Statistica Sinica **32** (2022), 1911-1938
doi:https://doi.org/10.5705/ss.202021.0068

A GENERALIZED HECKMAN MODEL WITH VARYING SAMPLE SELECTION BIAS AND DISPERSION PARAMETERS

Fernando de Souza Bastos[1,2], Wagner Barreto-Souza[2,3] and Marc G. Genton[3]

[1] *Universidade Federal de Viçosa,* [2] *Universidade Federal de Minas Gerais and* [3] *King Abdullah University of Science and Technology*

Abstract: Many proposals have emerged as alternatives to the Heckman selection model, mainly to address the non-robustness of its normal assumption, which is often illustrated using the 2001 Medical Expenditure Panel Survey data. 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 explain the sample selection bias phenomenon, rather than to simply form conclusions about its presence. Furthermore, our approach may attenuate the non-identifiability and multicollinearity problems faced by 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 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 a 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 the findings from prior studies.

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, using a bivariate normal distribution to relate the outcome of interest and a selection rule. A semiparametric alternative to this model, known

Corresponding author: Wagner Barreto-Souza, Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. E-mail: wagner.barretosouza@kaust.edu.sa.

as Heckman's two-step method, was proposed by Heckman (1979) to handle the non-robustness of the normal distribution in the presence of outliers.

The most discussed problem related to the Heckman model is its sensitivity to the assumption of normally distributed errors. Misspecification of the error distribution leads to inconsistent maximum likelihood estimators (MLEs), yielding biased estimates (Lai and Tsay (2018)). On the other hand, when the error terms are correctly specified, an estimation using a maximum likelihood or using procedures based on a 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 comprise large databases, in which heterogeneity is common. 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 heteroskedasticity than by an incorrect distribution of the error terms (Powell (1986)).

Nevertheless, even though there is a large body of recent research on sample selection models, few studies have been attempted to correct or minimize the impact of heteroskedasticity. This is one of the aims of this study. Chib, Greenberg and Jeliazkov (2009) proposed a semiparametric model for data with a sample selection bias. They considered nonparametric functions in their model, which allowed great flexibility in the way the covariates affect the response variables. They still presented a Bayesian method for the analysis of such models. Subsequently, Wiesenfarth and Kneib (2010) introduced 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, Genton and Ronchetti (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 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 the outcome equations, and flexible covariate effects. The modeling of zero-inflated count data with a sample selection bias is discussed by Wyszynski and Marra (2018). Mu and Zhang (2018) considered the semiparametric identification and estimation of a heteroskedastic binary choice model with endogenous dummy regressors, and without assuming any parametric restrictions on the distribution of the error term. This yields general multiplicative heteroskedasticity in both the selection and the 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, Kim and Kim (2019).

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

In the aforementioned works, the solution to dealing with departures from normality for continuous outcomes is to assume robust alternatives, such as the Student-t or skew-normal distributions. Another common approach is to consider nonparametric structures for the density of the error terms. Our proposal adopts a different approach, one that has not yet been explored in the literature.

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 explain the sample selection bias phenomenon, rather than to solely make conclusions about its presence. Note that our methodology can be straightforwardly adapted for existing sample selection methods, such as those proposed by Marchenko and Genton (2012), Ogundimu and Hutton (2016), and

Wyszynski and Marra (2018). In addition to the above, this study makes the following contributions to the literature:

- We explore the inferential aspects of the 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 the sample selection bias and the dispersion parameters may attenuate the non-identifiability and multicollinearity problems faced by existing sample selection models.

- A proper residual for sample selection models is proposed as a byproduct of our asymptotic analysis. This is another relevant contribution of our study, because 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 study 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 the findings in the literature. Moreover, we identify which covariates explain the presence of a sample selection bias in this important data set.

The remainder of this paper is organized as follows. In Section 2, we define the generalized Heckman (GH) sample selection model, and discuss the estimation of the parameters using the maximum likelihood method. Furthermore, diagnostics tools and residual analysis are discussed. Section 3 shows that the GH model satisfies regularity conditions that ensure consistency and asymptotic normality of the MLEs. In Section 4, we present Monte Carlo simulation results to evaluate the performance of the MLEs of our proposed model, and to check the behavior of other existing methodologies under misspecification. In Section 5, we apply our GH model to 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. GH 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 $\mu_{1i} = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, $\mu_{2i} = \boldsymbol{w}_i^\top \boldsymbol{\gamma}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top \in \mathbb{R}^q$ are vectors of unknown parameters with associated covariate vectors $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$, respectively, for $i = 1, \ldots, n$, and $\{(\epsilon_{1i}, \epsilon_{2i})\}_{i=1}^n$ is a sequence of independent bivariate normal random vectors. More specifically, we suppose that

$$\begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix} \sim \mathcal{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_i \sigma_i \\ \rho_i \sigma_i & 1 \end{pmatrix} \right], \tag{2.3}$$

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

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

respectively, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^\top \in \mathbb{R}^r$ and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_s)^\top \in \mathbb{R}^s$ are parameter vectors with associated covariate vectors $\boldsymbol{v}_i$ and $\boldsymbol{e}_i$, respectively, for $i = 1, \ldots, n$. 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 $Y_{2i}^* > 0$, and the variable Y_{2i}^* is latent. We only know if Y_{2i}^* is greater or less than zero. Equation (2.1) is the primary interest equation, and equation (2.2) represents the selection equation. In practice, we observe the variables

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

$$Y_i = Y_{1i}^* U_i, \tag{2.6}$$

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

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

The mixed distribution of Y_i is composed of the discrete component

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

and a continuous part given by the conditional density function

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

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and cumulative distribution functions, respectively, of the standard normal distribution. Additional details 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

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

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

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

Note that the maximum likelihood method for the classic Heckman model may suffer from multicollinearity problems when the selection equation has the same covariates as the regression equation. Other issues include no convergence of standard algorithms to the estimate parameters, multiple local maxima, and the MLE not existing in some cases; see, for instance, 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 following conditional expectation (for constant σ and ρ):

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

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

Even the two-step estimation can suffer from multicollinearity 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 by considering covariates in both the sample selection and the dispersion parameters. In our case, the conditional expectation $E(Y_i|U_i = 1)$ assumes the form

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

Owing to the nonlinearity of the functions $\tanh(\cdot)$ and $\exp(\cdot)$, we identify the parameters even when $\Lambda(\cdot)$ behaves like a linear function. Therefore, ordinary least squares 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 the sample selection bias and the dispersion parameters, consider just one common $\mathbb{R}$-valued covariate for all the regression structures, say x, with the intercepts equal to zero and the remaining parameters equal to one. 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 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 nonlinear behavior of the functions R_1 and R_2, in contrast to that of R_3 and Λ, where linear behavior is observed for certain ranges. In the $R_2(\cdot)$ case, ρ is constant. Even if R_2 is not close to a

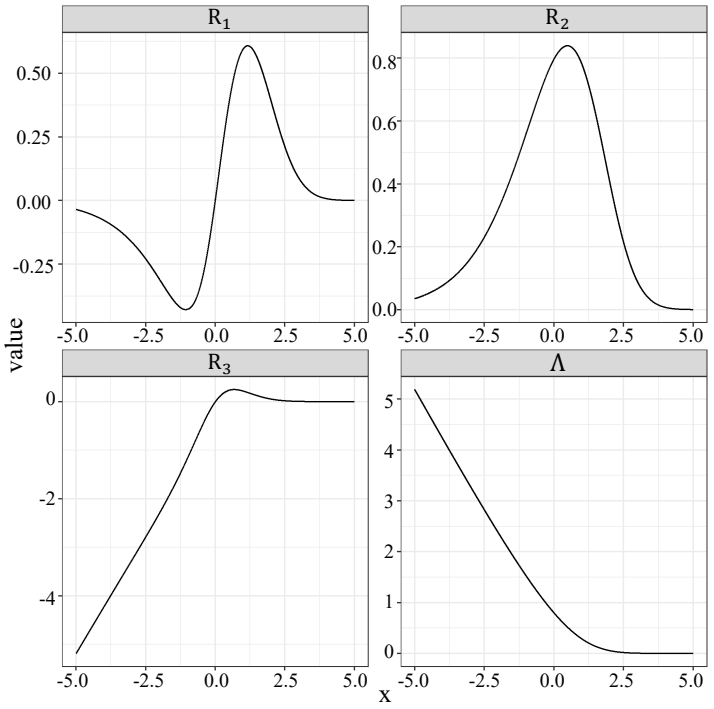

Figure 1. Plots of the functions $R_1(x)$, $R_2(x)$, $R_3(x)$, and $\Lambda(x)$ for $x \in (-5, 5)$.

linear function, not including covariates for ρ implies that this parameter can be absorbed into Λ when behaving linearly, so causing both non-identifiability and multicollinearity problems. Thus, we conclude that the best approach is to consider regression structures for both the sample selection bias and the dispersion parameters to avoid these issues.

We now discuss diagnostic techniques to detect observations that could exercise some influence on the parameter estimates or inference in general. Next, for the GH 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 GCD is defined by

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

where $\boldsymbol{M}$ is a nonnegative definite matrix that measures the weighted combination of the elements for the difference $\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}$, and $\widehat{\boldsymbol{\theta}}_{(i)}$ is the MLE of $\boldsymbol{\theta}$ when removing the ith observation. Many choices for $\boldsymbol{M}$ were considered by Cook and Weisberg (1982). We use the inverse variance-covariance matrix $\boldsymbol{M} = -\ddot{\mathcal{L}}(\widehat{\boldsymbol{\theta}})^{-1}$. To determine whether the ith case is potentially influential on inference about $\boldsymbol{\theta}$, we check whether 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. For the residual analysis, in the next section, we propose a proper residual for sample selection models, which is one of the aims of this study.

3. Asymptotic Properties and Score Residuals

Our aim in this section is to show that under some conditions, our proposed GH sample selection model satisfies the regularity conditions stated by Cox and Hinkley (1979). As a result, the MLEs discussed in the previous section are consistent and asymptotically normal distributed. As a byproduct 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 $\boldsymbol{\Theta}$ be the parameter space and $\ell_i(\boldsymbol{\theta}) = u_i \{\log \Phi(\zeta_i) + \log \phi(z_i) - \log \sigma_i\} + (1-u_i) \log \Phi(-\mu_{2i})$ be the contribution of the ith observation to the log-likelihood function, where ζ_i retains its definition from the previous section, for $i = 1, \ldots, n$.

Theorem 1. *The score function associated with the GH 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 (A.1), we define the ordinary score residual by $s_i = z_i - (\rho_i/\sqrt{1-\rho_i^2})(\phi(\zeta_i)/\Phi(\zeta_i))$ for the non-censored observations (where $u_i = 1$) and the standardized score residual by

$$
\begin{aligned}
S_i &= \frac{s_i - E(s_i|U_i = 1)}{\sqrt{\mathrm{Var}(s_i|U_i = 1)}} \\
&= \left(z_i - \frac{\rho_i}{\sqrt{1-\rho_i^2}} \frac{\phi(\zeta_i)}{\Phi(\zeta_i)} \right) \Bigg/ \sqrt{1 + \mu_{2i}\rho_i^2 \frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})} + \frac{\rho_i^2}{1-\rho_i^2}\Psi_i},
\end{aligned}
$$

where $\Psi_i = E(\phi^2(\zeta_i)/\Phi^2(\zeta_i)\big|U_i = 1) = (1/\sigma_i\Phi(\mu_{2i})) \int_{-\infty}^{\infty} \phi(z_i)\phi(\zeta_i)/\Phi(\zeta_i) dy_i$, 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

$$S_i^* = \frac{s_i - E(s_i)}{\sqrt{\mathrm{Var}(s_i)}} = \left(z_i - \frac{\rho_i}{\sqrt{1-\rho_i^2}}\frac{\phi(\zeta_i)}{\Phi(\zeta_i)}\right) \Bigg/ \sqrt{\Phi(\mu_{2i})\left(1 + \mu_{2i}\rho_i^2\frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})} + \frac{\rho_i^2}{1-\rho_i^2}\Psi_i\right)}, \quad (3.1)$$

for $i = 1, \ldots, n$. In practice, we replace the unknown parameters by their MLEs. The evaluation of the goodness-of-fit of our proposed GH model is performed using a score residual analysis. Based on this approach, discrepant observations are identified, in addition to making it 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 using 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, thus indicating a good fit. Otherwise, residuals outside the envelopes are possible outliers or indicate that the model is not properly specified. We apply the proposed score residual (3.1) to the MEPS data analysis. As shown, the residual analysis indicates that the normal assumption for the data is suitable, in contrast to the non-robustness of the Heckman model mentioned in the literature on sample selection models.

We now establish the consistency and asymptotic normality of the MLEs for our proposed GH model. For this, we need to assume some usual regularity conditions.

(C1) The parameter space $\boldsymbol{\Theta}$ is closed and compact, and the true parameter value, say $\boldsymbol{\theta}_0$, is an interior point of $\boldsymbol{\Theta}$.

(C2) The covariates are a sequence of independent and identically distributed (i.i.d.) 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 $E\left(\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{F}_n\|\right) < \infty$, where $\|\cdot\|$ is the Frobenius norm.

Remark 1. Conditions (C2) and (C3) enable us to apply a multivariate central limit theorem for i.i.d. random vectors to establish the asymptotic normality of the MLEs. These conditions are discussed, for instance, in Fahrmeir and Kaufmann (1985).

Theorem 2. *Under Conditions* (C1)–(C3)*, the MLE* $\widehat{\boldsymbol{\theta}}$ *of* $\boldsymbol{\theta}$ *for the GH model is consistent and satisfies the weak convergence:* $\mathbf{F}_n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$*, 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 normally distributed, which is a new finding of this study.

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 MLEs under the GH, classic Heckman, Heckman-Skew, and Heckman-t models when the assumption of either a 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, (I) with the 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 specify the following, respectively: (III) constant dispersion and varying correlation; (IV) varying dispersion and constant correlation; (V) both constant dispersion and constant correlation. To evaluate the sensitivity in the parameter estimation of the selection models at high censoring, in Scenario 6, we simulated from the GH model while (VI) varying both the sample selection bias and the dispersion parameters, with an average censorship of 50%.

Scenario 1 evaluates the performance of the GH model and compares it with that of its competitors when the assumption of a constant sample selection bias parameter and constant dispersion is not satisfied. Scenario 2 shows that despite the absence of the exclusion restriction, our model can yield satisfactory parameter estimates. Scenarios 3 and 4 justify the importance of using covariates to model the correlation and dispersion parameters, respectively. Scenario 5 illustrates some of the problems that the GH model can face, as in the case of the classic Heckman model. Finally, Scenario 6 demonstrates the sensitivity of the 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 are based on the following regression structures:

$$\mu_{1i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{4.1}$$

$$\mu_{2i} = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{3i}, \tag{4.2}$$

$$\log \sigma_i = \lambda_0 + \lambda_1 x_{1i}, \tag{4.3}$$

$$\operatorname{arctanh} \rho_i = \kappa_0 + \kappa_1 x_{1i}, \tag{4.4}$$

for $i = 1, \ldots, n$. All covariates were generated from a standard normal distribution, and were kept constant throughout the experiment. The responses were generated from the GH model according to each of the six configurations. We set the sample sizes $n = 500, 1000, 2000$ and $N = 1{,}000$ Monte Carlo replicates. Pilot simulations showed that the choice of parameters 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 Heckman-t and Heckman skew-normal models. Therefore, we developed an `R` package (to be submitted) able to fit our proposed GH model, as well as sample selection models by Marchenko and Genton (2012) and Ogundimu and Hutton (2016).

In the maximization procedure used to estimate the parameters based on the `optim` package, we consider as initial values for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, λ_0, and κ_0 the maximum likelihood estimates by the classic Heckman model. For the remaining parameters, we set $\lambda_i = 0$ and $\kappa_j = 0$, for $i = 1, \ldots, r$ and $j = 1, \ldots, s$. 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 one and two, respectively. These values were chosen after some pilot simulations.

4.2. Scenario 1: varying sample selection bias and dispersion parameters

Here, we consider (4.1)–(4.4) with $\boldsymbol{\beta} = (1.1, 0.7, 0.1)^\top$, $\boldsymbol{\gamma} = (0.9, 0.5, 1.1, 0.6)^\top$, $\boldsymbol{\lambda} = (-0.4, 0.7)^\top$, and $\boldsymbol{\kappa} = (0.3, 0.5)^\top$. The regressors are kept fixed throughout the simulations with $x_{1i}, x_{2i}, x_{3i} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, for all $i = 1, \ldots, n$.

In Table 1, we present the empirical mean and root mean squared error (RMSE) of the maximum likelihood estimates of the parameters based on the GH, classic Heckman, Student-t, and skew-normal sample selection models under Scenario 1. From this table, we observe good performance of the MLEs based on the GH model, even when 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. This suggests the consistency

of the MLEs, which is in line with our Theorem 2. On the other hand, even when the regression structures for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are correctly specified, we see that the MLEs for these parameters based on the classic Heckman, skew-normal, and 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 the 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 supported by Figures 2, 3, and 4, where box plots of the parameter estimates are presented for sample sizes $n = 500$, $n = 1{,}000$, and $n = 2{,}000$, respectively. We do not present the box plots of the estimates of γ_1, γ_2, and β_1, because they behaved similarly to other box plots.

We now provide some simulations to evaluate the size and power of the 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%.

Under the null hypothesis of an absence of a sample selection bias ($\rho_i = 0$, for all i, that is, $\kappa_0 = \kappa_1 = 0$), the likelihood ratio, gradient, and Wald tests present empirical values close to the nominal values only under the GH model. For the other models, the type-I error is inflated, indicating the presence of a selection bias. This suggests that the tests should be used with caution when testing the parameters of the sample selection models, and that some confounding may occur because of either the varying sample selection bias or the heteroskedasticity. Note that, even for the GH model, the Wald test presents a considerable inflated type-I error for $n = 500$.

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 GH, classic Heckman, Heckman-skew normal, and Heckman-t models, with significance levels at $1\%, 5\%$, and 10%. From these results, we observe that the tests under the GH model provide high power, mainly when the sample size increases. On the other hand, because the tests based on the other models do not provide correct nominal significance levels, the power of these tests is 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 MEPS available in `R` in the package

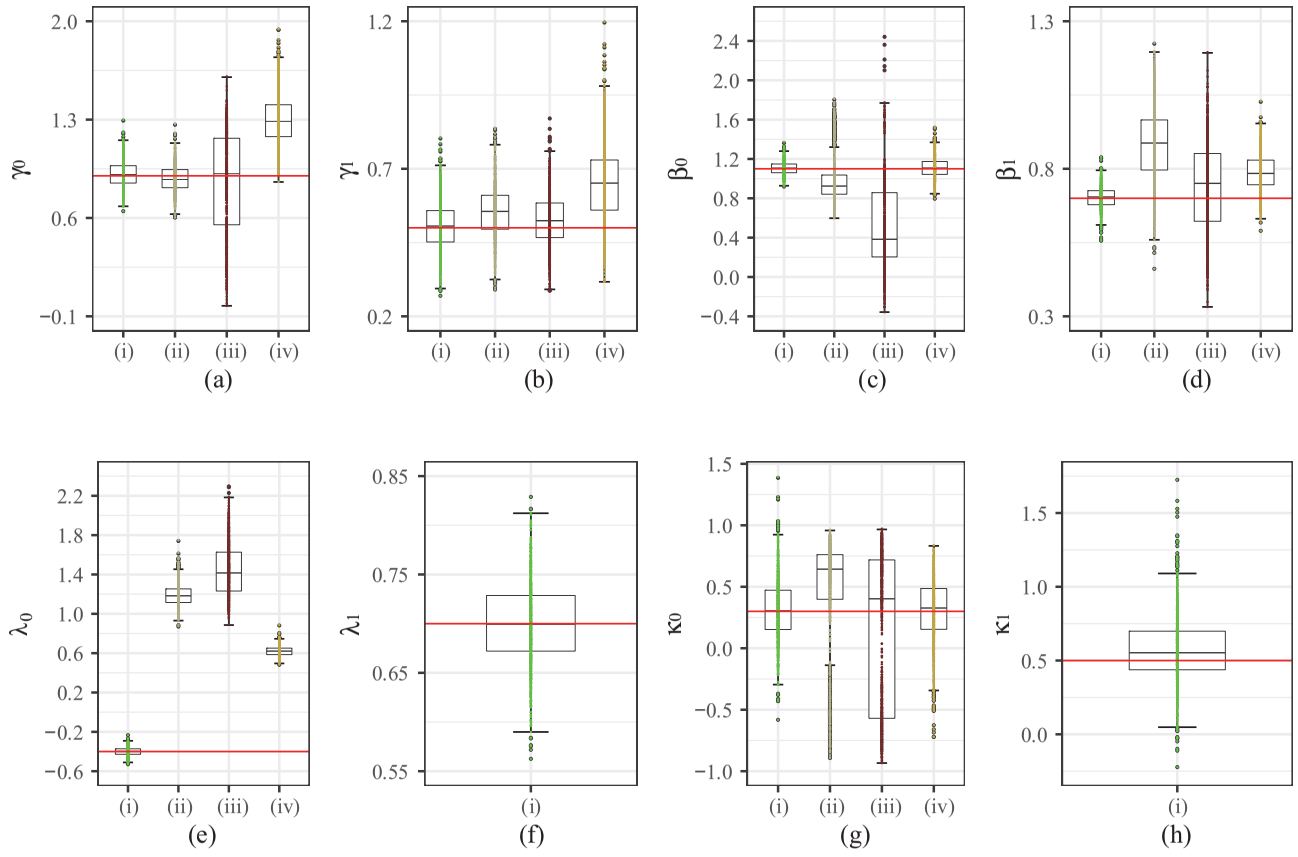

Figure 2. Box plots of the maximum likelihood estimates of the parameters (a) γ_0, (b) γ_1, (c) β_0, (d) β_1, (e) λ_0, (f) λ_1, (g) κ_0, and (h) κ_1 based on the (i) GH, (ii) classic Heckman, (iii) Heckman-Skew, and (iv) Heckman-t sample selection models. Sample size $n = 500$.

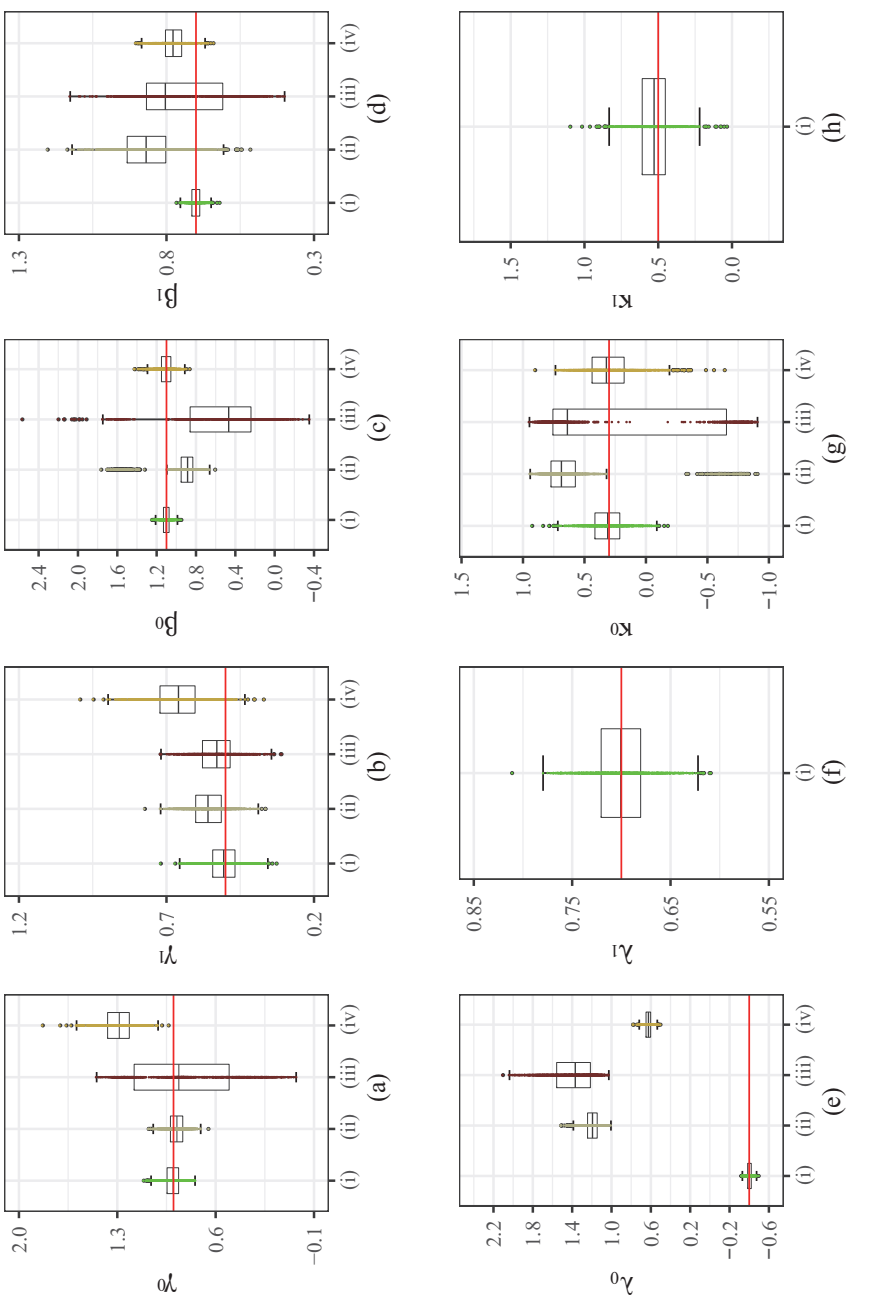

Figure 3. Box plots of the maximum likelihood estimates of the parameters (a) γ_0, (b) γ_1, (c) β_0, (d) β_1, (e) λ_0, (f) λ_1, (g) κ_0, and (h) κ_1 based on the (i) GH, (ii) classic Heckman, (iii) Heckman-Skew, and (iv) Heckman-t sample selection models. Sample size $n = 1{,}000$.

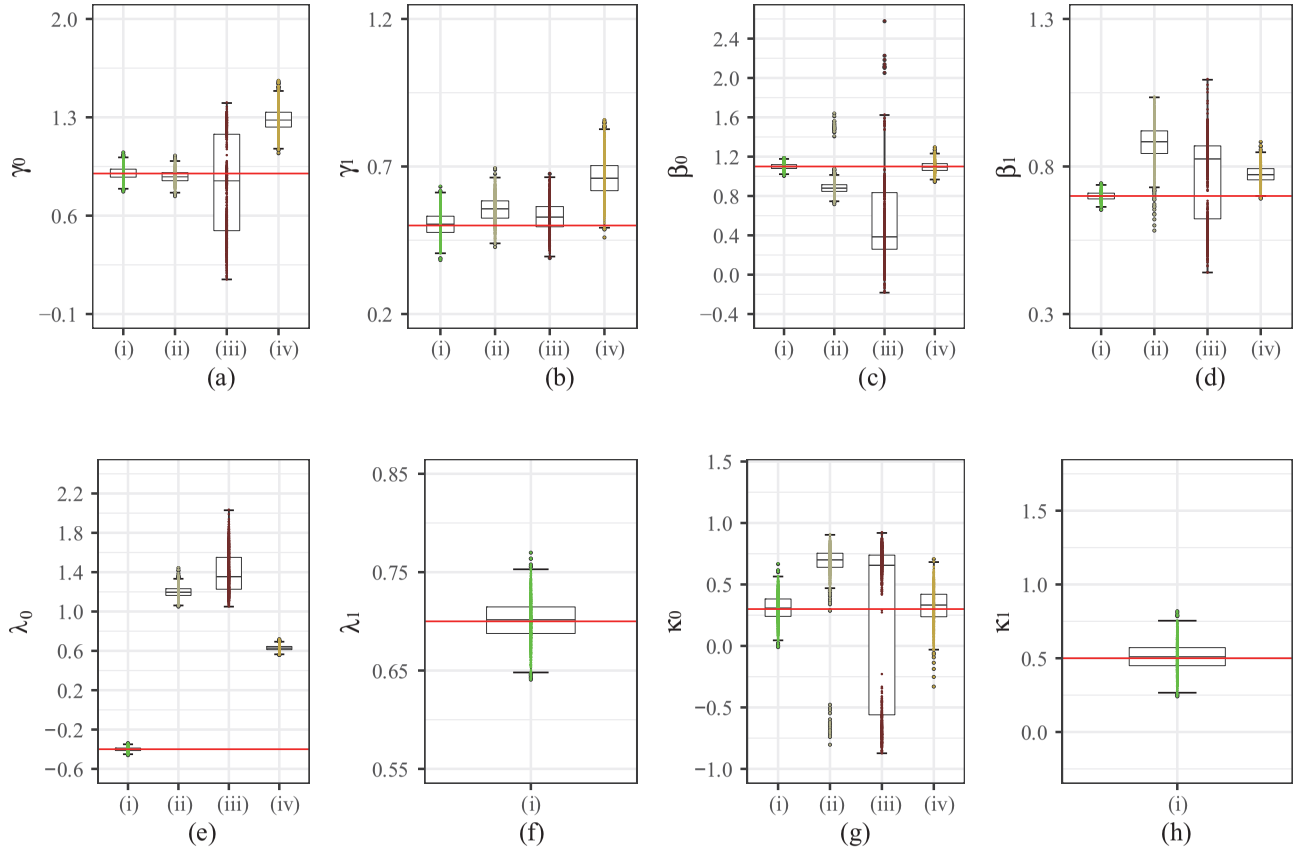

Figure 4. Box plots of the maximum likelihood estimates of the parameters (a) γ_0, (b) γ_1, (c) β_0, (d) β_1, (e) λ_0, (f) λ_1, (g) κ_0, and (h) κ_1 based on the (i) GH, (ii) classic Heckman, (iii) Heckman-Skew, and (iv) Heckman-t sample selection models. Sample size $n = 2{,}000$.

Table 1. Empirical mean and RMSE of the maximum likelihood estimates of the parameters based on the GH, classic Heckman, Student-t, and skew-normal sample selection models under Scenario 1.

			Generalized Heckman		Classic Heckman		Heckman Skew-Normal		Heckman Student-t	
Parameters		n	mean	RMSE	mean	RMSE	mean	RMSE	mean	RMSE
γ_0	0.9	500	0.912	0.093	0.880	0.095	0.854	0.364	1.300	0.435
		1,000	0.905	0.063	0.878	0.069	0.820	0.364	1.293	0.409
		2,000	0.903	0.042	0.877	0.048	0.804	0.348	1.286	0.395
γ_1	0.5	500	0.507	0.081	0.555	0.104	0.527	0.095	0.652	0.200
		1,000	0.506	0.056	0.558	0.085	0.530	0.075	0.663	0.185
		2,000	0.504	0.040	0.555	0.070	0.529	0.057	0.662	0.174
γ_2	1.1	500	1.118	0.104	1.053	0.125	1.001	0.169	1.540	0.480
		1,000	1.109	0.074	1.046	0.099	0.978	0.162	1.522	0.441
		2,000	1.107	0.053	1.033	0.090	0.980	0.146	1.515	0.425
γ_3	0.6	500	0.606	0.084	0.556	0.106	0.531	0.127	0.848	0.288
		1,000	0.604	0.065	0.539	0.098	0.504	0.130	0.828	0.253
		2,000	0.602	0.040	0.543	0.074	0.513	0.106	0.833	0.242
β_0	1.1	500	1.105	0.068	0.998	0.261	0.496	0.734	1.112	0.103
		1,000	1.102	0.044	0.954	0.273	0.553	0.701	1.108	0.078
		2,000	1.099	0.029	0.892	0.230	0.498	0.697	1.096	0.051
β_1	0.7	500	0.702	0.036	0.882	0.219	0.739	0.156	0.787	0.106
		1,000	0.701	0.020	0.860	0.191	0.753	0.158	0.777	0.087
		2,000	0.700	0.014	0.881	0.191	0.770	0.154	0.774	0.079
β_2	0.1	500	0.098	0.048	0.140	0.196	0.040	0.225	0.101	0.078
		1,000	0.099	0.029	0.174	0.196	0.069	0.237	0.104	0.057
		2,000	0.100	0.019	0.221	0.147	0.105	0.193	0.113	0.042
λ_0	-0.4	500	-0.400	0.042	0.174	0.577	0.366	0.771	-0.476	0.113
		1,000	-0.402	0.028	0.182	0.584	0.339	0.743	-0.470	0.092
		2,000	-0.401	0.019	0.182	0.582	0.331	0.734	-0.463	0.075
λ_1	0.7	500	0.699	0.042	–	–	–	–	–	–
		1,000	0.700	0.031	–	–	–	–	–	–
		2,000	0.701	0.020	–	–	–	–	–	–
κ_0	0.3	500	0.314	0.248	0.507	0.688	0.133	0.846	0.307	0.300
		1,000	0.311	0.154	0.588	0.714	0.223	0.918	0.302	0.235
		2,000	0.309	0.106	0.811	0.608	0.307	0.814	0.336	0.158
κ_1	0.5	500	0.573	0.240	–	–	–	–	–	–
		1,000	0.531	0.135	–	–	–	–	–	–
		2,000	0.510	0.092	–	–	–	–	–	–

`ssmrob` (Zhelonkin, Genton and Ronchetti (2014)). These data were also used by Cameron and Trivedi (2009), Marchenko and Genton (2012), and Zhelonkin, Genton and Ronchetti (2016) to fit the classic Heckman model, Heckman-t model,

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

	Generalized Heckman			Classic Heckman			Heckman Skew-normal			Heckman Student-t		
n	LR	G	W	LR	G	W	LR	G	W	LR	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
1,000	0.9	0.7	2.6	72.4	7.7	92.7	82.7	26.2	95.6	3.4	2.5	4.0
2,000	1.1	0.7	1.9	85.9	16.5	96.8	91.0	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
1,000	4.9	3.7	6.9	86.9	27.1	95.5	89.9	49.7	97.8	10.1	8.6	11.8
2,000	5.9	5.1	7.4	94.2	40.8	97.7	94.5	68.3	99.5	10.2	9.4	10.6
						$\alpha = 10\%$						
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
1,000	9.4	8.2	13.7	91.5	41.4	96.1	93.3	61.1	98.2	16.9	15.4	18.0
2,000	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 3. Empirical power of the likelihood ratio (LR), gradient (G), and Wald (W) tests (in percentage) for simulated data according to Scenario 1 under GH, classic Heckman, Heckman-skew normal, and Heckman-t models, with significance levels at $1\%, 5\%$, and 10%.

	Generalized Heckman			Classic Heckman			Heckman Skew-normal			Heckman Student-t		
n	LR	G	W	LR	G	W	LR	G	W	LR	G	W
						$\alpha = 1\%$						
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
1,000	95.3	95.2	96.9	93.7	30.8	99.2	88.0	14.0	98.4	20.9	16.0	26.6
2,000	99.9	99.9	99.9	99.6	74.4	100	98.2	28.4	99.4	48.3	46.0	52.1
						$\alpha = 5\%$						
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
1,000	99	98.9	99.4	98.1	57.4	99.9	94.5	33.5	99	41.9	37.9	46.2
2,000	100	100	100	99.9	87.0	100	98.9	46.8	99.7	68.8	66.9	70.9
						$\alpha = 10\%$						
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
1,000	99.8	99.8	99.8	99.1	69.6	100	96	45.5	99	52.4	49.4	54.7
2,000	100	100	100	100	91.6	100	99.2	55.9	99.7	78.4	77.8	79.4

and the robust version of the two-step method, respectively. The MEPS is a set of large-scale 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

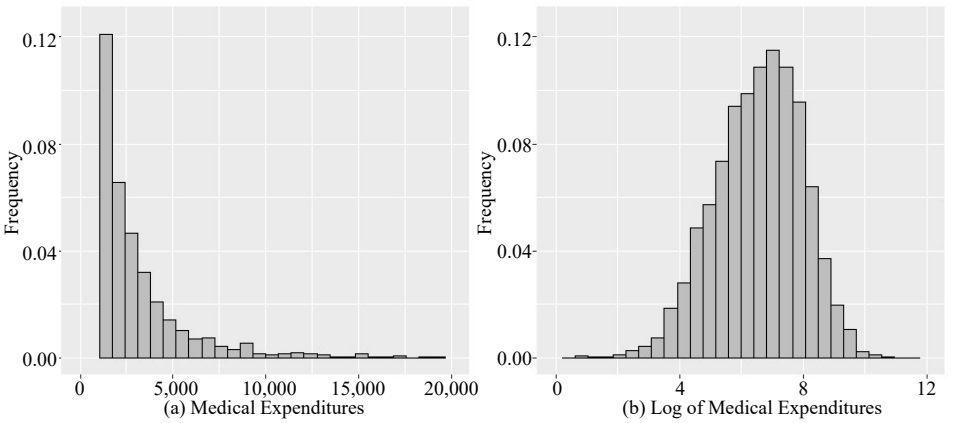

Figure 5. Histograms of the medical expenditure data (to the left) and of its logarithm (to the right).

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 the value one); `Educ` represents years of schooling; `Blhisp` is an indicator for ethnicity (black or Hispanic receive a value of one); `Totcr` is the total number of chronic diseases; `Ins` is the insurance status; and `Income` denotes an individual's income.

The variable of interest Y_{1i}^* represents the log-expenditure of the medical services of the ith individual. We consider the logarithm of the expenditure because it is highly skewed (i.e., see Figure 5). The variable Y_{2i}^* denotes the willingness of the ith individual to spend, and is not observed. We only observe $U_i = I\{Y_{2i}^* > 0\}$, which represents ith individual's decision on whether to spend on medical care.

According to Cameron and Trivedi (2009) and Zhelonkin, Genton and Ronchetti (2016), it is natural to fit a sample selection model to such data, because 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 to test $H_0 : \rho = 0$ 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 and Marchenko and Genton (2012) argued that a more robust model would find evidence of the

presence of sample selection bias in the data. These authors proposed using a Student-t sample selection model to deal with this problem. However, as illustrated in this application, this problem of the classic Heckman model may be due to the assumption of a constant sample selection bias and constant dispersion parameters, rather than the normal assumption itself. For those readers interested in an alternative approach to dealing with medical expenditures, we refer to Nawata and Kimura (2017), who consider a standard (type-I) tobit model with the power transformation.

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

$$\begin{aligned}
\mu_{1i} &= \beta_0 + \beta_1 \texttt{Age}_i + \beta_2 \texttt{Fem}_i + \beta_3 \texttt{Educ}_i + \beta_4 \texttt{Blhisp}_i + \beta_5 \texttt{Totchr}_i + \beta_6 \texttt{Ins}_i, \\
\mu_{2i} &= \gamma_0 + \gamma_1 \texttt{Age}_i + \gamma_2 \texttt{Fem}_i + \gamma_3 \texttt{Educ}_i + \gamma_4 \texttt{Blhisp}_i + \gamma_5 \texttt{Totchr}_i + \gamma_6 \texttt{Ins}_i + \\
&\quad + \gamma_7 \texttt{Income}_i, \\
\log \sigma_i &= \lambda_0 + \lambda_1 \texttt{Age}_i + \lambda_2 \texttt{Totchr}_i + \lambda_3 \texttt{Ins}_i, \\
\operatorname{arctanh} \rho_i &= \kappa_0 + \kappa_1 \texttt{Fem}_i + \kappa_2 \texttt{Totchr}_i,
\end{aligned}$$

for $i = 1, \ldots, 3328$. The primary equation has the same covariates as the selection equation, with the additional covariate `Income`, so that the 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 that were individually nonsignificant for ρ_i and σ_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 LR statistics equal to 4.207 with an associated p-value of 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 fits of the classic Heckman and GH models. From this table, we observe that the covariates `Fem` and `Totchr` are significant in terms of explaining the sample selection bias, using any significance level. We perform an LR test to check the absence ($H_0 : \kappa_0 = \kappa_1 = \kappa_2 = 0$) or presence of the sample selection bias. The LR statistic was 28.16, with corresponding p-value equal to 3×10^{-6}. Therefore, our proposed GH model is able to detect the presence of a sample selection bias, even under the normal assumption. We also performed the gradient and Wald tests, which confirmed this conclusion.

Furthermore, the covariates `age`, `totcr`, and `ins` are significant for the dis-

Table 4. Summary fits of the classic Heckman model (HM) and GH 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
Dispersion Parameter								
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
Sample Selection Bias Parameter								
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

persion parameter. For the selection equation, the covariate `Income` is not statistically significant (significance level at 5%) based on the GH model, in contrast to the classic Heckman model. However, it is important to keep it in order to satisfy the exclusion restriction. For the primary equation, 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 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-

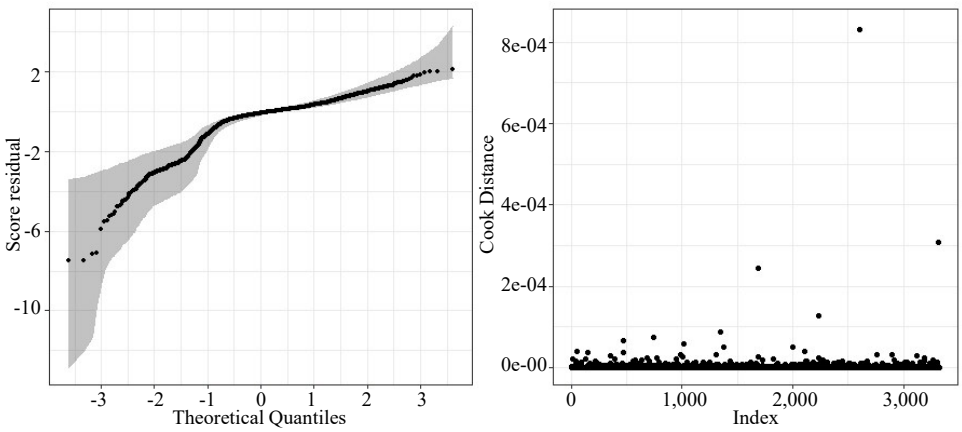

Figure 6. QQ-plot and its simulated envelope for the score residuals (left) and index plot of the GCD (right) for the GH model for the MEPS data.

plot of the score residuals given in (3.1) with simulated envelopes, and also a Cook distance plot for detecting a global influence. Based on this last plot, we do not detect any outlier observations, because all points are below the reference line $2p/n = 0.013$. We also investigate whether the highlighted point #2602 (above the line 8×10^{-4}) is influential. We fitted our model by removing this observation, and found no changes on the parameter estimates or different conclusions about the significance of the covariates. For the QQ-plot of the score residuals, the proposed model performs very well, because 96% of the points are inside the envelope. This confirms that the normal assumption for this particular data set is adequate, and that our GH model is suitable for the MEPS data analysis.

6. Conclusion

We have proposed a generalization of the Heckman model by allowing both the sample selection bias and the dispersion parameters to vary across covariates. We showed that the proposed model satisfies certain regularity conditions that ensure consistency and asymptotic normality of the MLEs. 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 GH model showed that the normal assumption for the data is suitable, in contrast to existing findings in the literature. Future research should address the following: (i) a generalization of other sample selection models, such as Student-t and skew-normal, to allow for varying sample selection bias and dispersion parameters; (ii) proper residuals for other sample selection models; and (iii) a deeper study of

influence analysis. An `R` package for fitting the GH, Student-t, and skew-normal models has been developed, and is available at `GitHub https://fsbmat-ufv.github.io/ssmodels/`.

A. Appendix

Proof of Theorem 1. We here show the results for the derivatives with respect to $\boldsymbol{\beta}$. 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

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

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

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

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

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 $\mu_{1i} + \rho_i\sigma_i(\phi(\mu_{2i})/\Phi(\mu_{2i}))$ (see Supplementary Material for more details) and get

$$\begin{aligned}
E\left(Z_i\big|U_i = 1\right) &= -\frac{\mu_{1i}}{\sigma_i} + \frac{1}{\sigma_i} E(Y_i|U_i = 1) \\
&= -\frac{\mu_{1i}}{\sigma_i} + \frac{1}{\sigma_i}\left(\mu_{1i} + \rho_i\sigma_i \frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})}\right) = \rho_i \frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})}.
\end{aligned}$$

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

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 $E((\partial^2 \ell_i/\partial\beta_j\beta_l)\big|U_i = 0) = -E\left((\partial \ell_i/\partial\beta_j)(\partial \ell_i/\partial\beta_l)\big|U_i = 0\right) = 0$.

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

$$E\left(\zeta_i \frac{\phi(\zeta_i)}{\Phi(\zeta_i)}\big|U_i = 1\right) = -\mu_{2i}(1-\rho_i^2)\frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})} \quad \text{and}$$
$$E\left(Z_i^2|U_i = 1\right) = 1 - \mu_{2i}\rho_i^2 \frac{\phi(\mu_{2i})}{\Phi(\mu_{2i})}.$$

By combining these results, we have that

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

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

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 $\boldsymbol{\beta}$'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 $\boldsymbol{\beta}$'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 $j,l,k=1,\ldots,p$, where $\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} g_i(\boldsymbol{\theta})$, 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 $E_{\boldsymbol{\theta}_0}\left(\zeta_i^{*2}(\phi^2(\zeta_i^*)/\Phi^2(\zeta_i^*))\right) < \infty$, where $E_{\boldsymbol{\theta}_0}(\cdot)$ denotes the expectation with respect to the true parameter vector value $\boldsymbol{\theta}_0$ and ζ_i^* is defined as ζ_i by replacing $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^*$. The proofs for the remaining terms follow from this one or in a similar way.

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

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

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

$$E_{\boldsymbol{\theta}_0}\left(\zeta_i^{*2}\frac{\phi^2(\zeta_i^*)}{\Phi^2(\zeta_i^*)}\right)$$
$$= E_{\boldsymbol{\theta}_0}\left(\zeta_i^{*2}\frac{\phi^2(\zeta_i^*)}{\Phi^2(\zeta_i^*)}I\left\{\zeta_i^* \leq \sqrt{2}\right\}\right) + E_{\boldsymbol{\theta}_0}\left(\zeta_i^{*2}\frac{\phi^2(\zeta_i^*)}{\Phi^2(\zeta_i^*)}I\left\{\zeta_i^* > \sqrt{2}\right\}\right)$$
$$\leq \sqrt{2}\left\{2\pi\Phi^2(\sqrt{2})\right\}^{-1} + E_{\boldsymbol{\theta}_0}\left(|\zeta_i^*|^3\right) < \infty,$$

with $E_{\boldsymbol{\theta}_0}(|\zeta_i^*|^3) < \infty$ being proved in the same lines that the first two moments presented in the Supplementary Material, which completes the proof of the desired result.

Supplementary Material

In the online Supplementary Material, we provide the score function, information matrix, and moments of the response variable, as well as additional simulated results.

Acknowledgments

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

References

Arabmazar, A. and Schmidt, P. (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 University Press, Oxford.

Azzalini, A., Kim, H. M. and Kim, H. J. (2019). Sample selection models for discrete and other non-Gaussian response variables. *Statistical Methods and Applications* **28**, 27–56.

Cameron, C. A. and Trivedi, P. K. (2009). *Microeconometrics Using Stata*. Revised Edition. Stata press, College Station.

Chib, S., Greenberg, E. and Jeliazkov, I. (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics* **18**, 321–348.

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

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Cox, D. R. and Hinkley, D. V. (1979). *Theoretical Statistics*. Chapman and Hall/CRC, Boca Raton.

Donald, S. G. (1995). Two-step estimation of heteroskedastic sample selection models. *Journal of Econometrics* **65**, 347–380.

Durrett, R. (2019). *Probability: Theory and Examples*. 5th Edition. Cambridge University Press, Cambridge.

Enders, C. K. (2010). *Applied Missing Data Analysis*. 1st Edition. Guilford Press, New York.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The 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 Econometrics* **11**, 247–258.

Kim, H., Roh, T. and Choi, T. (2019). Bayesian analysis of semiparametric Bernstein polynomial regression models for data with sample selection. *Statistics* **53**, 1082–1111.

Lai, H. P. and Tsay, W. J. (2018). Maximum simulated likelihood estimation of the panel sample selection model. *Econometric Reviews* **37**, 744–759.

Leung, S. F. and Yu, S. (1996). On the choice between sample selection and two-part models.

Journal of Econometrics **72**, 197–229.

Leung, S. F. and Yu, S. (2000). Collinearity and two-step estimation of sample selection models: Problems, origins, and remedies. *Computational Economics* **15**, 173–199.

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

Marra, G. and Wyszynski, K. (2016). Semi-parametric copula sample selection models for count responses. *Computational Statistics and Data Analysis* **104**, 110–129.

Mu, B. and Zhang, Z. (2018). Identification and estimation of heteroscedastic binary choice models with endogenous dummy regressors. *The Econometrics Journal* **21**, 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 Kimura, M. (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. *The Econometrics Journal* **12**, 217–229.

Ogundimu, E. O. and Hutton, J. L. (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*. R Foundation for Statistical Computing, Vienna.

Wiesenfarth, M. and Kneib, T. (2010). Bayesian geoadditive sample selection models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **59**, 381–404.

Wojtys, M., Marra, G. and Radice, R. (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 Marra, G. (2018). Sample selection models for count data in R. *Computational Statistics* **33**, 1385–1412.

Xie, F.-C. and Wei, B.-C. (2007). Diagnostics analysis for log-Birnbaum–Saunders regression models. *Computational Statistics and Data Analysis* **51**, 4692–4706.

Yamagata, T. and Orme, C. (2005). On testing sample selection bias under the multicollinearity problem. *Econometric Reviews* **24**, 467–481.

Zhelonkin, M., Genton, M. G. and Ronchetti, E. (2014). *R package ssmrob: Robust estimation and inference in sample selection models*. CRANR package version 0.4.

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

Fernando de Souza Bastos
Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Viçosa, Florestal, Brazil.
E-mail: fernando.bastos@ufv.br
Wagner Barreto-Souza
Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.
E-mail: wagner.barretosouza@kaust.edu.sa
Marc G. Genton
Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.
E-mail: marc.genton@kaust.edu.sa

(Received February 2021; accepted March 2021)