Statistica Sinica Preprint No: SS-2021-0068									
Title	Title A Generalized Heckman Model With Varying Sample								
	Selection Bias and Dispersion Parameters								
Manuscript ID	SS-2021-0068								
URL	http://www.stat.sinica.edu.tw/statistica/								
DOI	10.5705/ss.202021.0068								
<b>Complete List of Authors</b>	Fernando de Souza Bastos,								
	Wagner Barreto-Souza and								
	Marc G. Genton								
<b>Corresponding Author</b>	Wagner Barreto-Souza								
E-mail	wagner.barretosouza@kaust.edu.sa								
Notice: Accepted version subject to English editing.									

Statistica Sinica

# A Generalized Heckman Model With Varying Sample Selection Bias and Dispersion Parameters

Fernando de Souza Bastos<sup>\*†</sup>, Wagner Barreto-Souza<sup>\*†</sup> and Marc G. Genton<sup>\*</sup>

\*Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Viçosa, Florestal, Brazil

\*Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

<sup>†</sup>Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Abstract: Many proposals have emerged as alternatives to the Heckman selection model, mainly to address the non-robustness of its normal assumption. The 2001 Medical Expenditure Panel Survey data is often used to illustrate this non-robustness of the Heckman model. In this paper, we propose a generalization of the Heckman sample selection model by allowing the sample selection bias and dispersion parameters to depend on covariates. We show that the non-robustness of the Heckman model may be due to the assumption of the constant sample selection bias parameter rather than the normality assumption. Our proposed methodology allows us to understand which covariates are important to explain the sample selection bias phenomenon rather than to only form conclusions about its presence. Further, our approach may attenuate the non-identifiability and multicollinearity problems faced by the existing sample selection models. We explore the inferential aspects of the maximum likelihood estimators (MLEs) for our proposed generalized Heckman model. More specifically, we show that this model satisfies some regularity conditions such that it ensures consistency

Email: fernando.bastos@ufv.br

Corresponding author. Email: wagner.barretosouza@kaust.edu.sa Email: marc.genton@kaust.edu.sa and asymptotic normality of the MLEs. Proper score residuals for sample selection models are provided, and model adequacy is addressed. Simulated results are presented to check the finite-sample behavior of the estimators and to verify the consequences of not considering varying sample selection bias and dispersion parameters. We show that the normal assumption for analyzing medical expenditure data is suitable and that the conclusions drawn using our approach are coherent with findings from prior literature.

*Key words and phrases:* Asymptotics; Heteroscedasticity; Regularity conditions; Score residuals; Varying sample selection bias.

## 1. Introduction

Heckman (1974, 1976) introduced a model for dealing with the sample selection bias problem with the aid of a bivariate normal distribution to relate the outcome of interest and a selection rule. A semiparametric alternative to this model, known as Heckman's two-step method, was proposed by Heckman (1979), so that the nonrobustness of the normal distribution in the presence of outliers could be handled.

The most discussed problem regarding the Heckman model is its sensitivity to the normal assumption of errors. Misspecification of the error distribution leads to inconsistent maximum likelihood estimators, yielding biased estimates (Lai and Tsay, 2018). On the other hand, when the error terms are correctly specified, the estimation by maximum likelihood or by procedures based on likelihood produces consistent and efficient estimators (Leung and Yu, 1996; Enders, 2010).

However, even when the shape of the error density is correctly specified, the heteroskedasticity of the error terms can cause inconsistencies in the parameter estimates, as shown by Hurd (1979) and Arabmazar and Schmidt (1981). In response to this concern, Donald (1995) discussed how heteroskedasticity in sample selection models is relatively neglected and provided two reasons to motivate the importance of taking this into account in practice. The first reason is that typically, the data used to fit sample selection models comprises large databases, in which heterogeneity is commonly found. The second reason is that the estimates of the parameters obtained by fitting the usual selection models may in some cases be more severely affected by heteroscedasticity than by incorrect distribution of the error terms (Powell, 1986).

Nevertheless, even though there is a large body of recent research and studies on sample selection models, few studies have been carried out to correct or minimize the impact of heteroscedasticity. This is one of the aims of our paper. Chib et al. (2009) proposed a semiparametric model for data with sample selection bias. They considered nonparametric functions in their model, which allowed great flexibility in the way covariates affect response variables. They still presented a Bayesian method for the analysis of such models. Subsequently, Wiesenfarth and Kneib (2010) introduced another general estimation method based on Markov Chain Monte Carlo simulation techniques and used a simultaneous equation system that incorporates Bayesian versions of penalized smoothing splines.

Recent works on sample selection models have aimed to address robust alternatives to the Heckman model. In this direction, Marchenko and Genton (2012) proposed a Student-t sample selection model for dealing with the robustness to the normal assumption in the Heckman model. Zhelonkin et al. (2016) proposed a modified robust semiparametric alternative based on Heckman's two-step estimation method. They proved the asymptotic normality of the proposed estimators and provided the asymptotic covariance matrix. To deal with departures from normality due to skewness, Ogundimu and Hutton (2016) introduced the skew-normal sample selection model to mitigate the remaining effect of skewness after applying a logarithmic transformation to the outcome variable.

Another direction that has been explored in the last few years is the modeling of discrete data with sample selection. For instance, Marra and Wyszynski (2016) introduced sample selection models for count data, potentially allowing for the use of any discrete distribution, non-Gaussian dependencies between the selection and outcome equations, and flexible covariate effects. The modeling of zero-inflated count data with sample selection bias is discussed by Wyszynski and Marra (2018). Mu and Zhang (2018) considered the semiparametric identification and estimation of a heteroscedastic binary choice model with endogenous dummy regressors, and no parametric restriction on the distribution of the error term was assumed. This yields general multiplicative heteroscedasticity in both selection and outcome equations and multiple discrete endogenous regressors. A class of sample selection models for discrete and other non-Gaussian outcomes was recently proposed by Azzalini et al. (2019).

Wojtys et al. (2018) introduced a generalized additive model for location, scale and shape, which accounts for non-random sample selection, and Kim et al. (2019) proposed a Bayesian methodology to correct the bias of estimation of the sample selection models based on a semiparametric Bernstein polynomial regression model that incorporates the sample selection scheme into a stochastic monotone trend constraint, variable selection, and robustness against departures from the normality assumption.

In the aforementioned papers, the solution to deal with departures from normality for continuous outcomes is to assume robust alternatives like Student-*t* or skew-normal distributions. Another common approach is to consider non-parametric structures for the density of the error terms. Our proposal goes in a different direction, one which has not yet been explored in the literature.

In this paper, we propose a generalization of the Heckman sample selection model by allowing the sample selection bias and dispersion parameters to depend on covariates. We show that the non-robustness of the Heckman model may be due to the assumption of the constant sample selection bias parameter rather than the normality assumption. Our proposed methodology allows us to understand what covariates are important to explain the sample selection bias phenomenon rather than to solely make conclusions about its presence. It is worthy to mention that our methodology can be straightforwardly adapted for existing sample selection like those proposed by Marchenko and Genton (2012), Ogundimu and Hutton (2016), and Wyszynski and Marra (2018). We now highlight other contributions of our paper:

• We explore the inferential aspects of the maximum likelihood estimators (MLEs) for our proposed generalized Heckman model. More specifically, we show that this model satisfies regularity conditions so that it ensures consistency and asymptotic normality of the MLEs. In particular, we show that the Heckman model satisfies the regularity conditions, which is a new finding.

- Our approach on including covariates in both sample selection bias and dispersion parameters may attenuate the non-identifiability and multicollinearity problems faced by the existing sample selection models.
- Proper residual for sample selection models is proposed as a byproduct of our asymptotic analysis. This is another relevant contribution of our paper, since this point has not yet been thoroughly addressed.
- We develop an R package for fitting our proposed generalized Heckman model. It also includes the Student-*t* and skew-normal sample selection models, which have not been implemented in R (R Core Team, 2020) before. This makes the paper replicable and facilitates the use of our generalized Heckman model by practitioners. The package is available at the GitHub https://fsbmat-ufv. github.io/ssmodels/.
- We show that the normal assumption for analyzing medical expenditure data is suitable and that the conclusions drawn using our approach are consistent with findings from prior literature. Moreover, we identify which covariates are relevant for explaining the presence of sample selection bias in this important dataset.

This paper is organized as follows. In Section 2 we define the generalized Heckman (GH) sample selection model and discuss estimation of the parameters through the maximum likelihood method. Further, diagnostics tools and residual analysis are discussed. Section 3 is devoted to showing that the GH model satisfies regularity conditions that ensure consistency and asymptotic normality of the maximum likelihood estimators. In Section 4, we present Monte Carlo simulation results for evaluating the performance of the maximum likelihood estimators of our proposed model and for checking the behavior of other existing methodologies under misspecification. In Section 5 we apply our generalized Heckman model to the data on ambulatory expenditures from the 2001 Medical Expenditure Panel Survey and show that our methodology overcomes an existing problem in a simple way. Concluding remarks are addressed in Section 6.

## 2. Generalized Heckman Model

Assume that  $\{(Y_{1i}^*, Y_{2i}^*)\}_{i=1}^n$  are linearly related to covariates  $\boldsymbol{x}_i \in \mathbb{R}^p$  and  $\boldsymbol{w}_i \in \mathbb{R}^q$ through the following regression structures:

$$Y_{1i}^* = \mu_{1i} + \epsilon_{1i}, \tag{2.1}$$

$$Y_{2i}^* = \mu_{2i} + \epsilon_{2i}, \tag{2.2}$$

where  $\mu_{1i} = \boldsymbol{x}_i^{\top} \boldsymbol{\beta}, \ \mu_{2i} = \boldsymbol{w}_i^{\top} \boldsymbol{\gamma}, \ \boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^{\top} \in \mathbb{R}^p$  and  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_q)^{\top} \in \mathbb{R}^q$ are vectors of unknown parameters with associated covariate vectors  $\boldsymbol{x}_i$  and  $\boldsymbol{w}_i$ , for  $i = 1, \dots, n$ , and  $\{(\epsilon_{1i}, \epsilon_{2i})\}_{i=1}^n$  is a sequence of independent bivariate normal random vectors. More specifically, we suppose that

$$\begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix} \sim \mathcal{N}_2 \begin{bmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \sigma_i^2 & \rho_i \sigma_i \\ \rho_i \sigma_i & 1 \end{pmatrix} \end{bmatrix}, \quad (2.3)$$

with the following regression structures for the sample selection bias and dispersion parameters:

$$\operatorname{arctanh} \rho_i = \boldsymbol{v}_i^\top \boldsymbol{\kappa} \quad \text{and} \quad \log \sigma_i = \boldsymbol{e}_i^\top \boldsymbol{\lambda}, \tag{2.4}$$

where  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_r)^\top \in \mathbb{R}^r$  and  $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_s)^\top \in \mathbb{R}^s$  are parameter vectors, with associated covariate vectors  $\boldsymbol{v}_i$  and  $\boldsymbol{e}_i$ , for  $i = 1, \dots, n$ . The arctanh (inverse hyperbolic tangent) link function used for the sample selection bias parameter ensures that it belongs to the interval (-1, 1). The variable  $Y_{1i}^*$  is observed only if  $Y_{2i}^* > 0$ while the variable  $Y_{2i}^*$  is latent. We only know if  $Y_{2i}^*$  is greater or less than 0. The equation (2.1) is the primary interest equation and the equation (2.2) represents the selection equation. In practice, we observe the variables

$$U_i = I\{Y_{2i}^* > 0\}, \tag{2.5}$$

$$Y_i = Y_{1i}^* U_i, (2.6)$$

for i = 1, ..., n, where  $I\{Y_{2i}^* > 0\} = 1$  if  $Y_{2i}^* > 0$  and equals 0 otherwise.

Our generalized Heckman sample selection model is defined by (2.1)-(2.6). The classic Heckman model is obtained by assuming constant sample selection bias and dispersion parameters in (2.4).

The mixed distribution of  $Y_i$  is composed by the discrete component

$$P(U_i = u_i) = \Phi(\boldsymbol{w}_i^{\top} \boldsymbol{\gamma})^{u_i} \Phi(-\boldsymbol{w}_i^{\top} \boldsymbol{\gamma})^{1-u_i}, \quad u_i = 0, 1,$$
(2.7)

and by a continuous part given by the conditional density function

$$f(y_i|U_i=1) = \frac{1}{\sigma_i} \phi\left(\frac{y_i - \boldsymbol{x}_i^{\top} \boldsymbol{\beta}}{\sigma_i}\right) \frac{\Phi\left(\frac{\rho_i}{\sqrt{1 - \rho_i^2}} \left(\frac{y_i - \boldsymbol{x}_i^{\top} \boldsymbol{\beta}}{\sigma_i}\right) + \frac{\boldsymbol{w}_i^{\top} \boldsymbol{\gamma}}{\sqrt{1 - \rho_i^2}}\right)}{\Phi(\boldsymbol{w}_i^{\top} \boldsymbol{\gamma})}, \quad (2.8)$$

where  $\phi(\cdot)$  and  $\Phi(\cdot)$  denote the density and cumulative distribution functions of the standard normal, respectively. More details about this are provided in the Supplementary Material.

Let  $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\top}, \boldsymbol{\gamma}^{\top}, \boldsymbol{\lambda}^{\top}, \boldsymbol{\kappa}^{\top})^{\top}$  be the parameter vector. The log-likelihood function is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ u_i \log f(y_i | u_i = 1) + u_i \log \Phi(\mu_{2i}) + (1 - u_i) \log \Phi(-\mu_{2i}) \right\}$$
$$= \sum_{i=1}^{n} u_i \left\{ \log \Phi(\zeta_i) + \log \phi(z_i) - \log \sigma_i \right\} + \sum_{i=1}^{n} (1 - u_i) \log \Phi(-\mu_{2i}), \quad (2.9)$$

where  $u_i = 1$  if  $y_i$  is observed and  $u_i = 0$  otherwise,  $z_i \equiv \frac{y_i - \mu_{1i}}{\sigma_i}$ , and  $\zeta_i \equiv \frac{\mu_{2i} + \rho_i z_i}{\sqrt{1 - \rho_i^2}}$ , for i = 1, ..., n.

Expressions for the score function  $\mathbf{S}_{\theta} = \frac{\partial \mathcal{L}(\theta)}{\partial \theta}$  and the respective Hessian matrix are presented in the Supplementary Material. The maximum likelihood estimators (MLEs) are obtained as solution of the non-linear system of equations  $\mathbf{S}_{\theta} = \mathbf{0}$ , which does not have an explicit analytic form. We use the quasi-Newton algorithm of Broyden-Fletcher-Goldfarb-Shanno (BFGS) via the package optim of the software R (R Core Team, 2020) to maximize the log-likelihood function.

It is important to emphasize that the maximum likelihood method for the classic

Heckman model may suffer from possible multicollinearity problems when the selection equation has the same covariates as the regression equation. This and other issues such as no convergence of standard algorithms to estimate parameters, multiple local maxima, and the MLE not existing in some cases are discussed for instance in Nawata (1993), Nawata (1994), Yamagata and Orme (2005), Nawata (2007) and Marchenko and Genton (2012). To reduce the impact of this problem in parameter estimation, the exclusion restriction is suggested in the literature. According to this approach, at least one significant covariate included in the selection equation should not be included in the primary regression. The interested reader can find more details on the exclusion restriction procedure for the Heckman sample selection model in Heckman (1976), Leung and Yu (2000), and Newey (2009).

A robust alternative to the normal assumption is to use the two-step estimation procedure by Heckman (1979), which is based on the conditional expectation (for constant  $\sigma$  and  $\rho$ ):

$$E(Y_i|U_i = 1) = \boldsymbol{x}_i^{\top} \boldsymbol{\beta} + \rho \sigma \Lambda(\boldsymbol{w}_i^{\top} \boldsymbol{\gamma}), \qquad (2.10)$$

where  $\Lambda(x) = \frac{\phi(x)}{\Phi(x)}$  is the inverse Mills ratio function. The ordinary least square procedure based solely on (2.10) yields inconsistent estimators due to the extra term  $\rho\sigma\Lambda(\boldsymbol{w}_i^{\top}\boldsymbol{\gamma})$  when  $\rho \neq 0$  and then a second step is necessary. With this in mind, Heckman (1979) proposed to perform a probit regression in the first step so obtaining estimates for  $\boldsymbol{\gamma}$  and then plug them in (2.10). The second step considers (2.10) and additionally a least-squares residual variance or average-predicted probabilities from the probit model. For more details, please see Marchenko and Genton (2012).

Even the two-step estimation can suffer from a possible multicollinearity problem because of the close-to-linear behavior of the inverse Mills ratio function for a certain range. We now argue that this issue is attenuated if considering covariates in both sample selection and dispersion parameters. In our case, the conditional expectation  $E(Y_i|U_i = 1)$  assumes the form

$$E(Y_i|U_i = 1) = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \tanh(\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\kappa})\exp(\boldsymbol{e}_i^{\mathsf{T}}\boldsymbol{\lambda})\Lambda(\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{\gamma}).$$
(2.11)

Due to the non-linearity of the functions  $tanh(\cdot)$  and  $exp(\cdot)$ , the identification of the parameters is obtained even when  $\Lambda(\cdot)$  behaves like a linear function. Therefore, ordinary least square estimators can be proposed based on (2.11) for all parameters with no need for a second step.

To illustrate graphically the importance of including covariates in both sample selection bias and dispersion parameters, consider just one common  $\mathbb{R}$ -valued covariate for all the regression structures, say x, the intercepts equal to 0 and the remaining parameters equal to 1. Define the functions  $R_1(x) = \tanh(x) \exp(x) \Lambda(x)$  (varying sample selection bias and dispersion),  $R_2(x) = \exp(x) \Lambda(x)$  (constant sample selection bias and varying dispersion), and  $R_3(x) = \tanh(x) \Lambda(x)$  (constant dispersion and sample selection bias), for  $x \in \mathbb{R}$ . Figure 1 presents the plots of the functions  $R_1(\cdot), R_2(\cdot), R_3(\cdot)$ , and  $\Lambda(\cdot)$ , for  $x \in (-5, 5)$ . From this figure, we can observe a non-linear behavior of the functions  $R_1$  and  $R_2$  in contrast with  $R_3$  and  $\Lambda$ , where a linear behavior is observed for certain ranges. In the  $R_2(\cdot)$  case, we are thinking of  $\rho$  as constant. Even  $R_2$  being non-close to a linear function, the non-inclusion of covariates for  $\rho$  will imply that this parameter can be absorved into  $\Lambda$  when behaving linearly, so causing both non-identifiability and multicolinearity problems. Thus, we conclude that the best approach is to consider regression structures for both sample selection bias and dispersion parameters to avoid these issues.



**Figure 1:** Plots of the functions  $R_1(x)$ ,  $R_2(x)$ ,  $R_3(x)$ , and  $\Lambda(x)$  for  $x \in (-5, 5)$ .

We now discuss diagnostic techniques, which have been proposed to detect observations that could exercise some influence on the parameter estimates or inference in general. Next, for the generalized Heckman model, we describe the generalized Cook distance (GCD), and in the next section, we propose the score residual. Cook's distance is a method commonly used in statistical modeling to evaluate changes in the estimated vector of parameters when observations are deleted. It allows us to assess the effect of each observation on the estimated parameters. The methodology proposed by Cook (1977) suggests the deletion of each observation and the evaluation of the log-likelihood function without such a case. According to Xie and Wei (2007), the generalized Cook distance (GCD) is defined by

$$\operatorname{GCD}_{i}(\boldsymbol{\theta}) = \left(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}\right)^{\top} \boldsymbol{M}\left(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}\right), \quad i = 1, \dots, n,$$

where M is a nonnegative definite matrix, which measures the weighted combination of the elements for the difference  $\hat{\theta} - \hat{\theta}_{(i)}$ , and  $\hat{\theta}_{(i)}$  is the MLE of  $\theta$  when removing the *i*th observation. Many choices for M were considered by Cook and Weisberg (1982). We use the inverse variance-covariance matrix  $M = -\ddot{\mathcal{L}}(\hat{\theta})^{-1}$ . To determine whether the *i*th case is potentially influential on inference about  $\theta$ , we check if its associated GCD value is greater than 2p/n. In this case, this point would be a possible influential observation.

We illustrate the usage of the GCD in the analysis of the medical expenditure data in Section 5. Regarding the residual analysis, in the next section, we propose a proper residual for sample selection models, which is one of the aims of this paper.

#### 3. Asymptotic Properties and Score Residuals

Our aim in this section is to show that under some conditions, our proposed generalized Heckman sample selection model satisfies the regularity conditions stated by Cox and Hinkley (1979). As a consequence, the maximum likelihood estimators discussed in the previous section are consistent and asymptotically normal distributed. As a by-product of our findings here, we propose a score residual that is well-known to be approximately normal distributed. Proofs of the theorems stated in this section can be found in the Appendix.

Let  $\Theta$  be the parameter space and  $\ell_i(\theta) = u_i \{\log \Phi(\zeta_i) + \log \phi(z_i) - \log \sigma_i\} + (1 - u_i) \log \Phi(-\mu_{2i})$  be the contribution of the *i*th observation to the log-likelhood function, where  $\zeta_i$  retains its definition from previous section, for i = 1, ..., n.

**Theorem 1.** The score function associated to the generalized Heckman model has mean zero and satisfies the identity  $E\left(\mathbf{S}_{\boldsymbol{\theta}}\mathbf{S}_{\boldsymbol{\theta}}^{\top}\right) = -E\left(\partial\mathbf{S}_{\boldsymbol{\theta}}^{\top}/\partial\boldsymbol{\theta}\right).$ 

We now propose a new residual for sample selection models inspired from Theorem 1. From (6.17), we define the ordinary score residual by  $s_i = z_i - \frac{\rho_i}{\sqrt{1-\rho_i^2}} \frac{\phi(\zeta_i)}{\Phi(\zeta_i)}$ for the non-censored observations (where  $u_i = 1$ ) and the standardized score residual by

$$S_{i} = \frac{s_{i} - E(s_{i}|U_{i} = 1)}{\sqrt{\operatorname{Var}(s_{i}|U_{i} = 1)}} = \frac{z_{i} - \frac{\rho_{i}}{\sqrt{1 - \rho_{i}^{2}}} \frac{\phi(\zeta_{i})}{\Phi(\zeta_{i})}}{\sqrt{1 + \mu_{2i}\rho_{i}^{2}} \frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})} + \frac{\rho_{i}^{2}}{1 - \rho_{i}^{2}} \Psi_{i}},$$

where  $\Psi_i = E\left(\frac{\phi^2(\zeta_i)}{\Phi^2(\zeta_i)}|U_i=1\right) = \frac{1}{\sigma_i \Phi(\mu_{2i})} \int_{-\infty}^{\infty} \phi(z_i)\phi(\zeta_i)/\Phi(\zeta_i)dy_i$ , for  $i = 1, \ldots, n$  such that  $u_i = 1$ . Alternatively, a score residual based on all observations (including

the censored ones) can be defined by

$$S_{i}^{*} = \frac{s_{i} - E(s_{i})}{\sqrt{\operatorname{Var}(s_{i})}} = \frac{z_{i} - \frac{\rho_{i}}{\sqrt{1 - \rho_{i}^{2}}} \frac{\phi(\zeta_{i})}{\Phi(\zeta_{i})}}{\sqrt{\Phi(\mu_{2i})\left(1 + \mu_{2i}\rho_{i}^{2}\frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})} + \frac{\rho_{i}^{2}}{1 - \rho_{i}^{2}}\Psi_{i}\right)}},$$
(3.12)

for i = 1, ..., n. In practice, we replace the unknown parameters by their maximum likelihood estimates. The evaluation of the goodness-of-fit of our proposed generalized Heckman model will be performed through the score residual analysis. Based on this approach, discrepant observations are identified, besides being possible to evaluate the existence of serious departures from the assumptions inherent to the model. If the model is appropriate, plots of residuals versus predicted values should have a random behavior around zero. Alternatively, a common approach is to build residual graphics with simulated envelopes (Atkinson, 1985). In this case, it is not necessary to know about the distribution of the residuals, they just need to be within the region formed by the envelopes so indicating a good fit. Otherwise, residuals outside the envelopes are possible outliers or indicate that the model is not properly specified. We will apply the proposed score residual (3.12) to the MEPS data analysis. As will be shown, the residual analysis indicates that the normal assumption for the data is suitable in contrast with the non-robustness of the Heckman model mentioned in existing papers in the literature on sample selection models.

We now establish the consistency and asymptotic normality of the MLEs for our proposed generalized Heckman model. For this, we need to assume some usual regularity conditions. (C1) The parameter space  $\Theta$  is closed and compact, and the true parameter value, say  $\theta_0$ , is an interior point of  $\Theta$ .

(C2) The covariates are a sequence of iid random vectors, and  $\mathbf{F}_n$  is the information matrix conditional on the covariates.

(C3)  $E(\mathbf{F}_n)$  is well-defined and positive definite, and  $E(\max_{\theta \in \Theta} \|\mathbf{F}_n\|) < \infty$ , where  $\|\cdot\|$  is the Frobenius norm.

**Remark 1.** Conditions (C2) and (C3) enable us to apply a multivariate Central Limit Theorem for iid random vectors for establishing the asymptotic normality of the MLEs. These conditions are discussed for instance in Fahrmeir and Kaufmann (1985).

**Theorem 2.** Under Conditions (C1)–(C3), the maximum likelihood estimator  $\widehat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$ for the generalized Heckman model is consistent and satisfies the weak convergence:  $\mathbf{F}_n^{1/2}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \stackrel{d}{\longrightarrow} \mathcal{N}(\mathbf{0},\mathbf{I})$ , where  $\mathbf{I}$  is the identity matrix,  $\mathbf{F}_n$  is the conditional information matrix, and  $\boldsymbol{\theta}_0$  denotes the true parameter vector value.

**Remark 2.** An important consequence of Theorem 2 is that the classic Heckman model is regular under Conditions (C1)-(C3). Therefore, the MLEs for this model are consistent and asymptotically normal distributed, which is a new finding of this paper.

# 4. Monte Carlo Simulations

## 4.1 Simulation design

In this section, we develop Monte Carlo simulation studies to evaluate and compare the performance of the maximum likelihood estimators under the generalized Heckman, classic Heckman, Heckman-Skew, and Heckman-*t* models when the assumption of either constant sample selection bias parameter or constant dispersion is not satisfied. To do this, six different scenarios with relevant characteristics for a more detailed evaluation were considered. In Scenarios 1 and 2, we use models with both varying dispersion and correlation (sample selection bias parameter) and with (I) exclusion restriction and (II) without the exclusion restriction.

For Scenarios 3-6, the exclusion restriction is considered. More specifically, in Scenarios 3, 4, and 5, we have specified the following (III) constant dispersion and varying correlation; (IV) varying dispersion and constant correlation; (V) both constant dispersion and correlation. To evaluate the sensitivity in the parameter estimation of the selection models at high censoring, in Scenario 6 we simulated from the generalized Heckman model with (VI) varying both sample selection bias and dispersion parameters and an average censorship of 50%.

Scenario 1 aims to evaluate the performance of the generalized Heckman model and compare it with its competitors when the assumption of constant sample selection bias parameter and dispersion is not satisfied. Scenario 2 is devoted to demonstrating that despite the absence of exclusion restriction, our model can yield satisfactory parameter estimates. Scenarios 3 and 4 aim to justify the importance of modeling through covariates the correlation and dispersion parameters, respectively. Scenario 5 illustrates some problems that the generalized Heckman model can face as with the classic Heckman model. Finally, Scenario 6 was included to demonstrate the sensitivity of selection models to high correlation and high censoring. We here present the results from Scenario 1. The remaining results are presented in the Supplementary Material.

All scenarios were based on the following regression structures:

$$\mu_{1i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{4.13}$$

$$\mu_{2i} = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{3i}, \qquad (4.14)$$

$$\log \sigma_i = \lambda_0 + \lambda_1 x_{1i},\tag{4.15}$$

$$\operatorname{arctanh} \rho_i = \kappa_0 + \kappa_1 x_{1i}, \tag{4.16}$$

for i = 1, ..., n. All covariates were generated from a standard normal distribution and were kept constant throughout the experiment. The responses were generated from the generalized Heckman model according to each of the six configurations. We set the sample sizes n = 500, 1000, 2000 and N = 1000 Monte Carlo replicates. Pilot simulations showed that the choice of parameters used in the simulations does not affect the results, as long as they maintain the same average percentage of censorship.

We would like to highlight that there is no R package for fitting the Heckmant and Heckman skew-normal models. Therefore, we developed an R package (to be submitted) able to fit our proposed generalized Heckman model and also the sample selection models by Marchenko and Genton (2012) and Ogundimu and Hutton

#### (2016).

In the maximization procedure to estimate parameters based on the optim package, we consider as initial values for  $\beta$ ,  $\gamma$ ,  $\lambda_0$ , and  $\kappa_0$  the maximum likelihood estimates by the classic Heckman model. For the remaining parameters, we set  $\lambda_i = 0$ and  $\kappa_j = 0$  for i = 1, ..., r and j = 1, ..., s. The initial values for the degrees of freedom of the Heckman-t model and the skewness parameter for the Heckman-Skew model were set to 1 and 2, respectively. These values were chosen after some pilot simulations.

#### 4.2 Scenario 1: Varying sample selection bias and dispersion parameters

Here, we consider (4.13)-(4.16) with  $\boldsymbol{\beta} = (1.1, 0.7, 0.1)^{\top}, \boldsymbol{\gamma} = (0.9, 0.5, 1.1, 0.6)^{\top}$ , and also  $\boldsymbol{\lambda} = (-0.4, 0.7)^{\top}$  and  $\boldsymbol{\kappa} = (0.3, 0.5)^{\top}$ . The regressors are kept fixed throughout the simulations with  $x_{1i}, x_{2i}, x_{3i} \stackrel{iid}{\sim} \mathcal{N}(0, 1)$  for all  $i = 1, \ldots, n$ .

In Table 1, we present the empirical mean and root mean square error (RMSE) of the maximum likelihood estimates of the parameters based on the generalized Heckman, classic Heckman, Student-t, and skew-normal sample selection models under Scenario 1. From this table, we observe a good performance of the MLEs based on the generalized Heckman model, even for estimating the parameters related to the sample selection bias and dispersion. The bias and the RMSE under this model decrease for all the estimates as the sample size increases; therefore, suggesting the consistency of the MLEs, which is in line with our Theorem 2. On the other hand, even the regression structures for  $\beta$  and  $\gamma$  being correctly specified, we see that the MLEs for these parameters, based on the classic Heckman, skew-normal, and

#### 4.2 Scenario 1: Varying sample selection bias and dispersion parameters

Student-t, do not provide satisfactory estimates even for a large sample. This illustrates the importance of considering covariates for the sample selection bias and dispersion parameters. The mean estimates of the degrees of freedom and skewness for the Student-t and skew-normal sample selection models were 2.4 and 0.8, respectively.

The above comments are also supported by Figures 2, 3, and 4, where boxplots of the parameter estimates are presented for sample sizes n = 500, n = 1000, and n = 2000, respectively. We did not present the boxplots of the estimates of  $\gamma_1$ ,  $\gamma_2$ , and  $\beta_1$  since they behaved similarly to other boxplots.

We now provide some simulations to evaluate the size and power of likelihood ratio, gradient, and Wald tests. We consider Scenario 1 and present the empirical significance level of the tests in Table 2 for nominal significance levels at 1%, 5%, and 10%.

Table 1: Empirical mean and root mean square error (RMSE) of the maximum likelihood	1
estimates of the parameters based on the generalized Heckman, classic Heckman, Student-t	,
and skew-normal sample selection models under Scenario 1.	

			Gener	alized	C	assic	Hee	ckman Heckman			
			Heck	man	Hee	rkman	Skew	-Normal	l Student- <i>t</i>		
Para	meters	n	mean	RMSE	$\overline{\mathrm{mean}}$	RMSE	$\frac{\text{blew}}{\text{mean}}$	RMSE	mean	RMSE	
		500	0.912	0.093	0.880	0.095	0.854	0.364	1.300	0.435	
$\gamma_0$	0.9	1000	0.905	0.063	0.878	0.069	0.820	0.364	1.293	0.409	
10		2000	0.903	0.042	0.877	0.048	0.804	0.348	1.286	0.395	
		500	0.507	0.081	0.555	0.104	0.527	0.095	0.652	0.200	
$\gamma_1$	0.5	1000	0.506	0.056	0.558	0.085	0.530	0.075	0.663	0.185	
,		2000	0.504	0.040	0.555	0.070	0.529	0.057	0.662	0.174	
		500	1.118	0.104	1.053	0.125	1.001	0.169	1.540	0.480	
$\gamma_2$	1.1	1000	1.109	0.074	1.046	0.099	0.978	0.162	1.522	0.441	
		2000	1.107	0.053	1.033	0.090	0.980	0.146	1.515	0.425	
		500	0.606	0.084	0.556	0.106	0.531	0.127	0.848	0.288	
$\gamma_3$	0.6	1000	0.604	0.065	0.539	0.098	0.504	0.130	0.828	0.253	
		2000	0.602	0.040	0.543	0.074	0.513	0.106	0.833	0.242	
		500	1.105	0.068	0.998	0.261	0.496	0.734	1.112	0.103	
$\beta_0$	1.1	1000	1.102	0.044	0.954	0.273	0.553	0.701	1.108	0.078	
		2000	1.099	0.029	0.892	0.230	0.498	0.697	1.096	0.051	
		500	0.702	0.036	0.882	0.219	0.739	0.156	0.787	0.106	
$\beta_1$	0.7	1000	0.701	0.020	0.860	0.191	0.753	0.158	0.777	0.087	
		2000	0.700	0.014	0.881	0.191	0.770	0.154	0.774	0.079	
		500	0.098	0.048	0.140	0.196	0.040	0.225	0.101	0.078	
$\beta_2$	0.1	1000	0.099	0.029	0.174	0.196	0.069	0.237	0.104	0.057	
		2000	0.100	0.019	0.221	0.147	0.105	0.193	0.113	0.042	
		500	-0.400	0.042	0.174	0.577	0.366	0.771	-0.476	0.113	
$\lambda_0$	-0.4	1000	-0.402	0.028	0.182	0.584	0.339	0.743	-0.470	0.092	
		2000	-0.401	0.019	0.182	0.582	0.331	0.734	-0.463	0.075	
		500	0.699	0.042	_	_	_	_	_	_	
$\lambda_1$	0.7	1000	0.700	0.031	_	_	_	_	_	_	
		2000	0.701	0.020	—	—	—	—	_	_	
		500	0.314	0.248	0.507	0.688	0.133	0.846	0.307	0.300	
$\kappa_0$	0.3	1000	0.311	0.154	0.588	0.714	0.223	0.918	0.302	0.235	
		2000	0.309	0.106	0.811	0.608	0.307	0.814	0.336	0.158	
		500	0.573	0.240	_	—	_	_	_	_	
$\kappa_1$	0.5	1000	0.531	0.135	_	—	_	_	_	_	
		2000	0.510	0.092	_	—	_	_	_	_	

Statistica Sinica: Newly accepted Paper



# 4.2 Scenario 1: Varying sample selection bias and dispersion parameters

Statistica Sinica: Newly accepted Paper



# 4.2 Scenario 1: Varying sample selection bias and dispersion parameters

Statistica Sinica: Newly accepted Paper



# 4.2 Scenario 1: Varying sample selection bias and dispersion parameters

4.2	Scenario 1	l: \	/arying	sample	e selecti	on bias	and	disper	rsion	parameters
-----	------------	------	---------	--------	-----------	---------	-----	--------	-------	------------

	Generalized			Generalized Classic Heckman					Heckman				
	H	leckma	n	H	leckma	n	Sk	Skew-normal			Student-t		
n	LR	G	W	LR	G	W	LR	G	W	L	R G	W	
$\alpha = 1\%$													
500	1.5	1.1	3.5	30.3	1.6	61.4	41.9	6.4	71.3	2.	8 1.5	4.3	
1000	0.9	0.7	2.6	72.4	7.7	92.7	82.7	26.2	95.6	3.	4 - 2.5	4.0	
2000	1.1	0.7	1.9	85.9	16.5	96.8	9 <mark>1.</mark> 0	52.4	99.0	3.	2 2.6	3.9	
						$\alpha = 5$	%						
500	6.2	5.6	12.0	50.5	12.5	69.9	61.7	26.9	78.3	9.	9 7.7	11.6	
1000	4.9	3.7	6.9	86.9	27.1	95.5	89.9	49.7	97.8	10	.1 8.6	11.8	
2000	5.9	5.1	7.4	94.2	40.8	97.7	94.5	68.3	99.5	10	.2 9.4	10.6	
lpha=10%													
500	13.0	10.5	19.3	59.8	25.2	74.4	69.8	40.8	82.2	16	.2 14.8	18.4	
1000	9.4	8.2	13.7	91.5	41.4	96.1	93.3	61.1	98.2	16	.9 15.4	18.0	
2000	11.2	10.8	12.6	96.3	55.7	98.3	95.5	74.8	99.6	16	.1 15.8	16.4	

**Table 2:** Empirical significance level of the likelihood ratio (LR), gradient (G) and Wald (W) tests for  $H_0: \rho = 0$ .

Under the null hypothesis of absence of sample selection bias ( $\rho_i = 0$  for all *i*, that is  $\kappa_0 = \kappa_1 = 0$ ), the likelihood ratio, gradient, and Wald tests presented empirical values close to the nominal values only under the generalized Heckman model. For the other models, the type-I error was inflated and indicated the presence of selection bias. This suggests that the tests should be used with caution to test parameters of the sample selection models and that some confounding can be caused due to either varying sample selection bias or heteroskedasticity. It is important to point out that, even for the generalized Heckman model, the Wald test presents a considerable inflated type-I error for n = 500.

#### 4.2 Scenario 1: Varying sample selection bias and dispersion parameters

**Table 3:** Empirical power of the likelihood ratio (LR), gradient (G) and Wald (W) tests (in percentage) for simulated data according to Scenario 1 under generalized Heckman, classic Heckman, Heckman-Skew, and Heckman-t models, with significance level at 1%, 5% and 10%.

	Generalized			Classic			Heckman				Heckman			
	H	leckma	n	]	Heckma	n		Skew-normal			Student- $t$			
n	LR	G	W	LR	G	W	-	LR	G	W		LR	G	W
						$\alpha = 10^{\circ}$	%							
500	71.7	68.4	75.5	58.4	8.9	81.9		45.7	4.5	79.7		14.6	9.3	18.4
1000	95.3	95.2	96.9	93.7	30.8	99.2		88.0	14.0	98.4		20.9	16.0	26.6
2000	99.9	99.9	99.9	99.6	74.4	100		98.2	28.4	99.4		48.3	46.0	52.1
						$\alpha = 50$	%							
500	88.8	87.8	91.6	74.1	32.6	89.3		69.2	20.1	86.3		29.6	25.1	33.6
1000	99	98.9	99.4	98.1	57.4	99.9		94.5	33.5	99		41.9	37.9	46.2
2000	100	100	100	99.9	87.0	100		98.9	46.8	99.7		68.8	66.9	70.9
lpha=10%														
500	94.3	93.1	95.5	82.4	48.4	91.9		78.2	34.9	89.8		39.8	37.1	43.1
1000	99.8	99.8	99.8	99.1	69.6	100		96	45.5	99		52.4	49.4	54.7
2000	100	100	100	100	91.6	100		99.2	55.9	99.7		78.4	77.8	79.4

In Table 3, we present the empirical power of the likelihood ratio, gradient and Wald tests (in percentage) for simulated data according to Scenario 1 under generalized Heckman, classic Heckman, Heckman-Skew, and Heckman-t models, with significance level at 1%, 5% and 10%. From these results, we can observe that the tests, under the generalized Heckman model, provide high power, mainly when the sample size increases. On the other hand, since tests based on the other models do not provide the correct nominal significance level, the power of tests in these cases are not really comparable.

#### 5. MEPS Data Analysis

We present an application of the proposed model to a set of real data. Consider the outpatient expense data of the 2001 Medical Expenditure Panel Survey (MEPS) available in the R software in the package ssmrob (Zhelonkin et al., 2014). These data were also used by Cameron and Trivedi (2009), Marchenko and Genton (2012), and Zhelonkin et al. (2016) to fit the classic Heckman model, Heckman-t model, and the robust version of the two-step method, respectively. The MEPS is a set of largescale surveys of families, individuals, and their medical providers (doctors, hospitals, pharmacies, etc.) in the United States. It has data on the health services Americans use, how often they use them, the cost of these services, and how they are paid, as well as data on the cost and reach of health insurance available to American workers.

The sample is restricted to persons aged between 21 and 64 years and contains a variable response with 3,328 observations of outpatient costs, of which 526 (15.8 %) correspond to unobserved expenditure values identified as zero expenditure. It also includes the following explanatory variables: Age represents age measured in tens of years; Fem is an indicator variable for gender (female receives value 1); Educ represents years of schooling; Blhisp is an indicator for ethnicity (black or Hispanic receive a value of 1); Totcr is the total number of chronic diseases; Ins is the insurance status; and Income denotes the individual income.

The variable of interest  $Y_{1i}^*$  represents the log-expenditure on medical services of the *i*th individual. We consider the logarithm of the expenditure since it is highly skewed (i.e., see Figure 5, where plots of the expenditure and log-expenditure are



Figure 5: Histograms of the medical expenditure data (to the left) and of its logarithm (to the right).

presented). The variable  $Y_{2i}^*$  denoting the willingness of the *i*th individual to spend is not observed. We only observe  $U_i = I\{Y_{2i}^* > 0\}$ , which represents the decision or not of the *i*th individual to spend on medical care.

According to Cameron and Trivedi (2009) and Zhelonkin et al. (2016), it is natural to fit a sample selection model to such data, since the willingness to spend  $(Y_2^*)$  is likely to be related to the expense amount  $(Y_1^*)$ . However, after fitting the classic Heckman model and using the Wald statistics for testing  $H_0: \rho = 0$  against  $H_1: \rho \neq 0$ , the conclusion is that there is no statistical evidence (*p*-value > 0.1) to reject  $H_0$ , that is, there is no sample selection bias. Cameron and Trivedi (2009) suspected this conclusion on the absence of sample selection bias, and Marchenko and Genton (2012) argued that a more robust model could evidence the presence of sample selection bias in the data; these authors proposed a Student-*t* sample selection model to deal with this problem. However, as will be illustrated in this application, this problem of the classic Heckman model can be due to the assumption of constant sample selection bias and constant dispersion parameters rather than the normal assumption itself. For those readers interested in an alternative approach for dealing with medical expenditures, we refer to Nawata and Kimura (2017), where a standard (type I) Tobit model with the power transformation is considered.

After a preliminary analysis, we consider the following regression structures for our proposed generalized Heckman model:

$$\begin{split} \mu_{1i} &= \beta_0 + \beta_1 \texttt{Age}_i + \beta_2 \texttt{Fem}_i + \beta_3 \texttt{Educ}_i + \beta_4 \texttt{Blhisp}_i + \beta_5 \texttt{Totchr}_i + \beta_6 \texttt{Ins}_i, \\ \mu_{2i} &= \gamma_0 + \gamma_1 \texttt{Age}_i + \gamma_2 \texttt{Fem}_i + \gamma_3 \texttt{Educ}_i + \gamma_4 \texttt{Blhisp}_i + \gamma_5 \texttt{Totchr}_i + \gamma_6 \texttt{Ins}_i + \\ &+ \gamma_7 \texttt{Income}_i, \end{split}$$

 $\log \sigma_i = \lambda_0 + \lambda_1 \texttt{Age}_i + \lambda_2 \texttt{Totchr}_i + \lambda_3 \texttt{Ins}_i,$ 

 $\operatorname{arctanh} \rho_i = \kappa_0 + \kappa_1 \operatorname{Fem}_i + \kappa_2 \operatorname{Totchr}_i,$ 

for i = 1, ..., 3328. The primary equation has the same covariates of the selection equation with the additional covariate Income, so that exclusion restriction is in force. To select the covariates for the sample selection bias and dispersion parameters, we fitted a full model by incorporating all the covariates in the regression structures and then checked those individually non-significant for  $\rho_i$  and  $\sigma_i$ . We performed a likelihood ratio (LR) test to check the joint removal of these covariates; the null hypothesis states that the coefficients associated with such covariates are null. We obtain the LR statistics equal to 4.207 with associated *p*-value 0.756. Therefore, we have statistical evidence that the aforementioned reduced model is preferable to the full model (we cannot reject the null hypothesis).

In Table 4, we present the summary of the fits of the classic Heckman and generalized Heckman models. From this table, we can observe that the covariates **Fem** and **Totchr** are significant to explain the sample selection bias, by using any significance level. We perform a likelihood (LR) test for checking the absence ( $H_0$ :  $\kappa_0 = \kappa_1 = \kappa_2 = 0$ ) or presence of sample selection bias. The LR statistic was 28.16 with corresponding *p*-value equal to  $3 \times 10^{-6}$ . Therefore, our proposed generalized Heckman model is able to detect the presence of sample selection bias even under normal assumption. We also performed the gradient and Wald tests, which confirmed this conclusion.

Further, the covariates age, totcr, and ins were significant for the dispersion parameter. For the selection equation, the covariate Income is not significant (significance level at 5%) based on the generalized Heckman model in contrast with the classic Heckman model. Anyway, it is important to keep it in order to satisfy the exclusion restriction. Regarding the primary equation, we observe that the covariate Ins is only significant under the classic Heckman model. Another interesting point is that Educ is strongly significant for the primary equation under our proposed generalized Heckman model.

We conclude this application by checking the goodness-of-fit of the fitted generalized sample selection Heckman model. In Figure 6, we provide the QQ-plot of the score residuals given in (3.12) with simulated envelopes, and also a Cook dis-

**Table 4:** Summary fits of the classic Heckman model (HM) and generalized Heckman model (GHM). The GHM summary fit contains estimates with their respective standard errors, z-value, *p*-value, and inferior and superior bounds of the 95% confidence interval.

Selection Equation												
covariates	HM-est.	p-value	GHM-est.	stand. error	z-value	p-value	Inf.	Sup.				
Intercept	-0.676	0.000	-0.590	0.187	-3.162	0.002	-0.956	-0.224				
Age	0.088	0.001	0.086	0.026	3.260	0.001	0.034	0.138				
Fem	0.663	0.000	0.630	0.060	10.544	0.000	0.513	0.747				
Educ	0.062	0.000	0.057	0.011	4.984	0.000	0.035	0.079				
Blhisp	-0.364	0.000	-0.337	0.060	-5.644	0.000	-0.454	-0.220				
Totchr	0.797	0.000	0.758	0.069	11.043	0.000	0.624	0.893				
Ins	0.170	0.007	0.173	0.061	2.825	0.005	0.053	0.293				
Income	0.003	0.040	0.002	0.001	1.837	0.066	0.000	0.005				
Primary Equation												
covariates	HM-est.	p-value	GHM-est.	stand. error	z-value	p-value	Inf.	Sup.				
Intercept	5.044	0.000	5.704	0.193	29.553	0.000	5.326	6.082				
Age	0.212	0.000	0.184	0.023	7.848	0.000	0.138	0.230				
Fem	0.348	0.000	0.250	0.059	4.252	0.000	0.135	0.365				
Educ	0.019	0.076	0.001	0.010	0.129	0.897	-0.019	0.021				
Blhisp	-0.219	0.000	-0.128	0.058	-2.221	0.026	-0.242	-0.015				
Totchr	0.540	0.000	0.431	0.031	14.113	0.000	0.371	0.490				
Ins	-0.030	0.557	-0.103	0.051	-1.999	0.046	-0.203	-0.002				
			Dispers	ion Paramete	r							
covariates	HM-est.	p-value	GHM-est.	stand. error	z-value	p-value	Inf.	Sup.				
Intercept	1.271		0.508	0.057	8.853	0.000	0.396	0.621				
Age		-	-0.025	0.013	-1.986	0.047	-0.049	0.000				
Totchr	-	-	-0.105	0.019	-5.475	0.000	-0.142	-0.067				
Ins		_	-0.107	0.028	-3.864	0.000	-0.161	-0.053				
		Sa	mple Select	tion Bias Par	rameter							
covariates	HM-est.	p-value	GHM-est.	stand. error	z-value	p-value	Inf.	Sup.				
Intercept	-0.131	0.375	-0.648	0.114	-5.668	0.000	-0.872	-0.424				
Fem	_	_	-0.403	0.136	-2.973	0.003	-0.669	-0.137				
Totchr	_	_	-0.438	0.186	-2.353	0.019	-0.803	-0.073				



**Figure 6:** QQ-plot and its simulated envelope for the score residuals (left) and index plot of the GCD (right) for the generalized Heckman model for the medical expenditure panel survey data.

tance plot for detecting global influence. Based on this last plot, we do not detect any outlier observation, since all points are below the reference line 2p/n = 0.013. Anyway, we investigate if the highlighted point #2602 (above the line  $8 \times 10^{-4}$ ) is influential. We fitted our model by removing this observation and no changes either on the parameter estimates or different conclusions about the significance of covariates were obtained. Regarding the QQ-plot of the score residuals, we observe a great performance of our model, since 96% of the points are inside of the envelope. This confirms that the normal assumption for this particular dataset is adequate and that our generalized Heckman model is suitable for the MEPS data analysis.

# 6. Concluding Remarks

In this paper, a generalization of the Heckman model was proposed by allowing both sample selection bias and dispersion parameters to vary across covariates. We showed that the proposed model satisfies certain regularity conditions that ensure consistency and asymptotic normality of the maximum likelihood estimators. Furthermore, a proper score residual for sample selection models was proposed. These finding are new contributions on this topic. The MEPS data analysis based on the generalized Heckman model showed that the normal assumption for the data is suitable in contrast with existing findings in the literature. Future research should address (i) generalization of other sample selection models such as Student-t and skew-normal to allow varying sample selection models; and (iii) deeper study of influence analysis. An R package for fitting our proposed generalized Heckman, Student-t, and skew-normal models has been developed and available at the GitHub https://fsbmat-ufv.github.io/ssmodels/.

# Acknowledgments

We would like to thank the anonymous Referee for the insightful and valuable comments which contributed to an improvement of the paper. W. Barreto-Souza and M. Genton would like to acknowledge support for their work by the *King Abdullah University of Science and Technology.* 

# Appendix

Proof of Theorem 1. We here show the results for the derivatives with respect to  $\beta$ . The results involving the other derivatives follow similarly and therefore they are omitted.

For  $i = 1, \ldots, n$  and  $j = 1, \ldots, p$ , we have that

$$\frac{\partial \ell_i}{\partial \beta_j} = \left\{ -\frac{\rho_i}{\sqrt{1-\rho_i^2}} \frac{\phi(\zeta_i)}{\Phi(\zeta_i)} + z_i \right\} x_{ij} u_i / \sigma_i.$$
(6.17)

By using basic properties of conditional expectation, it follows that  $E\left(\frac{\partial \ell_i}{\partial \beta_j}\right) = E\left[E\left(\frac{\partial \ell_i}{\partial \beta_j}|U_i\right)\right]$  and it is immediate that  $E\left(\frac{\partial \ell_i}{\partial \beta_j}|U_i=0\right) = 0$ . Let us now compute the conditional expectations involved in  $E\left(\frac{\partial \ell_i}{\partial \beta_j}|U_i=1\right)$ .

Here it is worth to remember the notations  $z_i = \frac{y_i - \mu_{1i}}{\sigma_i}$  and  $\zeta_i = \frac{\mu_{2i} + z_i \rho_i}{\sqrt{1 - \rho_i^2}}$ , for i = 1, ..., n. We now use the conditional density function given in (2.8) to obtain that

$$E\left(\frac{\phi(\zeta_i)}{\Phi(\zeta_i)}|U_i=1\right) = \frac{1}{\sigma_i\Phi(\mu_{2i})} \int_{-\infty}^{\infty} \phi(\zeta_i)\phi(z_i)dy_i = \frac{1}{2\pi\sigma_i\Phi(\mu_{2i})} \int_{-\infty}^{\infty} \exp\{-(\zeta_i^2+z_i^2)/2\}dy_i$$
$$= \frac{e^{-\mu_{2i}^2/2}}{2\pi\sigma_i\Phi(\mu_{2i})} \int_{-\infty}^{\infty} \exp\{-\frac{(y_i-\mu_{1i}+\sigma_i\mu_{2i}\rho_i)^2}{2\sigma_i^2(1-\rho_i^2)}\}dy_i = \sqrt{1-\rho_i^2}\frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})},$$

where the last equality follows by identifying a normal kernel in the integral. On the other hand, we use the fact that  $Z_i$  given  $U_i = 1$  has mean equal to  $\mu_{1i} + \rho_i \sigma_i \frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})}$ 

(see Supplementary Material for more details) and get

$$E\left(Z_{i}|U_{i}=1\right) = -\frac{\mu_{1i}}{\sigma_{i}} + \frac{1}{\sigma_{i}}E(Y_{i}|U_{i}=1) = -\frac{\mu_{1i}}{\sigma_{i}} + \frac{1}{\sigma_{i}}\left(\mu_{1i} + \rho_{i}\sigma_{i}\frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})}\right) = \rho_{i}\frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})}.$$

With the results above, we obtain that  $E\left(\frac{\phi(\zeta_i)}{\Phi(\zeta_i)}|U_i\right) = 0$  almost surely and therefore  $E\left(\frac{\partial \ell_i}{\partial \beta_j}\right) = 0$  for  $j = 1, \dots, p$ .

We now concentrate our attention to prove the identity stated in the theorem. It follows that

$$\frac{\partial^2 \ell_i}{\partial \beta_j \beta_l} = -\left\{ \frac{\rho_i^2}{1 - \rho_i^2} \left[ \zeta_i \frac{\phi(\zeta_i)}{\Phi(\zeta_i)} + \frac{\phi^2(\zeta_i)}{\Phi^2(\zeta_i)} \right] + 1 \right\} \frac{x_{ij} x_{il}}{\sigma_i^2} u_i$$

and

$$\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_l} = \left\{ z_i^2 - 2z_i \frac{\rho_i}{\sqrt{1 - \rho_i^2}} \frac{\phi(\zeta_i)}{\Phi(\zeta_i)} + \frac{\rho_i^2}{1 - \rho_i^2} \frac{\phi^2(\zeta_i)}{\Phi^2(\zeta_i)} \right\} \frac{x_{ij} x_{il}}{\sigma_i^2} u_i,$$

where we have used that  $u_i^2 = u_i$  (since  $u_i \in \{0, 1\}$ ) in the last equality. It is immediate that  $E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \beta_l} | U_i = 0\right) = -E\left(\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_l} | U_i = 0\right) = 0.$ 

Following in a similar way as before, after some algebra we obtain that

$$E\left(\zeta_{i}\frac{\phi(\zeta_{i})}{\Phi(\zeta_{i})}|U_{i}=1\right) = -\mu_{2i}(1-\rho_{i}^{2})\frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})} \quad \text{and} \quad E\left(Z_{i}^{2}|U_{i}=1\right) = 1-\mu_{2i}\rho_{i}^{2}\frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})}$$

By combining these results, we have that

$$-E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \beta_l} | U_i = 1\right) = E\left(\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_l} | U_i = 1\right)$$
$$= \left\{ 1 + \mu_{2i} \rho_i^2 \frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})} + \frac{\rho_i^2}{1 - \rho_i^2} E\left(\frac{\phi^2(\zeta_i)}{\Phi^2(\zeta_i)} | U_i = 1\right) \right\} \frac{x_{ij} x_{il}}{\sigma_i^2}.$$

Since the conditional expectations coincide, the marginal expectations also coincide so giving the desired result.  $\Box$ 

*Proof of Theorem 2.* Conditions (C1)–(C3) and Theorem 1 give us the consistency of the MLEs. To establish the asymptotic normality of the estimators, we need to show that the third derivatives of the log-likelihood function are bounded by integrable functions not depending on the parameters.

We will show here that this is possible for the derivatives involving the  $\beta$ 's. The other cases follow in a similar way as discussed in the proof of Theorem 1 and therefore they are omitted.

By computing the third derivatives with respect to the  $\beta$ 's and using the triangular inequality, we have that

$$\left| \frac{\partial^{3} \ell_{i}}{\partial \beta_{j} \partial \beta_{l} \partial \beta_{k}} \right| \leq \frac{\rho_{i}^{2}}{\sigma_{i}^{3} (1 - \rho_{i}^{2})^{3/2}} \left\{ (1 + z_{i}^{2}) \frac{\phi(\zeta_{i})}{\Phi(\zeta_{i})} + \zeta_{i}^{2} \frac{\phi^{2}(\zeta_{i})}{\Phi^{2}(\zeta_{i})} + 2\zeta_{i} \frac{\phi^{2}(\zeta_{i})}{\Phi^{2}(\zeta_{i})} + 2\frac{\phi^{2}(\zeta_{i})}{\Phi^{3}(\zeta_{i})} \right\} x_{ij} x_{il} x_{ik}$$
$$\equiv g_{i}(\boldsymbol{\theta}) \leq g_{i}(\boldsymbol{\theta}^{*}),$$

for j, l, k = 1, ..., p, where  $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} g_i(\boldsymbol{\theta})$ , which is well-defined due to Assumption (C1).

We now need to show that the expectations of the terms in  $g_i(\boldsymbol{\theta}^*)$  are finite.

Let us show that  $E_{\theta_0}\left(\zeta_i^{*2}\frac{\phi^2(\zeta_i^*)}{\Phi^2(\zeta_i^*)}\right) < \infty$ , where  $E_{\theta_0}(\cdot)$  denotes the expectation with respect to the true parameter vector value  $\theta_0$  and  $\zeta_i^*$  is defined as  $\zeta_i$  by replacing  $\theta$  by  $\theta^*$ . The proofs for the remaining terms follow from this one or in a similar way.

For  $\zeta_i^* \leq \sqrt{2}$ , it follows that  $\phi^2(\zeta_i^*)/\Phi^2(\zeta_i^*) \leq \{2\pi\Phi^2(\sqrt{2})\}^{-1}$ . Now, consider  $\zeta_i^* > \sqrt{2}$ . Theorem 1.2.6 from Durrett (2019) gives us the following inequality for x > 0:

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)e^{-x^2/2} \le \int_x^\infty e^{-y^2/2}dy.$$

Using this inequality and under  $\zeta_i^* > \sqrt{2}$ , we obtain that  $\frac{\phi^2(\zeta_i^*)}{\Phi^2(\zeta_i^*)} \le \frac{\zeta_i^{*3}}{\zeta_i^{*2} - 1} \le \zeta_i^*$ . These results imply that

$$E_{\theta_{0}}\left(\zeta_{i}^{*2}\frac{\phi^{2}(\zeta_{i}^{*})}{\Phi^{2}(\zeta_{i}^{*})}\right) = E_{\theta_{0}}\left(\zeta_{i}^{*2}\frac{\phi^{2}(\zeta_{i}^{*})}{\Phi^{2}(\zeta_{i}^{*})}I\left\{\zeta_{i}^{*} \leq \sqrt{2}\right\}\right) + E_{\theta_{0}}\left(\zeta_{i}^{*2}\frac{\phi^{2}(\zeta_{i}^{*})}{\Phi^{2}(\zeta_{i}^{*})}I\left\{\zeta_{i}^{*} > \sqrt{2}\right\}\right)$$

$$\leq \sqrt{2}\left\{2\pi\Phi^{2}(\sqrt{2})\right\}^{-1} + E_{\theta_{0}}\left(|\zeta_{i}^{*}|^{3}\right) < \infty,$$

with  $E_{\theta_0}(|\zeta_i^*|^3) < \infty$  being proved in the same lines that the first two moments presented in the Supplementary Material, which completes the proof of the desired result.  $\Box$ 

## References

Arabmazar, A. and P. Schmidt (1981). Further evidence on the robustness of the tobit estimator to heteroskedasticity. Journal of Econometrics 17, 253–258.

Atkinson, A. C. (1985). Plots, Transformations, and Regression. Oxford: Oxford University Press.

- Azzalini, A., H. M. Kim, and H. J. Kim (2019). Sample selection models for discrete and other non-gaussian response variables. *Statistical Methods and Applications* 19, 27–56.
- Cameron, C. A. and P. K. Trivedi (2009). *Microeconometrics Using Stata* (Revised ed.). College Station: TX: Stata press.
- Chib, S., E. Greenberg, and I. Jeliazkov (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. Journal of Computational and Graphical Statistics 18(2), 321–348.
- Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics 19, 15–18.
- Cook, R. D. and S. Weisberg (1982). Residuals and Influence in Regression. New York: Chapman and Hall.
- Cox, D. R. and D. V. Hinkley (1979). Theoretical Statistics. Chapman and Hall/CRC.
- Donald, S. G. (1995). Two-step estimation of heteroskedastic sample selection models. *Journal of Econo*metrics 65, 347–380.
- Durrett, R. (2019). Probability: Theory and Examples (5th ed.). Cambridge University Press.
- Enders, C. K. (2010). Applied Missing Data Analysis (1 ed.). New York, NY: Guilford Press.
- Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Annals of Statistics 13, 342–368.
- Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. Econometrica 42, 679-694.
- 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.
- Hurd, M. (1979). Estimation in truncated samples when there is heteroscedasticity. *Journal of Economet*rics 11, 247–258.

- Kim, H., T. Roh, and T. Choi (2019). Bayesian analysis of semiparametric Bernstein polynomial regression models for data with sample selection. *Statistics* 53, 1–30.
- Lai, H. P. and W. J. Tsay (2018). Maximum simulated likelihood estimation of the panel sample selection model. *Econometric Reviews* 37(7), 744–759.
- Leung, S. F. and S. Yu (1996). On the choice between sample selection and two-part models. Journal of Econometrics 72, 197–229.
- Leung, S. F. and S. Yu (2000). Collinearity and two-step estimation of sample selection models: Problems, origins, and remedies. *Computational Economics* 15, 173–199.
- Marchenko, Y. V. and M. G. Genton (2012). A Heckman selection-t model. Journal of the American Statistical Association 107, 304–317.
- Marra, G. and K. Wyszynski (2016). Semi-parametric copula sample selection models for count responses. Computational Statistics and Data Analysis 104, 110–129.
- Mu, B. and Z. Zhang (2018). Identification and estimation of heteroscedastic binary choice models with endogenous dummy regressors. *Econometrics Journal* 21(2), 218–246.
- Nawata, K. (1993). A note on the estimation of models with sample-selection biases. *Economics Letters 42*, 15–24.
- Nawata, K. (1994). Estimation of sample selection bias models by the maximum likelihood estimator and Heckman's two-step estimator. *Economics Letters* 45, 33–40.
- Nawata, K. (2007). A Monte Carlo analysis of the type II Tobit maximum likelihood estimator when the true model is the type I Tobit model. *Economics Bulletin 3*, 1–10.
- Nawata, K. and M. Kimura (2017). Evaluation of medical costs of kidney diseases and risk factors in Japan. *Health 9*, 1734–1749.

- Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal 12*, 217–229.
- Ogundimu, E. O. and J. L. Hutton (2016). A sample selection model with skew-normal distribution. Scandinavian Journal of Statistics 43, 172–190.
- Powell, J. L. (1986). Symmetrically trimmed least squares estimation for tobit models. *Econometrica* 54, 1435–1460.
- R Core Team (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- Wiesenfarth, M. and T. Kneib (2010). Bayesian geoadditive sample selection models. Journal of the Royal Statistical Society – Series C 59(3), 381–404.
- Wojtys, M., G. Marra, and R. Radice (2018). Copula based generalized additive models for location, scale and shape with non-random sample selection. *Computational Statistics and Data Analysis 127*, 1–14.
- Wyszynski, K. and G. Marra (2018). Sample selection models for count data in R. Computational Statistics 33, 1385–1412.
- Xie, F.-C. and B.-C. Wei (2007). Diagnostics analysis for log-birnbaum-saunders regression models. Computational Statistics and Data Analysis 51(9), 4692–4706.
- Yamagata, T. and C. Orme (2005). On testing sample selection bias under the multicollinearity problem. *Econometric Reviews* 24, 467–481.
- Zhelonkin, M., M. G. Genton, and E. Ronchetti (2014). R package ssmrob: Robust estimation and inference in sample selection models. CRAN. R package version 0.4.
- Zhelonkin, M., M. G. Genton, and E. Ronchetti (2016). Robust inference in sample selection models. Journal of the Royal Statistical Society – Series B 78, 805–827.