = \int_{-\infty}^{\infty} g_e((\frac{y_i - \mu_1}{\sigma_1})^2 + y^2) dy. \tag{2.17}$$

By combining Proposition 2.4, Equations (2.11) and (2.17), we have

$$G_i(x) = \int_{-\infty}^{\infty} f_{Z_1i, Z_2i}(\frac{y_i - \mu_1}{\sigma_1}, w_i) dw_i = \int_{-\infty}^{\infty} g_e((\frac{y_i - \mu_1}{\sigma_1})^2 + w_i^2) dw_i.$$

Substituting the above formulas of $H_i$, $f_{Z_i}$, and $G_i$ into Theorem 2.5 we complete the proof.

\[\square\]

2.2 Maximum likelihood estimation

By combining Equation (2.4) with Theorem 2.5 the following formula for the PDF of $Y_i$ is valid:

$$f_{Y_i}(y_i; \theta) = (1 - H_i(\mu_2))^{1-u_i} H_i(\mu_2)^{u_i} \left[ \frac{1}{\sigma_1} f_{Z_1i} \left( \frac{y_i - \mu_1}{\sigma_1} \right) \frac{G_i(\tau_i + \alpha_i(\frac{y_i - \mu_1}{\sigma_1}))}{H_i(\mu_2)} \right]^{u_i},$$

where $\alpha_i = \rho_i / \sqrt{1 - \rho_i^2}$, $\tau_i = \mu_2 / \sqrt{1 - \rho_i^2}$, $u_i = 1$ if $u_i^* > 0$ and $u_i = 0$ otherwise, $g_1(\mu_1i) = x_1^T \beta$, $g_2(\mu_2i) = w_i^T \gamma$, $h_1(\sigma_i) = z_i^T \lambda$ and $h_2(\rho_i) = v_i^T \kappa$.

The log-likelihood of the symmetric generalized Heckman model for $\theta = (\beta^T, \gamma^T, \lambda^T, \kappa^T)^T$ is given by

$$\ell(\theta) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \theta)$$

$$= \sum_{i=1}^{n} u_i \left[ -\log(\sigma_i) + \log f_{Z_1i} \left( \frac{y_i - \mu_1}{\sigma_1} \right) + \log G_i \left( \tau_i + \alpha_i \left( \frac{y_i - \mu_1}{\sigma_1} \right) \right) \right]$$

$$+ \sum_{i=1}^{n} (1 - u_i) \log(1 - H_i(\mu_2)). \tag{2.18}$$

To obtain the ML estimate of $\theta$, we maximize the log-likelihood function \[2.18\] by equating the score vector $\ell(\theta)$ 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

$$
0 = \frac{\partial \ell(\theta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{u_i}{\sigma_i} \left[ \frac{f'_{Z_{i1}}(\frac{y_i - \mu_{i1}}{\sigma_i})}{f_{Z_{i1}}(\frac{y_i - \mu_{i1}}{\sigma_i})} - \alpha_i f_{Z_{i2} | Z_{i1}}(\alpha_i \frac{y_i - \mu_{i1}}{\sigma_i} + \tau_i | \frac{y_i - \mu_{i1}}{\sigma_i}) \right] \frac{\partial \mu_{i1}}{\partial \beta_j};
$$

$$
0 = \frac{\partial \ell(\theta)}{\partial \gamma_r} = \sum_{i=1}^{n} \left[ \frac{u_i}{\sqrt{1 - \rho_i^2}} \frac{f_{Z_{i2} | Z_{i1}}(\alpha_i \frac{y_i - \mu_{i1}}{\sigma_i} + \tau_i | \frac{y_i - \mu_{i1}}{\sigma_i})}{G_i(\alpha_i \frac{y_i - \mu_{i1}}{\sigma_i} + \tau_i)} + \frac{f_{\rho_i, Z_{i1} | Z_{i2}}(\rho_i \frac{Z_{i1}}{\sqrt{1 - \rho_i^2}}, Z_{i2} | - \mu_{2i})}{1 - H_i(\mu_{2i})} \right] \frac{\partial \rho_i}{\partial \gamma_r};
$$

$$
0 = \frac{\partial \ell(\theta)}{\partial \lambda_s} = \sum_{i=1}^{n} \frac{u_i}{\sigma_i} \left[ \frac{(y_i - \mu_{i1}) f'_{Z_{i1}}(\frac{y_i - \mu_{i1}}{\sigma_i})}{f_{Z_{i1}}(\frac{y_i - \mu_{i1}}{\sigma_i})} - \alpha_i (y_i - \mu_{i1}) f_{Z_{i2} | Z_{i1}}(\alpha_i \frac{y_i - \mu_{i1}}{\sigma_i} + \tau_i | \frac{y_i - \mu_{i1}}{\sigma_i}) \right] \frac{\partial \sigma_i}{\partial \lambda_s};
$$

$$
0 = \frac{\partial \ell(\theta)}{\partial \kappa_m} = \sum_{i=1}^{n} \frac{u_i}{\sqrt{1 - \rho_i^2}} \left[ \frac{(y_i - \mu_{i1}) f'_{Z_{i1}}(\frac{y_i - \mu_{i1}}{\sigma_i})}{1 - \rho_i^2} - \mu_{2i} \rho_i \right] \frac{f_{Z_{i2} | Z_{i1}}(\alpha_i \frac{y_i - \mu_{i1}}{\sigma_i} + \tau_i | \frac{y_i - \mu_{i1}}{\sigma_i})}{G_i(\alpha_i \frac{y_i - \mu_{i1}}{\sigma_i} + \tau_i)} \frac{\partial \rho_i}{\partial \kappa_m};
$$

where

$$
\frac{\partial \mu_{i1}}{\partial \beta_j} = \frac{x_{ij}}{g_i'(\mu_{i1})}, \quad j = 1, \ldots, k; \quad i = 1, \ldots, n;
$$

$$
\frac{\partial \mu_{2i}}{\partial \gamma_r} = \frac{w_{ir}}{g_i'(\mu_{2i})}, \quad r = 1, \ldots, l; \quad i = 1, \ldots, n;
$$

$$
\frac{\partial \sigma_i}{\partial \lambda_s} = \frac{z_{is}}{h_i'(\sigma_i)}, \quad s = 1, \ldots, p; \quad i = 1, \ldots, n;
$$

$$
\frac{\partial \rho_i}{\partial \kappa_m} = \frac{v_{im}}{h_i'(\rho_i)}, \quad m = 1, \ldots, q; \quad i = 1, \ldots, n.
$$

### 3 Generalized Heckman-Student-t model

The generalized Heckman-normal model proposed by Bastos et al. [2021] is a special case of (2.3) 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 heavier tails than the bivariate normal distribution. From (2.2), if $Y = (Y_1, Y_2)^\top$ follows a Bt distribution, then the associated PDF is given by

$$
f(y; \mu, \Sigma, \nu) = \frac{1}{|\Sigma|^{1/2} Z_{gcd}} \left\{ 1 + \frac{(y - \mu)^\top \Sigma^{-1}(y - \mu)}{\nu} \right\}^{-(\nu+2)/2}, \quad (3.1)
$$

where $\nu$ is the number of degrees of freedom. Here, the density generator of the Bt distribution is given by $g_s(x) = (1 + x/\nu)^{-(\nu+2)/2}$, $|\Sigma| = \sigma^2(1 - \rho^2)$ and $Z_{gcd} = [\nu \pi \Gamma(\nu/2)]/\Gamma((\nu + 2)/2)$ is a normalization constant. Therefore, if $(Y_i^*, U_i^*)$ follow a Bt distribution, then, by Corollary 2.8.
the PDF of $Y^*_i|U^*_i > 0$ is written as

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

where $f_{\nu}$ and $F_{\nu}$ are the PDF and CDF, respectively, of a univariate Student-t distribution with $\nu$ degrees of freedom, $\alpha_i = \rho_i/\sqrt{1-\rho^2_i}$, and $\tau_i = \mu_{2i}/\sqrt{1-\rho^2_i}$. The log-likelihood for $\theta = (\beta^T, \gamma^T, \lambda^T, \kappa^T, \nu)^T$ is given by

$$\ell(\theta) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \theta) = \sum_{i=1}^{n} \log \left\{ (F_{\nu}(-\mu_{2i}))^{1-u_i} (F_{\nu}(\mu_{2i}))^{u_i} (f_{Y^*_i|U^*_i > 0}(y_i; \mu_{1i}, \sigma^2_i, \alpha_i, \tau_i, \nu))^{u_i} \right\}$$

$$= \sum_{i=1}^{n} u_i \left[ -\log(\sigma_i) + \log f_{\nu}\left(\frac{y_i - \mu_{1i}}{\sigma_i}\right) + \log F_{\nu+1}\left(\sqrt{\frac{\nu+1}{\nu+\left(\frac{y_i - \mu_{1i}}{\sigma_i}\right)^2}} (\tau_i + \alpha_i(\frac{y_i - \mu_{1i}}{\sigma_i}))\right) \right]$$

$$+ \sum_{i=1}^{n} (1-u_i) \log F_{\nu}(-\mu_{2i}),$$

(3.2)

where $u_i = 1$ if $u^*_i > 0$ and $u_i = 0$ otherwise, $\mu_{1i}, \mu_{2i}, \sigma_i$ and $\rho_i$ are as in (2.3). The ML estimate of $\theta$ is obtained by maximizing the log-likelihood function (3.2), that is, by equating the score vector $\ell'(\theta)$ (given in Subsection 2.2) to zero, providing the likelihood equations. They are solved using an iterative procedure for non-linear optimization, such as the BFGS quasi-Newton method.

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

$$\mu_{1i} = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i},$$

$$\mu_{2i} = \gamma_1 + \gamma_2 x_{1i} + \gamma_3 x_{2i} + \gamma_4 x_{3i},$$

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

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

(4.1) (4.2) (4.3) (4.4)
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 \( \text{NREP} = 1000 \) Monte Carlo replicates for each sample size. In the structure presented in (4.1) - (4.4), \( \mu_{1i} \) is the primary interest equation, while \( \mu_{2i} \) 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

\[
\hat{\text{Bias}}(\hat{\theta}) = \frac{1}{\text{NREP}} \sum_{i=1}^{\text{NREP}} \hat{\theta}^{(i)} - \theta \quad \text{and} \quad \hat{\text{MSE}}(\hat{\theta}) = \frac{1}{\text{NREP}} \sum_{i=1}^{\text{NREP}} (\hat{\theta}^{(i)} - \theta)^2, \tag{4.5}
\]

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

We consider the following sets of true parameter values for the regression structure in (4.1)-(4.4):

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

To keep the censoring proportion around 50%, in Scenario 1 a threshold greater than zero was used, so \( U_i^* > a \). According to Bastos et al. (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 proportion was maintained around 30%. In Scenario 3, the censoring rate around 50% was obtained by changing the parameters of the selection equation \( \mu_{2i} \).

The ML estimation results for the Scenarios 1), 2) and 3) are presented in Tables 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 1-3 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: Bias and MSE for the indicated ML estimates of the generalized Heckman-$t$ model parameters (Scenario 1).

| Parameters | $n$  | Censoring | Generalized Heckman-$t$ | Generalized Heckman-$t$ |
|------------|------|-----------|-------------------------|-------------------------|
|            | 500  | 30.8878   | 0.0034                  | 0.0053                  |
| $\beta_1$  | 1000 | 30.9352   | 0.0019                  | 0.0025                  |
|            | 2000 | 31.0035   | 0.0021                  | 0.0011                  |
| $\beta_2$  | 500  | 30.8878   | 0.0037                  | 0.0014                  |
|            | 1000 | 30.9352   | 0.0020                  | 0.0007                  |
|            | 2000 | 31.0035   | 0.0016                  | 0.0003                  |
| $\beta_3$  | 500  | 30.8878   | 0.0009                  | 0.0002                  |
|            | 1000 | 30.9352   | -0.0003                 | 0.0008                  |
|            | 2000 | 31.0035   | 0.0008                  | 0.0004                  |
| $\gamma_1$ | 500  | 30.8878   | 0.0163                  | 0.0137                  |
|            | 1000 | 30.9352   | 0.0055                  | 0.0065                  |
|            | 2000 | 31.0035   | -0.0008                 | 0.0032                  |
| $\gamma_2$ | 500  | 30.8878   | 0.0066                  | 0.0091                  |
|            | 1000 | 30.9352   | 0.0064                  | 0.0044                  |
|            | 2000 | 31.0035   | -0.0008                 | 0.0032                  |
| $\gamma_3$ | 500  | 30.8878   | 0.0177                  | 0.0194                  |
|            | 1000 | 30.9352   | 0.0081                  | 0.0085                  |
|            | 2000 | 31.0035   | 0.0032                  | 0.0041                  |
| $\gamma_4$ | 500  | 30.8878   | 0.0091                  | 0.0110                  |
|            | 1000 | 30.9352   | 0.0081                  | 0.0085                  |
|            | 2000 | 31.0035   | 0.0028                  | 0.0026                  |
| $\kappa_1$ | 500  | 30.8878   | 0.0139                  | 0.0564                  |
|            | 1000 | 30.9352   | 0.0048                  | 0.0048                  |
|            | 2000 | 31.0035   | -0.0005                 | 0.0102                  |
| $\kappa_2$ | 500  | 30.8878   | 0.0589                  | 0.0554                  |
|            | 1000 | 30.9352   | 0.0054                  | 0.0225                  |
|            | 2000 | 31.0035   | 0.0120                  | 0.0083                  |
| $\lambda_1$ | 500  | 30.8878   | 0.0011                  | 0.0049                  |
|            | 1000 | 30.9352   | 0.0210                  | 0.018                   |
|            | 2000 | 31.0035   | 0.0018                  | 0.0011                  |
| $\lambda_2$ | 500  | 30.8878   | 0.0017                  | 0.0031                  |
|            | 1000 | 30.9352   | 0.0006                  | 0.0023                  |
|            | 2000 | 31.0035   | 0.0007                  | 0.0007                  |
| $\nu$      | 500  | 30.8878   | 0.3633                  | 2.3387                  |
|            | 1000 | 30.9352   | 0.1506                  | 0.5565                  |
|            | 2000 | 31.0035   | 0.0833                  | 0.2377                  |
Table 2: Bias and MSE for the indicated ML estimates of the generalized Heckman-t model parameters (Scenario 2).

| n   | Parameters | Censoring | Generalized Heckman-t Parameters | Censoring | Generalized Heckman-t Parameters |
|-----|------------|-----------|----------------------------------|-----------|----------------------------------|
|     |            | Bias      | MSE                              |            | Bias                             |
| 500 | $\beta_1$ | 1.0       | 0.0048                           | 0.0045     | 31.017                           |
| 1000| $\beta_1$ | 1.0       | 0.0031                           | 0.0022     | 30.9395                          |
| 2000| $\beta_1$ | 1.0       | 0.0013                           | 0.0009     | 30.9098                          |
| 500 | $\beta_2$ | 0.7       | 0.0013                           | 0.0009     | 31.017                           |
| 1000| $\beta_2$ | 0.7       | 0.0002                           | 0.0046     | 30.9268                          |
| 2000| $\beta_2$ | 0.7       | 0.0002                           | 0.0001     | 30.9098                          |
| 500 | $\beta_3$ | 1.1       | 0.0071                           | 0.0141     | 31.017                           |
| 1000| $\beta_3$ | 1.1       | 0.0115                           | 0.0081     | 30.9395                          |
| 2000| $\beta_3$ | 1.1       | 0.0076                           | 0.0020     | 30.9098                          |
| 500 | $\gamma_1$| 0.9       | 0.0165                           | 0.0814     | 30.9395                          |
| 1000| $\gamma_1$| 0.9       | 0.0036                           | 0.0031     | 30.9098                          |
| 2000| $\gamma_1$| 0.9       | 0.0151                           | 0.0254     | 30.9395                          |
| 500 | $\gamma_2$| 0.5       | 0.0072                           | 0.0090     | 31.017                           |
| 1000| $\gamma_2$| 0.5       | 0.0076                           | 0.0101     | 30.9395                          |
| 2000| $\gamma_2$| 0.5       | 0.0076                           | 0.0020     | 30.9098                          |
| 500 | $\gamma_3$| 1.1       | 0.0130                           | 0.0180     | 31.017                           |
| 1000| $\gamma_3$| 1.1       | 0.0151                           | 0.0254     | 30.9395                          |
| 2000| $\gamma_3$| 1.1       | 0.0020                           | 0.0041     | 30.9098                          |
| 500 | $\gamma_4$| 0.6       | 0.0067                           | 0.0102     | 31.017                           |
| 1000| $\gamma_4$| 0.6       | 0.0064                           | 0.0195     | 30.9395                          |
| 2000| $\gamma_4$| 0.6       | 0.0001                           | 0.0023     | 30.9098                          |
| 500 | $\kappa_1$| 0.7       | 0.0299                           | 0.0560     | 31.017                           |
| 1000| $\kappa_1$| 0.7       | 0.0197                           | 0.0628     | 30.9395                          |
| 2000| $\kappa_1$| 0.7       | 0.0031                           | 0.0089     | 30.9098                          |
| 500 | $\kappa_2$| 0.3       | 0.052                            | 0.0702     | 31.017                           |
| 1000| $\kappa_2$| 0.3       | 0.0197                           | 0.0576     | 30.9395                          |
| 2000| $\kappa_2$| 0.3       | 0.0097                           | 0.0097     | 30.9098                          |
| 500 | $\lambda_1$| -0.2      | 0.0004                           | 0.0053     | 31.017                           |
| 1000| $\lambda_1$| -0.2      | 0.0019                           | 0.0041     | 30.9395                          |
| 2000| $\lambda_1$| -0.2      | 0.0019                           | 0.0012     | 30.9098                          |
| 500 | $\lambda_2$| 1.2       | 0.0032                           | 0.0038     | 31.017                           |
| 1000| $\lambda_2$| 1.2       | 0.0018                           | 0.0020     | 30.9395                          |
| 2000| $\lambda_2$| 1.2       | 0.0018                           | 0.0007     | 30.9098                          |
| 500 | $\nu$     | 4         | 0.4404                           | 2.3837     | 31.017                           |
| 1000| $\nu$     | 4         | 0.1418                           | 0.6321     | 30.9395                          |
| 2000| $\nu$     | 4         | 0.1234                           | 0.2899     | 30.9098                          |

16
Table 3: Bias and MSE for the indicated ML estimates of the generalized Heckman-t model parameters (Scenario 3).

| n  | Parameters | Censoring | Generalized Heckman-t Parameters | Censoring | Generalized Heckman-t Parameters |
|----|------------|-----------|----------------------------------|-----------|----------------------------------|
| 500| β₁ 1.1     | 49.8604 -0.0037 0.0094 | β₁ 1.1  | 49.9824 -0.0109 0.0072 |
| 1000| 50.0265 -0.0012 0.0017 | 9.9075 -0.0049 0.003 |
| 2000| 49.925 -0.0036 0.0011 |
| 500| β₂ 0.7     | 49.8604 -0.0025 0.0016 | β₂ 0.7  | 49.9824 -0.0057 0.0014 |
| 1000| 49.9469 -0.0006 0.0005 | 9.9075 -0.0026 0.0005 |
| 2000| 49.925 -0.0025 0.0002 |
| 500| β₃ 0.1     | 49.8604 0.0001 0.0016 | β₃ 0.1  | 49.9824 0.0017 0.0011 |
| 1000| 49.9469 0.0001 0.0005 | 9.9075 0.0004 0.0004 |
| 2000| 49.925 -0.0002 0.0002 |
| 500| γ₁ 0       | 49.8604 0.0025 0.0065 | γ₁ 0    | 49.9824 0.0012 0.0063 |
| 1000| 49.9469 0.0036 0.003 | 9.9075 0.0029 0.004 |
| 2000| 49.925 0.0022 0.0014 |
| 500| γ₂ 0.5     | 49.8604 0.0067 0.0078 | γ₂ 0.5  | 49.9824 0.0051 0.0071 |
| 1000| 49.9469 0.0061 0.0035 | 9.9075 0.0036 0.0051 |
| 2000| 49.925 0.0003 0.0015 |
| 500| γ₃ 1.1     | 49.8604 0.0009 0.0161 | γ₃ 1.1  | 49.9824 0.0095 0.0155 |
| 1000| 49.9469 0.0047 0.008 | 9.9075 0.0116 0.0185 |
| 2000| 49.925 0.0046 0.0035 |
| 500| γ₄ 0.6     | 49.8604 0.0010 0.0091 | γ₄ 0.6  | 49.9824 0.0048 0.0083 |
| 1000| 49.9469 0.0014 0.0042 | 9.9075 0.0056 0.0086 |
| 2000| 49.925 0.0038 0.002 |
| 500| κ₁ -0.3    | 49.8604 -0.0105 0.01 | κ₁ -0.7 | 49.9824 -0.0166 0.0437 |
| 1000| 49.9469 -0.0107 0.0173 | 9.9075 -0.0061 0.0183 |
| 2000| 49.925 -0.0034 0.0078 |
| 500| κ₂ -0.3    | 49.8604 -0.0302 0.0394 | κ₂ -0.7 | 49.9824 -0.0831 0.0609 |
| 1000| 49.9469 -0.0189 0.0163 | 9.9075 -0.0317 0.0195 |
| 2000| 49.925 -0.0214 0.0086 |
| 500| λ₁ -0.4    | 49.8604 0.0032 0.0067 | λ₁ -0.4 | 49.9824 -0.0041 0.0072 |
| 1000| 49.9469 0.0003 0.0035 | 9.9075 -0.006 0.0034 |
| 2000| 49.925 -0.0033 0.0018 |
| 500| λ₂ 1.2     | 49.8604 0.0062 0.004 | λ₂ 1.2  | 49.9824 0.0039 0.0038 |
| 1000| 49.9469 0.0057 0.0019 | 9.9075 0.0029 0.0018 |
| 2000| 49.925 0.0016 0.0008 |
| 500| ν 4        | 49.8604 0.5755 3.9653 | ν 4     | 49.9824 0.546 8.083 |
| 1000| 49.9469 0.2316 0.8686 | 9.9075 0.1677 0.793 |
| 2000| 49.925 0.0841 0.4107 |
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 is new and is analyzed for the first time here.

5.1 Outpatient expense

In this subsection, a real data set corresponding to outpatient expense from the 2001 Medical Expenditure Panel Survey (MEPS) 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 \( Y_i^* = \ln(ambx) \), while the latent variable \( U_i^* = dambexp \) is the willingness of the individual to spend; \( U_i = 1_{\{U_i^* > 0\}} \) 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 \( revenue \) denotes the individual income.

Table 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 1(a). The boxplot shown in Figure 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.

| 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 |

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

\[
\ln(ambx) = \beta_0 + \beta_1 age_i + \beta_2 fem_i + \beta_3 educ_i + \beta_4 blhisp_i + \beta_5 totcr_i + \beta_6 ins_i, \tag{5.1}
\]

\[
dambexp = \gamma_0 + \gamma_1 age_i + \gamma_2 fem_i + \gamma_3 educ_i + \gamma_4 blhisp_i + \gamma_5 totcr_i + \gamma_6 ins_i + \gamma_7 income_i, \tag{5.2}
\]
We initially compare the adjustments of the generalized Heckman-\(t\) (GH\(t\)) model, in terms of Akaike (AIC) and Bayesian information (BIC), with the adjustments of the classical Heckman-normal (CHN) \cite{Heckman1976} and generalized Heckman-normal (GHN) \cite{Bastos2021} models; see Table 5. The AIC and BIC values reveal that the GH\(t\) model provides the best adjustment, followed by the GHN model.

Table 5: AIC and BIC of the indicated Heckman models.

|          | CHN    | GHN    | GH\(t\) |
|----------|--------|--------|---------|
| AIC      | 11706.44 | 11660.29 | 11635.05 |
| BIC      | 11810.31 | 11794.71 | 11775.59 |

Table 5 presents the estimation results of the GHN and GH\(t\) models. From this table, we observe the following results: the explanatory variables \textit{totchr} and \textit{ins} that model the dispersion are significant, in both models, indicating the presence of heteroscedasticity in the data. The explanatory variable \textit{age} is only significant in the GHN model. For the correlation term, the covariates \textit{fem} and \textit{totchr} are significant for both models, which indicates the presence of selection bias in the data.

In the outcome equation, when we look at both models, \textit{age}, \textit{fem}, \textit{blhisp} and \textit{totchr} are significant at the 5\% level, and \textit{educ} is not significant. The explanatory variable \textit{ins} is significant at the 5\% and 10\% levels in the GHN and GH\(t\) models, respectively; see Table 5. We can interpret the estimated coefficients in terms of the effect on the expenditure on medical services;
see Weisberg (2014, p.82). For example, a 1-year increase in age rises by \((\exp(0.1838) - 1) \times 100 = 20.18\%\) and \((\exp(0.1895) - 1) \times 100 = 20.86\%\) the expected value of the expenditure on medical services according to the GHN and \(\text{GH}_t\), respectively. Moreover, a 1-unit increase in \(\text{totchr}\) rises by \((\exp(0.1240) - 1) \times 100 = 53.82\%\) and \((\exp(0.4464) - 1) \times 100 = 56.27\%\) the expected value of the response according to the GHN and \(\text{GH}_t\) models, respectively.

In the case of the selection equation, we observe that, for both models, the explanatory variables age, \(\text{fem}\), educ, blhisp, \(\text{totchr}\) and ins are significant at the 5% level and revenue is significant at the 10% level; see Table 6. The interpretation is made in terms of odds ratio, which is obtained by exponentiating the estimated explanatory variable coefficient. For example, for the \(\text{GH}_t\) case, the odds ratio for age is \(\exp(0.0930) = 1.0975\), suggesting that each additional year of age raises the likelihood of an individual having expenditures on medical services by \(((1.0975 - 1) \times 100) = 9.75\%\).

| Table 6: Estimation results of the GHN and \(\text{GH}_t\) models. |
|---------------------------------|-----------------|-----------|-------|-------|-------|-------|-------|-------|
|                               | Variables       | Estimates  | Std. Error | \(t\) Value | \(p\)-value | \(GHN\) | \(\text{GH}_t\) | \(GHN\) | \(\text{GH}_t\) |
|                               | (Intercept)     | -0.5903    | 0.1867     | 0.2014 | -3.1620 | <0.001 | -3.1800 | <0.001 | -3.1400 | <0.001 |
|                               | age             | 0.0863     | 0.0264     | 0.0288 | 3.2610  | <0.001 | 3.2230  | <0.001 | 3.2130  | <0.001 |
|                               | \(\text{fem}\)  | 0.6299     | 0.0597     | 0.0681 | 1.0543  | <0.001 | 1.0395  | <0.001 | 1.0295  | <0.001 |
|                               | educ            | 0.0569     | 0.0114     | 0.0122 | 4.9830  | <0.001 | 4.8110  | <0.001 | 4.6990  | <0.001 |
|                               | blhisp          | -0.3368    | 0.0596     | 0.0647 | -5.6430 | <0.001 | -5.7590 | <0.001 | -5.8190 | <0.001 |
|                               | \(\text{totchr}\) | 0.7585    | 0.0686     | 0.0858 | 1.1043  | <0.001 | 1.0168  | <0.001 | 1.0320  | <0.001 |
|                               | ins             | 0.1727     | 0.0611     | 0.0665 | 2.8240  | <0.001 | 2.7990  | <0.001 | 2.8030  | <0.001 |
|                               | revenue         | 0.0022     | 0.0012     | 0.0013 | 1.8380  | <0.001 | 1.8750  | 0.0661 | 1.8610  | 0.0610 |

Outcome equation

|                               | Variables       | Estimates  | Std. Error | \(t\) Value | \(p\)-value | \(GHN\) | \(\text{GH}_t\) | \(GHN\) | \(\text{GH}_t\) |
|--------------------------------|-----------------|-----------|-----------|-------|-------|-------|-------|-------|-------|
|                               | (Intercept)     | 5.7041    | 0.1930     | 0.1912 | 2.9553 | <0.001 | 2.9316  | <0.001 | 2.9280  | <0.001 |
|                               | age             | 0.1838    | 0.0234     | 0.0230 | 7.8460 | <0.001 | 8.2350  | <0.001 | 8.2030  | <0.001 |
|                               | \(\text{fem}\)  | 0.2498    | 0.0587     | 0.0580 | 4.2530 | <0.001 | 4.4000  | <0.001 | 4.3760  | <0.001 |
|                               | educ            | 0.0013    | 0.0101     | 0.0100 | 0.1290 | 0.8970 | 0.6240  | 0.5329 | 0.6850  | 0.5516 |
|                               | blhisp          | -0.1283   | 0.0577     | 0.0569 | -2.2210 | <0.001 | -2.3630 | <0.001 | -2.3320 | <0.001 |
|                               | \(\text{totchr}\) | 0.4306    | 0.0305     | 0.0297 | 1.4115 | <0.001 | 1.5002  | <0.001 | 1.4860  | <0.001 |
|                               | ins             | -0.1027   | 0.0513     | 0.0501 | -2.000 | <0.001 | -1.9470 | 0.0456 | -1.8800 | 0.0516 |

Dispersion

|                               | Variables       | Estimates  | Std. Error | \(t\) Value | \(p\)-value | \(GHN\) | \(\text{GH}_t\) | \(GHN\) | \(\text{GH}_t\) |
|--------------------------------|-----------------|-----------|-----------|-------|-------|-------|-------|-------|-------|
|                               | (Intercept)     | 0.5081    | 0.0573     | 0.0643 | 8.8550 | <0.001 | 6.4820  | <0.001 | 6.3390  | <0.001 |
|                               | age             | -0.0249   | 0.0125     | 0.0136 | -1.9870 | 0.0469 | -1.5360 | 0.1246 | -1.4830 | 0.1426 |
|                               | \(\text{totchr}\) | -0.1046   | 0.0191     | 0.0208 | -5.4760 | <0.001 | -5.3720 | <0.001 | -5.3090 | <0.001 |
|                               | ins             | -0.1070   | 0.0277     | 0.0303 | -3.8630 | <0.001 | -3.6810 | <0.001 | -3.6090 | <0.001 |

Correlation

|                               | Variables       | Estimates  | Std. Error | \(t\) Value | \(p\)-value | \(GHN\) | \(\text{GH}_t\) | \(GHN\) | \(\text{GH}_t\) |
|--------------------------------|-----------------|-----------|-----------|-------|-------|-------|-------|-------|-------|
|                               | (Intercept)     | -0.6475   | 0.1143     | 0.1118 | -5.666 | <0.001 | -5.413  | <0.001 | -5.3720 | <0.001 |
|                               | \(\text{fem}\)  | -0.4040   | 0.1356     | 0.1489 | -2.978 | <0.001 | -2.835  | <0.001 | -2.7820 | <0.001 |
|                               | \(\text{totchr}\) | -0.4379   | 0.1862     | 0.2102 | -2.351 | <0.001 | -2.378  | 0.0187 | -2.3990 | 0.0174 |
|                               | \(\nu\)         | -12.3230  | -2.7570    | 4.4690 | - <0.001 | - <0.001 | - <0.001 | - <0.001 | - <0.001 |

Figure 2 displays the quantile versus quantile (QQ) plots of the martingale-type (MT) residuals for the GHN and \(\text{GH}_t\) models. This residual is given by...
\[ r_i^{MT} = \text{sign}(r^{Mi}) \sqrt{-2 \left( r^{Mi} + u_i \log(u_i - r^{Mi}) \right)}, \quad i = 1, \ldots, n. \] (5.5)

where \( r^{Mi} = u_i + \log(\hat{S}(t_i)) \), \( \hat{S}(t_i) \) is the fitted survival function, and \( u_i = 0 \) or 1 indicating that case \( i \) is censored or not, respectively; see [Therneau et al. (1990)]. The MT residual is asymptotically standard normal, if the model is correctly specified whatever the specification of the model is. From Figure 2 we see clearly that the GH\( t \) model provides better fit than GHN model.

![QQ plot for the MT residuals for the GHN and GH\( t \) models.](image)

(a) GHN

(b) GH\( t \)

**Figure 2:** QQ plot for the MT residuals for the GHN and GH\( t \) models.

### 5.2 Investments in education

In this subsection, data on education investments are used to illustrate the proposed methodology. We consider the investments made by municipalities of two Brazilian states: Sao Paulo (SP) and Minas Gerais (MG). This data set was obtained from the Brazilian National Fund for Educational Development (FNDE) website\(^1\), which is a federal agency under the Ministry of Education. The origin of these investments comes from the Fund for the Maintenance and Development of Basic Education and the Valorization of Education Professionals (FUNDEB). The resources are distributed to 27 federative units (26 states plus the Federal District), according to the number of students enrolled in their basic education network. This rule is established for the previous year’s school census data, e.g., 2018 resources were based on 2017 student numbers. This method helps to better distribute resources across the country, as it takes into account the size of education networks.

The variable of interest is education investments with 1503 observations, of which 102 (7%) correspond to unobserved investment values identified as zero investment. The explanatory

---

\(^1\) FNDE website: [link to website]

variables considered in the study were: \textit{revenue} represents per capita revenue collected by the municipality; \textit{gnp} is the Gross National Product of the municipality; \textit{distribute} is a dummy variable indicating if the municipality receives resources from the Financial Compensation for Exploration of Mineral Resources (CFEM); this resource must be destined to investments in the areas of health, education and infrastructure for the community; \textit{sp} is an indicator variable for state (\textit{sp} receives value 1); \textit{enrollment} is the school census enrollment numbers.

As in the previous study, the response variable, investment in education, is in the logarithm scale $Y_i^* = \ln\text{invest}$. The latent variable ($U_i^* = \text{dinvest}$) denotes the willingness of the \textit{i}th municipality to invest education; $U_i = 1_{\{U_i^* > 0\}}$ corresponds to the decision or not of the \textit{i}th municipality to invest in education.

The descriptive statistics for the investments in education are reported in Table 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 3(a). The boxplot shown in Figure 3(b) indicates potential outliers. Therefore, we observe that a Student-$t$ model is a reasonable assumption.

| 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|

We then analyze the education investment data using the \textit{GHt} model, expressed as

$$dinvest = \beta_0 + \beta_1 revenue_i + \beta_2 sp_i + \beta_3 enrollment_i,$$  
(5.6)

$$\ln\text{invest} = \gamma_0 + \gamma_1 revenue_i + \gamma_2 sp_i + \gamma_3 enrollment_i + \gamma_4 gnp_i,$$  
(5.7)

$$\log \sigma_i = \lambda_0 + \lambda_1 revenue_i + \lambda_2 sp_i,$$  
(5.8)

$$\arctanh \rho_i = \kappa_0 + \kappa_1 revenue_i + \kappa_2 distribute_i.$$  
(5.9)

Table 8 reports the AIC and BIC of the CHN, GHN and \textit{GHt} models. From Table 8 we observe that the \textit{GHt} model provides better adjustment than other models based on the values of AIC and BIC.

| CHN    | GHN    | \textit{GHt} |
|--------|--------|--------------|
| AIC    | 8269.058| 5873.579    | 5719.590    |
| BIC    | 8306.147| 5953.055    | 5804.365    |

1\text{Filtered data are available at https://repositorio.shinyapps.io/plataforma_de_dados_municipais.}
2\text{https://dados.gov.br/dataset/sistema-arrecadacao.}
3\text{http://www.ipeadata.gov.br/Default.aspx.}
4\text{https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=resultados.}
4\text{https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar.}
Table 9 presents the estimation results of the GHN and GH\textit{t} models. From this table, we note that the explanatory variables \textit{revenue} and \textit{sp} associated with the dispersion are significant, in both models, suggesting the presence of heteroscedasticity. For the explanatory variables related to the correlation parameter, \textit{revenue} is significant at the 5\% and 10\% levels in the GHN and GH\textit{t} models, respectively, while \textit{distribute} is significant at the 10\% and 5\% levels in the GHN and GH\textit{t} models, respectively. Therefore, there is evidence of the presence of sample selection bias in the data.

From Table 9, we observe that in the outcome equation \textit{revenue}, \textit{sp}, \textit{enrollment} and \textit{gnp} are significant, according to the GHN and GH\textit{t} models. Note that increasing \textit{revenue} by one unit is associated with \((\exp(-0.2683) - 1) \times 100 = -23.53\%\) and \((\exp(-0.2167) - 1) \times 100 = -19.48\%\) decrease in the average investment in education according to the GHN and GH\textit{t} models, respectively. Note also that when the municipality is located in the state of Sao Paulo (\textit{sp}), the average investment in education increases by \((\exp(1.0714) - 1) \times 100 = 191.95\%\) and \((\exp(1.0063) - 1) \times 100 = 173.55\%,\) according to the GHN and GH\textit{t} models, respectively. These results show that the municipalities of the state of Sao Paulo invest much more in education than the municipalities of the state of Minas Gerais.

In the selection equation, we observe that the explanatory variables \textit{revenue}, \textit{sp} and \textit{enrollment} are significant in the GH\textit{t} model. On the other hand, \textit{revenue} and \textit{sp} are not significant in the GHN model; see Table 9. We observe that, for the GH\textit{t} case, the odds ratio for \textit{sp} is \(\exp(0.3674) = 1.4439\), that is, the likelihood of a municipality to spend on medical services increases by \(((1.4439 - 1) \times 100) = 44.39\%\) when the municipality is located in the state of Sao Paulo.

Figure 4 displays the QQ plots of the MT residuals. This figure indicates that the MT
Table 9: Estimation results of the GHN and GH\(t\) models.

| Variables    | Probit selection equation | Outcome equation | Dispersion | Correlation |
|--------------|---------------------------|------------------|-----------|-------------|
|              | Estimate | Std. Error | t Value | p-Value | Estimate | Std. Error | t Value | p-Value | Estimate | Std. Error | t Value | p-Value |
| (intercept)  | GHN      | 1.0523     | 0.9955  | 10.3870 | 14.8880  | <0.01     | <0.001   |
|              | GH\(t\)  | 1.2222     | 0.1032  | 1.0750  | 3.5600   | 0.2820    |          |
| revenue      | GHN      | -0.0058    | 0.0223  | -0.2630 | -3.2330  | 0.7920    |          |
|              | GH\(t\)  | -0.0530    | 0.0164  | -1.0750 | 3.5600   | 0.2820    |          |
| sp           | GHN      | 0.1065     | 0.3674  | 1.0750  | 3.5600   | 0.2820    |          |
|              | GH\(t\)  | 0.3674     | 0.0990  | 1.0750  | 3.5600   | 0.2820    |          |
| enrollment   | GHN      | 0.0311     | 0.0032  | 9.6380  | 20.6150  | <0.001    | <0.001   |
|              | GH\(t\)  | 0.0537     | 0.0026  | 9.6380  | 20.6150  | <0.001    | <0.001   |
| revenue      | GHN      | -0.2683    | 0.0332  | -8.0640 | -7.9030  | <0.001    | <0.001   |
|              | GH\(t\)  | -0.2167    | 0.0274  | -8.0640 | -7.9030  | <0.001    | <0.001   |
| sp           | GHN      | 1.0714     | 0.1003  | 10.6810 | 11.4190  | <0.001    | <0.001   |
|              | GH\(t\)  | 1.0063     | 0.0881  | 10.6810 | 11.4190  | <0.001    | <0.001   |
| enrollment   | GHN      | 0.0031     | 0.0005  | 5.6670  | 8.8410   | <0.001    | <0.001   |
|              | GH\(t\)  | 0.0226     | 0.0025  | 5.6670  | 8.8410   | <0.001    | <0.001   |
| gnp          | GHN      | 0.0151     | 0.0023  | 8.4630  | 27.4250  | <0.001    | <0.001   |
|              | GH\(t\)  | 0.0151     | 0.0023  | 8.4630  | 27.4250  | <0.001    | <0.001   |

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.
Figure 4: QQ plot for the MT residuals for the GHN and GHt models.

References

Abdous, B., F. A.-L. G. K. (2005). Extreme behaviour for bivariate elliptical distributions. *The Canadian Journal of Statistics*, 33:317–334.

Balakrishnan, N. and Lai, C. D. (2009). *Continuous Bivariate Distributions*. Springer-Verlag, New York, NY.

Bastos, F. S. and Barreto-Souza, W. (2021). Birnbaum-Saunders sample selection model. *Journal of Applied Statistics*, 48:1896–1916.

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*.

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

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

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*, 5:475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.
Lachos, V. H., Prates, M. O., and Dey, D. K. (2021). Heckman selection-t model: Parameter estimation via the em-algorithm. *Journal of Multivariate Analysis*, 184:104737.

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

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

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

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

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

Paarsch, H. J. (1984). A monte carlo comparison of estimators for censored regression models. pages 24:197–213.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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*, pages 18:395–415.

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

Weisberg, S. (2014). *Applied Linear Regression*. John Wiley & Sons, Hoboken, New Jersey, fourth edition edition.