Journal of Statistical Software

August 2021, Volume 99, Issue 4. *doi: 10.18637/jss.v099.i04*

Robust Analysis of Sample Selection Models through the R Package ssmrob

Mikhail Zhelonkin
Erasmus University Rotterdam

Elvezio Ronchetti
University of Geneva

Abstract

The aim of this paper is to describe the implementation and to provide a tutorial for the R package **ssmrob**, which is developed for robust estimation and inference in sample selection and endogenous treatment models. The sample selectivity issue occurs in practice in various fields, when a non-random sample of a population is observed, i.e., when observations are present according to some selection rule. It is well known that the classical estimators introduced by Heckman (1979) are very sensitive to small deviations from the distributional assumptions (typically the normality assumption on the error terms). Zhelonkin, Genton, and Ronchetti (2016) investigated the robustness properties of these estimators and proposed robust alternatives to the estimator and the corresponding test. We briefly discuss the robust approach and demonstrate its performance in practice by providing several empirical examples. The package can be used both to produce a complete robust statistical analysis of these models which complements the classical one and as a set of useful tools for exploratory data analysis. Specifically, robust estimators and standard errors of the coefficients of both the selection and the regression equations are provided together with a robust test of selectivity. The package therefore provides additional useful information to practitioners in different fields of applications by enhancing their statistical analysis of these models.

Keywords: endogenous treatment model, R, robust estimation, robust inference, sample selection models, two-step estimator.

1. Introduction

The present paper has three purposes. First, we introduce the R package **ssmrob** (Zhelonkin, Genton, and Ronchetti 2021) for the robust analysis of data with sample selection; second, we discuss some practical aspects about the use of robust methods in general and in sample selection models in particular; third, we propose a robust estimator for the endogenous treat-

ment model. As advocated by several authors (see e.g., Athey and Imbens 2017), in order to increase transparency and credibility of research, it is reasonable to complement the reported results by a supplementary analysis. Several measures have been proposed, see Andrews, Gentzkow, and Shapiro (2017) and Athey and Imbens (2015). We believe, that reporting the results of a robust analysis should become a normal practice, if one uses parametric estimators. This will not only safeguard the statistical analysis from misleading implications possibly due to deviations from the assumptions on the model, but also enrich the analysis by offering additional useful information on the structure of the data.

The paper can be read in different ways. Those readers already familiar with robust statistics and interested in applying directly the new robust analysis can check the models covered here in Section 2 and then go directly to the implementation and the use of the package in Section 5 and 6, respectively. Those who would like to have a short introduction to the basic concepts of robust statistics and a general discussion on its role in this setup, including its relationship to parametric and semi-parametric methods, can read Section 2 and 3. Section 4 contains the description of the robust estimators and tests, which is useful to understand their structure and the options in the functions presented in Section 5 and 6. A user willing to use only the default options of the package can skip this section. Finally, the practical use of the package is discussed in Section 5 and Section 6, where two empirical applications are discussed in details by comparing the classical statistical analysis with its full robust counterpart.

2. Sample selection models and main functions

We consider three models, but for simplicity of exposition we will focus on the standard Heckman (1979) framework.

Heckman's model (Tobit-2 model)

$$y_{1i} = I(x_{1i}^\top \beta_1 + e_{1i} > 0), \tag{1}$$

$$y_{2i} = \begin{cases} x_{2i}^\top \beta_2 + e_{2i}, & \text{if } y_{1i} = 1, \\ \text{NA}, & \text{if } y_{1i} = 0, \end{cases} \tag{2}$$

where x_{ji} is a vector of explanatory variables, β_j is a $p_j \times 1$ vector of parameters, $j = 1, 2$, e_{ji} are the error terms which follow a bivariate normal distribution with variances $\sigma_1^2 = 1$, σ_2^2, and correlation ρ, and I is the indicator function. The variance parameter σ_1^2 is set to be equal to 1 to ensure identifiability. Here (1) is the selection equation, defining the observability rule, and (2) is the equation of interest or outcome equation. Notice that sometimes instead of NA (not available) zeros are used, although this notational practice can be misleading. The system (1)–(2) is also known as Tobit-2 model according to Amemiya (1984) classification. In the analysis of the dataset in Section 6.2 we use this model, where y_{1i} in the selection equation defines whether or not the i'th individual enters into the labor force and y_{2i} in the outcome equation represents its log-wage. x_{1i} and x_{2i} are vectors of covariates which include age, education status, experience, and squared experience.

Switching regressions model (Tobit-5 model)

It is a natural extension of Heckman's model. In this case instead of NA in (2) we have a

second regime. The selection equation (1) remains the same. The outcome equation can be written as follows:

$$y_{2i} = \begin{cases} x_{21i}^\top \beta_{21} + e_{21i}, & \text{if } y_{1i} = 1, \\ x_{22i}^\top \beta_{22} + e_{22i}, & \text{if } y_{1i} = 0, \end{cases}$$

where x_{2ji} are the vectors of explanatory variables, β_{2j} are q_j vectors of parameters, $j = 1, 2$, the error terms e_{21i} and e_{22i} together with e_{1i} follow a multivariate normal distribution

$$\begin{pmatrix} e_{1i} \\ e_{21i} \\ e_{22i} \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix} \right\}. \tag{3}$$

Endogenous treatment model (ETM)

It has the same selection equation (1), but the outcome equation becomes

$$y_{2i} = x_{2i}^\top \beta_2 + \alpha y_{1i} + e_{2i}, \tag{4}$$

where the dependent variable y_1 appears as an explanatory variable. The error terms follow a bivariate normal distribution with the same covariance structure as in the Tobit-2 model. Because of non-zero correlation between the errors, y_1 becomes endogenous. The parameter α is the average treatment effect. In this case y_1 is a treatment variable, for instance the decision to enroll in a job training program.

These three models are the central tools for the analysis of data with non-random sampling and play an important role in policy evaluation and treatment effect estimation in observational studies.

Package **ssmrob** (Zhelonkin *et al.* 2021) is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=ssmrob`. The main functions in the package are `ssmrob()` and `etregrob()`. The function `etregrob()` is the estimator of ETM, which returns an object of class '`etregrob`'. Function `ssmrob()` works as a wrapper. In the current version (version 1.0) there are two options: the Heckman's selection model (Tobit-2) and the switching regressions model with probit selection mechanism (Tobit-5). If the Tobit-2 model is chosen, then `heckitrob()` is called; if the Tobit-5 model is chosen then `heckit5rob()` is called. The function `ssmrob()` returns the object of class '`heckitrob`' or '`heckit5rob`' for Tobit-2 or Tobit-5, respectively.

3. Estimation

In this section we briefly review different estimation approaches (for a thorough review, see e.g., Vella 1998) and discuss situations where each method is preferable from the robust statistics perspective. We focus on the Tobit-2 model, but the general arguments hold for two other models as well.

3.1. Parametric estimation

Without any doubt, the parametric approach is currently the most popular approach for the estimation of sample selection models in practice. It is straightforward to write the likelihood

function which now can be relatively easily maximized, although it was quite computationally difficult in the seventies, when the model was introduced. Heckman (1979) proposed an appealing two-step estimator, which became standard because of its simplicity and easy interpretation. Consider the conditional expectation of y_{2i} given x_{2i} and the selection rule

$$E(y_{2i} \mid x_{2i}, y_{1i} = 1) = x_{2i}^\top \beta_2 + E(e_{2i} \mid e_{1i} > -x_{1i}^\top \beta_1). \tag{5}$$

The expectation on the right hand side of (5) is in general not equal to zero, which leads to the following regression

$$y_{2i} = x_{2i}^\top \beta_2 + \lambda(x_{i1}^\top \beta_1)\beta_\lambda + \nu_i, \tag{6}$$

where $\beta_\lambda = \rho\sigma_2$, $\lambda(x_{i1}^\top \beta_1) = \phi(x_{i1}^\top \beta_1)/\Phi(x_{i1}^\top \beta_1)$ is the conditional expectation on the right hand side of (5) called the inverse Mills ratio (IMR), ν_i is the zero expectation error term and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and cumulative distribution function of the standard normal distribution, respectively. Heckman suggested to estimate β_1 by probit maximum likelihood estimator (MLE) and to use ordinary least squares (OLS) in regression (6), where $\lambda(\cdot)$ is correcting for the selection bias. A similar two-step estimation structure can be used for switching regressions and ETM. The first step is probit MLE, the second step is OLS with corresponding conditional expectation as a selection bias correction.

Nowadays both Heckman's two-step estimator and the full information maximum likelihood estimator (FIML) are standard methods implemented in many software packages including SAS (SAS Institute Inc. 2014), using `proc qlim`, Stata (StataCorp 2017), using `heckman` and `etregress`, and R (R Core Team 2021) using package **sampleSelection** by Toomet and Henningsen (2008).

However, in spite of their simplicity and appealing interpretation, both estimators are very sensitive to the normality assumption on the error terms. The two-step estimator is considered to be slightly more robust than FIML (Cameron and Trivedi 2009, p. 544). In spite of the fact that the distributional assumptions for both estimators are the same, in the situation when there are measurement errors in the outcome equation, the two-step estimator remains consistent, while FIML is not (Stapleton and Young 1984). Another interesting case is when there are outliers only in the outcome equation. Then, in the two-step estimator at least the selection equation will be estimated correctly and the second step can become biased, while using FIML, parameters in both equations can become biased. In general, as we show below, both estimators can be arbitrarily biased, even when the assumed model F is approximately correct, say e.g., 99% of observations come from F and 1% from some arbitrary distribution G.

A natural way to mitigate the sensitivity problem is to use a more flexible parametric family of distributions. For instance, Lee (1983) and Marchenko and Genton (2012) proposed to use the t distribution, which allows longer tails and their adjustment using the degrees of freedom, Smith (2003) proposed to use a copula-based approach, and Ogundimu and Hutton (2016) proposed skew-normal selection model. Although, these approaches add more flexibility to the standard normal model, they do not provide full protection against possible deviations from the central model, i.e., when the true data generating distribution lies in a full neighborhood of the assumed parametric model; see Section 3.2.

3.2. Robust estimation

The robust approach offers a reasonable compromise between the fully parametric approach explained in Section 3.1 and the semiparametric approach discussed in Section 3.3. We still

assume the classical normal model F as the central model, but we believe that the true data generating distribution lies in a neighborhood of it, i.e., $F_\epsilon = (1-\epsilon)F + \epsilon G$, where G is some unknown arbitrary distribution and ϵ is (typically) small. We then derive (robust) estimators that are consistent for the central model F and remain stable when the true data generating distribution is F_ϵ lying in a neighborhood of F, i.e., their bias will always be bounded no matter the distribution F_ϵ. Notice that the classical estimators (FIML and Heckman's two-step) are also consistent for the central model, but can have an infinite bias when the underlying distribution of the data lies in a small neighborhood of the central model. The robust estimators identify the same parameters as the classical parametric estimators and therefore retain the interpretation as in the classical case. The price to pay for robustness is some loss of efficiency at the central model. In the default tuning that is implemented in the package, this is approximately 20%. One might argue that the two-step estimator is inefficient itself, however even with minor contamination (1%), the robust estimator becomes more efficient than the classical one. We demonstrate this issue in the examples in Section 6. Details about the robust approach are provided in Section 4 and a complete discussion can be found in Zhelonkin *et al.* (2016).

3.3. Semi- and nonparametric estimation

If the parametric assumptions are not satisfied even approximately or if we are completely uncertain about the data generating distribution, the natural alternative is to use semi- and nonparametric estimators. The goal of robust methods is to estimate the parameters of the central model F in order to retain their interpretation in spite of the fact that the data were generated by some F_ϵ. In a fully nonparametric setup the goal would be to estimate characteristics of the distribution F_ϵ, such as its expectation instead of the expectation of the central model F.

The literature on semi- and nonparametric estimation on models with sample selectivity is large, see Ahn and Powell (1993), Newey (2009) and Chapter 8 in the book by Pagan and Ullah (1999). Here we only briefly discuss the estimators which preserve the linearity in the parameters and relax the distributional assumptions. For the treatment of nonlinear predictors we refer to the paper by Wojtyś, Marra, and Radice (2016) and their package **SemiParSampleSel**, see also Das, Newey, and Vella (2003) and references therein.

Similarly to parametric estimation there are one-step FIML-like estimators and semiparametric two-step estimators. An example of one-step estimator is a method proposed by Gallant and Nychka (1987). The method is based on Hermite series approximation of the true density. The authors provide the consistency result, but not the (asymptotic) distribution theory. Another strategy is to use a two-step approach. Newey (2009) proposed first to estimate the selection equation as a semiparametric single index model (Klein and Spady 1993) and then to use a series expansion to correct for sample selection in the second step with OLS. However, there is a possible identification problem. The single index model does not identify β_1: the intercept is not identified and other components of β_1 are identified only up to a proportionality factor. For the identification of parameters in the outcome equation we need to assume exclusion restrictions, i.e., there must be a variable in x_1 that is not included in x_2. If $x_1 = x_2$ then the identification of the model depends entirely on the functional form and the distributional assumptions (Manski 1989). In practice it is often difficult to find such a variable, and as mentioned by (Vella 1998, p. 135) some economic models require the same

explanatory variables to appear in both equations. Moreover, the intercept in β_2 is incorporated in the series expansion and needs a separate estimation. Heckman (1990) and Andrews and Schafgans (1998) proposed methods to estimate the intercept, which require the so-called identification at infinity.

If the practitioner switches to semi- nonparametric methods, then the exclusion restriction must be imposed, and not all the structural parameters can be identified. This creates a delicate trade-off between the parametric assumptions and the nonparametric identification assumptions. What is more restrictive in practice is a complicated question, which is beyond the scope of this paper. It must also be mentioned that the exclusion restriction is also desirable (but formally not compulsory) for the fully parametric setup since the inverse Mills ratio is quasi-linear on a wide range of its support. The FIML estimator also suffers from this problem; see Leung and Yu (2000) for a thorough discussion.

Most empirical work is based on the parametric approach and there are several reasons for that. The simplest one is that nonparametric methods are technically demanding. Heckman and Vytlacil (2007) mentioned the issues of the sensitivity of semiparametric estimators to the choices of smoothing parameters, trimming parameters and bandwidths. Another reason is that the parametric framework allows to estimate the structural parameters corresponding to the economic model and to evaluate the entire model and not only to estimate some parameters.

3.4. Quantile regression approach

Quantile regression (QR) (Koenker 2005) is often considered a robust alternative to OLS in linear regression. In the original paper by Koenker and Bassett (1978) the robustness issue was one of the central points for the introduction of QR together with the advantage (compared to standard OLS) of providing the estimation of all the quantiles (and not just the expectation) of the conditional distribution of the response variable given the covariates.

In the sample selection literature several QR methods have been proposed (Buchinsky 1998; Huber and Melly 2015; Arellano and Bonhomme 2017). The main focus of these papers is the the estimation of quantile effects, but the robustness properties of the estimators are not discussed. Although the QR approach is a powerful tool for statistical modeling, its robustness is not automatically guaranteed, at least by the currently available methods. Indeed there are situations, where the QR estimator can break down in the presence of even a very small deviation from the assumed model. We will get back to this point with more details at the end of Section 4.1.

4. A general robust approach

In this section we describe a general way to obtain robust estimators and tests for models with sample selectivity.

4.1. Robust estimation

The robustness issues with Heckman's two-stage estimator can be naturally described by studying the behavior of the estimator when the true error generating distribution is not the central normal model F but some perturbation $F_\epsilon = (1-\epsilon)F + \epsilon G$, where G is an

unknown arbitrary distribution, and ϵ is the contamination proportion. Our estimator in this framework will be consistent for the model F, but slightly biased for F_ϵ, any distribution in a ϵ-neighborhood of F. Notice that the classical estimator can have an infinite bias at a F_ϵ, see below.

The robust approach controls the worst possible bias that can occur when the data generating distribution belongs to that neighborhood. It turns out that the worst possible bias over all G can be linearly approximated by

$$\epsilon \cdot \sup_z ||IF(z; T, F)||,$$

where $IF(z; T, F)$ is the so-called influence function (IF), a derivative of the estimator viewed as functional of the underlying distribution; see Hampel (1974) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986). If the estimator has an unbounded influence function, the corresponding worst bias in the ϵ-neighborhood of F will be infinite.

To check the IF of Heckman's estimator, remember that β_1 is estimated in the first stage by MLE in a probit model, whereas β_2 is estimated by OLS in regression (6), with $\lambda(\cdot)$ correcting for the selection bias.

The population versions of the estimating equations defining the two-stage estimator are given by:

$$\int \Psi_1\{(x_1, y_1); S(F)\} dF = 0, \tag{7}$$

$$\int \Psi_2[(x_2, y_2); \lambda\{(x_1, y_1); S(F)\}, T(F)] dF = 0, \tag{8}$$

where

$$\Psi_1\{(x_1, y_1); S(F)\} = \{y_1 - \Phi(x_1^\top \beta_1)\} \frac{\phi(x_1^\top \beta_1)}{\Phi(x_1^\top \beta_1)\{1 - \Phi(x_1^\top \beta_1)\}} x_1, \tag{9}$$

$$\Psi_2[(x_2, y_2); \lambda\{(x_1, y_1); S(F)\}, T(F)] = (y_2 - x_2^\top \beta_2 - \lambda\beta_\lambda) \begin{pmatrix} x_2 \\ \lambda \end{pmatrix} y_1, \tag{10}$$

are the score functions of the first and second stage estimators, respectively, $S(F)$ and $T(F)$ are the population counterparts of the estimators of the parameters β_1 and β_2 respectively, and $\lambda\{(x_1, y_1); S(F)\}$ denotes the dependence of λ on $S(F)$, while $T(F)$ depends directly on F and indirectly on F through $S(F)$.

The two-stage estimator is the solution of the empirical counterpart of the above system of estimating equations. In particular, (7) with (9) corresponds to the estimating equation of the MLE in a probit model, whereas (8) with (10) is the estimation equation for OLS in the regression model (6).

It can be shown that the IF of T is proportional to the score functions Ψ_1 and Ψ_2, i.e., it is linear in $\{y_1 - \Phi(x_1^\top \beta_1)\}$, $(y_2 - x_2^\top \beta_2 - \lambda\beta_\lambda)$, x_1, x_2, λ. Therefore it is unbounded and this causes the non-robustness problems of the estimator; see Proposition 1 in Zhelonkin *et al.* (2016).

It is then clear that in order to robustify the classical Heckman's two-stage estimator, we need to bound the two score functions, which amounts to perform a robust probit estimation in the first stage and a robust regression in the second stage. To do that, a thresholding

of the linear functions mentioned above is necessary and this is achieved by "huberizing" (i.e., applying the Huber function (12) to) the errors $\{y_1 - \Phi(x_1^\top \beta_1)\}$ and $(y_2 - x_2^\top \beta_2 - \lambda\beta_\lambda)$, and by downweighting x_1 and x_2. This operation defines new scores functions, so-called of Mallows type. For the first stage, following Cantoni and Ronchetti (2001), we define:

$$\Psi_1^R\{z_1; S(F)\} = \psi_{c_1}(r)\frac{1}{V^{1/2}(\mu)}\omega_1(x_1)\mu' - \alpha(\beta_1), \tag{11}$$

where $z_1 = (x_1, y_1)$, $r = \frac{y_1-\mu}{V^{1/2}(\mu)}$ are the Pearson residuals, $\psi_{c_1}(\cdot)$ is the Huber function defined by

$$\psi_{c_1}(r) = \begin{cases} r, & |r| \leq c_1, \\ c_1 \text{ sign}(r), & |r| > c_1, \end{cases} \tag{12}$$

$\alpha(\beta_1) = \frac{1}{n}\sum_{i=1}^n E\{\psi_{c_1}(r_i)\frac{1}{V^{1/2}(\mu_i)}\}\omega_1(x_{1i})\mu_i'$, $\mu_i = \Phi(x_{1i}^\top\beta_1)$, $\mu_i' = \frac{\partial}{\partial\beta_1}\mu_i$, $V(\mu_i) = \Phi(x_{1i}^\top\beta_1)\{1-\Phi(x_{1i}^\top\beta_1)\}$, $\omega_1(\cdot)$ is a weight based on the inverse of the robust Mahalanobis distance computed by means of high breakdown robust estimators of location and scatter of the x_{1i}. Several options are implemented in the package; see the argument `heckitrob.control()` in Section 5. The constant $\alpha(\beta_1)$ ensures that the new modified estimating equation remains unbiased, i.e., $E[\Psi_1^R\{z_1; S(F)\}] = 0$.

The modification of the classical score function entails an efficiency loss at the normal model. This can be viewed as an insurance premium, which provides protection against possible deviations from the model and their consequences on the bias of the estimator. Then, the tuning constant c_1 can be chosen to ensure a given level of asymptotic efficiency at the normal model. A typical value is 1.345.

For the second stage, we have a Mallows type robust score function:

$$\Psi_2^R(z_2; \lambda, T) = \psi_{c_2}\{(y_2 - x_2^\top\beta_2 - \lambda\beta_\lambda)/\sigma\}\omega(x_2, \lambda)y_1, \tag{13}$$

where $c_2 = 1.345$, $z_2 = (x_2, y_2)$, the weight function $\omega(\cdot,\cdot)$ is based on the robust Mahalanobis distance $d(x_2, \lambda)$, e.g.,

$$\omega(x_2, \lambda) = \begin{cases} 1 & \text{if } d(x_2, \lambda) < c_m, \\ \frac{c_m}{d(x_2,\lambda)} & \text{if } d(x_2, \lambda) \geq c_m, \end{cases} \tag{14}$$

with c_m the 95% quantile of the χ^2-distribution. In the situation, when the exclusion restriction is not available an additional modification of weights is used in order to reduce the loss of efficiency. We split the covariate space in two subspaces, then calculate the weights separately, and finally combine them. Details are given in Zhelonkin *et al.* (2016), p. 814.

Notice that the original QR estimator (Koenker and Bassett 1978) has a bounded IF with respect to the dependent variable, but its IF is unbounded in the space of explanatory variables. In a sample selection setting there is also the first estimation stage. If one uses probit MLE, then the final QR won't be robust in any case. If the robust probit is used, then the introduction of robustness weights in the covariates space is required, and to the best of our knowledge, this is still an open question. Buchinsky (1998) and Huber and Melly (2015) used the semiparametric single index model in the first step and the original QR in the second step. This leads to identification issues discussed in Section 3.3, and this procedure is in general not robust to outliers in the covariates space. Moreover, it is less efficient than our

robust estimator (Zhelonkin *et al.* 2016). To summarize, the QR approach is a very useful tool for modeling quantile effects. However, one should be careful about using it as a robust alternative, since a fully robust QR estimator in sample selection model is not available yet.

4.2. Robust inference

The robust two-stage estimator defined by (7)–(8), where the score functions are given by (11)–(13) is an M-estimator and its asymptotic variance is readily available (see Zhelonkin *et al.* (2016), p. 814, Formula (19)), which can be consistently estimated by means of the heteroscedasticity-consistent variance estimator; see Eicker (1967), Huber (1967) and White (1980). This allows us to construct a t test to test sample selection bias, i.e., $H_0 : \beta_\lambda = 0$ vs $H_A : \beta_\lambda \neq 0$, based on the robust estimator of β_λ and the corresponding estimator of its standard error.

4.3. Robust ETM

The two-step estimator of ETM consists of a probit MLE for the selection equation (1) and OLS for the following regression

$$y_{2i} = x_{2i}^\top \beta_2 + \alpha y_{1i} + \beta_\lambda \lambda^C + \tilde{\nu}_i, \tag{15}$$

where $\tilde{\nu}_i$ is a zero mean error term, λ^C is the inverse Mills ratio for the complete sample, defined by

$$\lambda^C\{z_1; S(F)\} = y_1 \left\{ \frac{\phi(x_1^\top \beta_1)}{\Phi(x_1^\top \beta_1)} \right\} + (1 - y_1) \left\{ \frac{-\phi(x_1^\top \beta_1)}{1 - \Phi(x_1^\top \beta_1)} \right\}.$$

The second option is to represent (15) in the form of switching regressions and estimate the average treatment effect α as a difference between the intercepts of two states. Both estimators have unbounded IF's and are not robust. The IF of the former is derived in Appendix A. The IF of the switching regressions estimator is discussed in the supplementary material of Zhelonkin *et al.* (2016).

The structure of the IF is similar to that of the Tobit-2 estimator. Hence, we apply the same principles as in Section 4.1 for the construction of the robust estimator for ETM. The first step is a robust probit with the score function (11). The second step is a Mallows type M-estimator with score function as in (13)

$$\Psi_2^R(z_2; \lambda^C, T) = \psi_{c_2}\{(y_2 - x_2^\top \beta_2 - \alpha y_1 - \lambda^C \beta_\lambda)/\sigma\} \omega(x_2, \lambda^C),$$

where $\psi_{c_2}(\cdot)$ is the Huber function defined by (12), and $\omega(\cdot, \cdot)$ is the weight function defined by (14). The combination of these score functions bounds the IF, making the estimator (locally) robust. A simulation study illustrating the performance of the robust estimator as well as the classical two-step and FIML estimators is presented in Appendix B.

5. Implementation and description of the functions

The package is written completely in R. It imports the packages **sampleSelection** (Toomet and Henningsen 2008), **robustbase** (Todorov and Filzmoser 2009; Maechler *et al.* 2021), and **MASS** (Venables and Ripley 2002). The package **mvtnorm** (Genz *et al.* 2020) is used for

the examples based on simulated data, which requires the simulation of the errors from a multivariate normal distribution. All these packages are available from the Comprehensive R Archive Network at http://CRAN.R-project.org/.

We give a description of the functions implemented in the package and discuss some aspects of the implementation.

```
ssmrob(outcome, selection, data, control = heckitrob.control())
```

This function is a wrapper which, depending on the type of arguments in `selection` and `outcome`, chooses the model and calls the corresponding estimator. The argument `outcome` is a simple formula for the case of Tobit-2, or a list of two formulas for the case of switching regressions. The argument `selection` is a formula for the selection equation. The argument `control` defines the accuracy and the robustness tuning parameters.

```
heckitrob.control(acc = 1e-04, test.acc = "coef", maxit = 50,
                  weights.x1 = c("none", "hat", "robCov", "covMcd"),
                  weights.x2 = c("none", "hat", "robCov", "covMcd"),
                  tcc = 1.345, t.c = 1.345)
```

This function provides the tuning parameters for the robust two-stage estimator. The parameters `acc` and `test.acc` control for the accuracy of estimation. The maximum number of iterations is defined by `maxit`. The leverage weights for the first stage ($\omega_1(x_1)$ in (11)) and for the second stage ($\omega(x_2, \lambda)$ in (13)) are defined by `weights.x1` and `weights.x2`, respectively. If `"none"` is chosen, which is the default option, then the weights are equal to 1. If `"hat"` is chosen, then weights on the design of the form $\sqrt{1 - h_{ii}}$ are used, where h_{ii} are the diagonal elements of the hat matrix. If `"robCov"` is chosen, then weights based on the robust Mahalanobis distance of the design matrix are used, where the covariance matrix is estimated by the `rob.cov()` method from package **MASS** (Venables and Ripley 2002) using the minimum volume ellipsoid estimator (Rousseeuw 1985). Similarly, if `"covMcd"` is chosen, the covariance is estimated by the minimum covariance determinant estimator (Rousseeuw and Van Driessen 1999). The arguments `tcc` and `t.c` are the tuning constants c_1 and c_2 for the Huber-functions of the first (11) and second (13) stage estimators, respectively.

```
heckitrob(outcome, selection, data, control = heckitrob.control())
```

This function presents the robust two-stage estimator of the Tobit-2 model. The arguments `outcome` and `selection` must be formulas. Note that, if `tcc` and `t.c` (tuning constants c_1 and c_2) are large and the leverage weights are ones, then the estimator converges to the classical Heckman's two-stage estimator.

```
heckit5rob(outcome1, outcome2, selection, data,
           control = heckitrob.control())
```

Similarly to the previous function, but `heckit5rob()` estimates the switching regressions model. The set of the tuning parameters for the second step estimator is used for both states.

```
etregrob(outcome, selection, data, control = heckitrob.control())
```

This function presents the two-step estimator for ETM, described in Section 4.3. It uses the same control function as the Tobit-2 and Tobit-5 models.

The computation of the asymptotic variance matrices is performed by the functions `heck2steprobVcov()`, `heck5twosteprobVcov()` and `etreg2steprobVcov()` for the Tobit-2, Tobit-5 and ETM models, respectively. They are used inside of the corresponding estimator functions. The output can be obtained by using the `vcov()` method on the object of '`heckitrob`', '`heckit5rob`' or '`etregrob`' classes.

The package provides the generic functions. The function `print()` prints the estimation results. The function `summary()` calculates and prints the summary of the estimation with standard errors, t values and p values, and returns an object of class '`summary.heckitrob`' or '`summary.heckit5rob`', '`summary.etregrob`' for Tobit-2, Tobit-5, and ETM respectively. The function `coef()` extracts the estimated coefficients, the function `vcov()` returns a list of two or three variance-covariance matrices, one for the selection equation and one or two (depending on the model) for the outcome equation. The functions `fitted()` and `residuals()` return one vector or a list that contains two vectors of fitted values or residuals, also depending on the model (one vector for Tobit-2 and ETM, two vectors for Tobit-5). The function `model.matrix()` has an argument `part` with `"outcome"` as a default value, which produces a design matrix of the outcome equation for Tobit-2 and ETM and a list with two matrices for Tobit-5. If `part = "selection"` then a design matrix for the selection equation is returned. Finally, the function `nobs()` returns the number of observations.

6. Using the ssmrob package

In this section we provide illustrative examples. First we treat a simulated example for sample selection model (1)–(2). It illustrates the behavior of the classical and robust estimators when the data generating process (DGP) is known. Second, we examine two empirical applications. Both have already been analyzed in the literature and are well known. In the example about wage offers there is no robustness problem, and the results of the estimation by classical and robust procedures are close. In the second example (ambulatory expenditures), on the contrary, the distributional assumptions are violated and the robust estimator is different from the classical estimator. The behavior of robust and classical estimators for Tobit-5 and endogenous treatment models is similar and their discussion can be found in Appendix.

To start, we load the package **ssmrob**.

```
R> library("ssmrob")
```

6.1. Tobit-2 model

We simulate the data from model (1)–(2):

```
R> library("mvtnorm")
R> N <- 5000
R> beta1 <- c(0, 1.0, 1.0, 0.75)
R> beta2 <- c(0, 1.5, 1.0, 0.5)
R> set.seed(2)
R> x1 <- rmvnorm(N, mean = c(0, -1, 1), sigma = diag(c(1, 0.5, 1)))
```

```
R> x2 <- x1
R> x2[, 3] <- rnorm(N, 1, 1)
R> covmtrx <- matrix(c(1, 0.7, 0.7, 1), 2, 2)
R> eps <- rmvnorm(N, mean = rep(0, 2), sigma = covmtrx)
R> y1 <- ifelse(cbind(1, x1) %*% beta1 + eps[, 1] > 0, 1, 0)
R> y2 <- ifelse(y1 == 1, cbind(1, x2) %*% beta2 + eps[, 2], NA)
```

We set the sample size equal to 5000. The explanatory variables (`x1` and `x2`) are generated from a multivariate normal distribution. We choose them overlapping, such that the first two variables match and the third one differs. This provides the exclusion restriction. The errors are generated from a bivariate normal distribution with correlation equal to 0.7, which leads to $\beta_\lambda = 0.7$. Finally, using the explanatory variables and the errors we compute the response variables (`y1` and `y2`). This simulation design is taken from Zhelonkin *et al.* (2016). It is a compromise between the complexity of real applications and simplicity of illustrative examples. The dataset generated from the explained procedure is not contaminated, and from the two outputs below we can see that the classical two-step estimates (**sampleSelection**) and the robust estimates are close to the true parameters and to each other. The standard errors of the robust estimator are slightly larger than that of the classical. The results by the classical estimator are:

```
R> library("sampleSelection")
R> summary(selection(y1 ~ x1, y2 ~ x2, method = "2step"))

--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
5000 observations (2740 censored and 2260 observed)
11 free parameters (df = 4990)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02135    0.04496   0.475    0.635
x11          1.01536    0.03053  33.262   <2e-16 ***
x12          1.04424    0.03873  26.965   <2e-16 ***
x13          0.78171    0.02723  28.708   <2e-16 ***
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03607    0.04363   0.827    0.408
x21          1.47018    0.02945  49.921   <2e-16 ***
x22          0.96962    0.03677  26.371   <2e-16 ***
x23          0.47070    0.01882  25.015   <2e-16 ***
Multiple R-Squared:0.6466,        Adjusted R-Squared:0.646
   Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio  0.67460    0.05718    11.8   <2e-16 ***
sigma          0.99812         NA      NA       NA
rho            0.67587         NA      NA       NA
---
```

```
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
-------------------------------------------
```

The robust estimation gives the following output:

```
R> rob.ctrl <- heckitrob.control(weights.x1 = "hat", weights.x2 = "covMcd")
R> summary(ssmrob(y1 ~ x1, y2 ~ x2, control = rob.ctrl))

Call:
ssmrob(selection = y1 ~ x1, outcome = y2 ~ x2, control = rob.ctrl)

Heckman selection model / robust 2-step M-estimation
5000 observations: 2740 censored and 2260 observed
Probit selection equation:
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 0.003156    0.04735  0.06664  9.47e-01
x11         1.004134    0.03360 29.88000 3.15e-196 ***
x12         1.032265    0.04198 24.59000 1.66e-133 ***
x13         0.782265    0.02989 26.17000 5.51e-151 ***
Outcome equation:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   0.0362    0.04619  0.7837  4.33e-01
x21           1.4775    0.03076 48.0300  0.00e+00 ***
x22           0.9780    0.03947 24.7800 1.63e-135 ***
x23           0.4677    0.02007 23.3100 3.63e-120 ***
IMR1          0.6815    0.05532 12.3200  7.10e-35 ***
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma  1.000731
```

The default values of the robustness weights `weights.x1` and `weights.x2` are `"none"`, which means that $\omega_1(x_1)$ in (11) and $\omega(x_2, \lambda)$ in (13) are equal to 1. If one expects outliers in the explanatory variables, then the weights need to be introduced, as we did above by setting `weights.x1 = "hat"` and `weights.x2 = "covMcd"`. This choice is based on our personal experience and good performance in simulations (see the study in Zhelonkin *et al.* 2016), although one can use minimum volume ellipsoid based weights as well (option `"robCov"`).

In order to study the robustness of the estimator we introduce a contamination. With probability 0.01 we generate outliers from the degenerate distribution putting mass one at the point $x_1 = (-2, -2, -1)$, $x_2 = (-2, -2, -1)$, $(y_1, y_2) = (1, 0)$. It generates leverage outliers, which also emerge in the outcome equation since $y_1 = 1$. This contamination corresponds to the situation when some observations are extremely unlikely to be observed, but somehow appear in the sample.

```
R> uni <- runif(N, 0, 1)
R> for(i in 1:N)
+    if(uni[i] < 0.01)
```

```
+    {
+    x1[i,] <- c(-2, -2, -1)
+    x2[i,] <- c(-2, -2, -1)
+    y1[i] <- 1
+    y2[i] <- 0
+    }
```

The results of the robust estimation are as follows:

```
R> summary(ssmrob(y1 ~ x1, y2 ~ x2, control = rob.ctrl))

Call:
ssmrob(selection = y1 ~ x1, outcome = y2 ~ x2, control = rob.ctrl)

Heckman selection model / robust 2-step M-estimation
5000 observations: 2710 censored and 2290 observed
Probit selection equation:
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 0.002461    0.04769  0.05161  9.59e-01
x11         1.004201    0.03377 29.74000 2.54e-194 ***
x12         1.031806    0.04227 24.41000 1.26e-131 ***
x13         0.780878    0.03001 26.02000 2.65e-149 ***
Outcome equation:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -0.01202    0.04813 -0.2497  8.03e-01
x21          1.49916    0.03215 46.6300  0.00e+00 ***
x22          0.99972    0.04086 24.4600 3.53e-132 ***
x23          0.45297    0.02024 22.3800 6.71e-111 ***
IMR1         0.82872    0.06244 13.2700  3.38e-40 ***
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma  1.061732
```

The robust estimator is stable and does not deviate a lot from the true values of the parameters. For comparison we present the output of the classical estimator from the **sampleSelection** package.

```
R> summary(selection(y1 ~ x1, y2 ~ x2, method = "2step"))

--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
5000 observations (2710 censored and 2290 observed)
11 free parameters (df = 4990)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  0.08551    0.04077   2.097    0.036 *
x11          0.62956    0.02192  28.715   <2e-16 ***
x12          0.64723    0.03037  21.315   <2e-16 ***
x13          0.45687    0.02064  22.132   <2e-16 ***
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.47503    0.06504  -7.304 3.24e-13 ***
x21          1.60039    0.04026  39.754  < 2e-16 ***
x22          1.11239    0.04911  22.649  < 2e-16 ***
x23          0.42948    0.01891  22.706  < 2e-16 ***
Multiple R-Squared:0.6316,        Adjusted R-Squared:0.631
   Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio  1.75790    0.07602   23.12   <2e-16 ***
sigma          1.53525         NA      NA       NA
rho            1.14503         NA      NA       NA
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
--------------------------------------------
```

The parameter estimates are clearly biased. It is important to notice that both estimation stages are affected and the biases are clearly larger than those from the robust estimator. In practice this divergence between estimators should indicate a problem with the data. Moreover, the estimator of the inverse Mills ratio increases to 1.76, while with the robust estimator it is 0.83, which is much closer to the true value of 0.7. Note that `rho` is not a sample correlation. It is not bounded between -1 and 1 and can easily exceed these limits, see Greene (1981) for details. This effect can also occur due to contamination, as we see that `rho` is 1.145. Hence, one should treat this quantity with caution, while using two-step estimator.

When the exclusion restriction is not available, the influence of the contamination is stronger. The biases are larger as well as the loss of efficiency. One can verify this by repeating the analysis above erasing the line `> x2[, 3] <- rnorm(N, 1, 1)` in the data generation step. From this we can conclude that the contamination and the absence of exclusion restriction make the classical estimator particularly unstable.

We can also obtain the classical estimates using the `ssmrob()` function by using large values, e.g., 1000, of the tuning parameters `tcc` and `t.c`, and by setting the leverage weights `weights.x1` and `weights.x2` to `"none"`.

6.2. Wage offer data

The first dataset is an example from Wooldridge (2002). We consider the Example 17.6 (p. 565) about the wage offer for married women, with potential selectivity bias into the labor force. A Heckman model is used, as presented in Section 2. The dataset consists of 753 observations, with 325 (43.2%) truncated observations. The selection equation defining the labor force participation includes the following variables as `age`, education status (`educ`), non-wife income (`nwifeinc`), experience (`exper`), squared experience (`expersq`), number of children less than 6 years old (`kidslt6`), and number of children older than 6 years (`kidsge6`).

In the equation of interest the log-wage offer depends on education, experience, and squared experience. We apply the classical estimator and we obtain the following output:

```
R> data("MROZ.RAW", package = "ssmrob")
R> selectEq <- inlf ~ nwifeinc + educ + exper + expersq + age + kidslt6 +
+    kidsge6
R> outcomeEq <- lwage ~ educ + exper + expersq
R> summary(selection(selectEq, outcomeEq, data = MROZ.RAW, method = "2step"))

--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
753 observations (325 censored and 428 observed)
15 free parameters (df = 739)
Probit selection equation:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept)  0.270077   0.508593   0.531  0.59556
nwifeinc    -0.012024   0.004840  -2.484  0.01320 *
educ         0.130905   0.025254   5.183 2.81e-07 ***
exper        0.123348   0.018716   6.590 8.34e-11 ***
expersq     -0.001887   0.000600  -3.145  0.00173 **
age         -0.052853   0.008477  -6.235 7.61e-10 ***
kidslt6     -0.868328   0.118522  -7.326 6.21e-13 ***
kidsge6      0.036005   0.043477   0.828  0.40786
Outcome equation:
             Estimate  Std. Error  t value Pr(>|t|)
(Intercept) -0.5781032  0.3050062  -1.895  0.05843 .
educ         0.1090655  0.0155230   7.026 4.83e-12 ***
exper        0.0438873  0.0162611   2.699  0.00712 **
expersq     -0.0008591  0.0004389  -1.957  0.05068 .
Multiple R-Squared:0.1569,        Adjusted R-Squared:0.149
Error terms:
              Estimate  Std. Error t value Pr(>|t|)
invMillsRatio  0.03226    0.13362   0.241    0.809
sigma          0.66363         NA      NA       NA
rho            0.04861         NA      NA       NA
--------------------------------------------
```

Using the robust estimator we obtain results similar to those obtained using the classical estimator.

```
R> summary(ssmrob(selectEq, outcomeEq, data = MROZ.RAW))

Call:
ssmrob(selection = selectEq, outcome = outcomeEq, data = MROZ.RAW)

Heckman selection model / robust 2-step M-estimation
```

```
753 observations: 325 censored and 428 observed
Probit selection equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.185086  0.5215843  0.3549 7.23e-01
nwifeinc    -0.013812  0.0051413 -2.6870 7.22e-03 **
educ         0.131747  0.0263492  5.0000 5.73e-07 ***
exper        0.123029  0.0192493  6.3910 1.64e-10 ***
expersq     -0.001906  0.0006134 -3.1070 1.89e-03 **
age         -0.050790  0.0087215 -5.8240 5.76e-09 ***
kidslt6     -0.840733  0.1223745 -6.8700 6.41e-12 ***
kidsge6      0.039740  0.0453318  0.8766 3.81e-01
Outcome equation:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4720491  0.2595474 -1.8190 6.90e-02 .
educ         0.1114265  0.0132375  8.4170 3.85e-17 ***
exper        0.0366974  0.0134698  2.7240 6.44e-03 **
expersq     -0.0007016  0.0003697 -1.8980 5.77e-02 .
IMR1        -0.0495793  0.1338460 -0.3704 7.11e-01
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma  0.6655772
```

The parameter estimates obtained by classical and robust estimators are very close. The standard deviations are also close, and the test statistics are similar. The significance of the parameters remains the same. Finally, we can conclude that there is no evidence of violation of distributional assumptions and that the classical estimator provides reliable results. The test for sample selection bias is non-significant for both estimators. In absence of selection bias the data can be estimated by OLS (see Wooldridge 2002, Table 17.1).

6.3. Ambulatory expenditures data

The second example is an example considered in the book by Cameron and Trivedi (2009), p. 544–547. The data on ambulatory expenditures comes from the 2001 Medical Expenditure Panel Survey. The sample size is 3328 observations, with 526 (15.8%) zero expenditures. The set of explanatory variables contains `age`, gender (`female`), education status (`educ`), ethnicity (`blhisp`), number of chronic diseases (`totchr`), insurance status (`ins`) and `income`, where `income` is used only in selection equation as an exclusion restriction variable. Other variables enter both the selection equation and the outcome equation. First we carry out the analysis without exclusion restriction. We apply the classical estimator and we obtain the following output:

```
R> data("MEPS2001", package = "ssmrob")
R> MEPS2001 <- MEPS2001 * 1
R> selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins
R> outcomeEq <- lnambx ~ age + female + educ + blhisp + totchr + ins
R> summary(selection(selectEq, outcomeEq, data = MEPS2001, method = "2step"))
```

```
--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
3328 observations (526 censored and 2802 observed)
17 free parameters (df = 3312)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.71771    0.19247  -3.729 0.000195 ***
age          0.09732    0.02702   3.602 0.000320 ***
female       0.64421    0.06015  10.710  < 2e-16 ***
educ         0.07017    0.01134   6.186 6.94e-10 ***
blhisp      -0.37449    0.06175  -6.064 1.48e-09 ***
totchr       0.79352    0.07112  11.158  < 2e-16 ***
ins          0.18124    0.06259   2.896 0.003809 **
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.30257    0.29414  18.028  < 2e-16 ***
age          0.20212    0.02430   8.319  < 2e-16 ***
female       0.28916    0.07369   3.924 8.89e-05 ***
educ         0.01199    0.01168   1.026    0.305
blhisp      -0.18106    0.06585  -2.749    0.006 **
totchr       0.49833    0.04947  10.073  < 2e-16 ***
ins         -0.04740    0.05315  -0.892    0.373
Multiple R-Squared:0.1926,        Adjusted R-Squared:0.1906
   Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio  -0.4802     0.2907  -1.652   0.0986 .
sigma           1.2932         NA      NA       NA
rho            -0.3713         NA      NA       NA
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
--------------------------------------------
```

Using the robust two-stage estimator we obtain:

```
R> meps.ctrl <- heckitrob.control(tcc = 3.2)
R> summary(ssmrob(selectEq, outcomeEq, data = MEPS2001, control = meps.ctrl))

Call:
ssmrob(selection = selectEq, outcome = outcomeEq, data = MEPS2001,
control = meps.ctrl)

Heckman selection model / robust 2-step M-estimation
3328 observations: 526 censored and 2802 observed
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.74914    0.19507  -3.840 1.23e-04 ***
```

```
age          0.10541    0.02770   3.806 1.41e-04 ***
female       0.68741    0.06226  11.040 2.41e-28 ***
educ         0.07012    0.01147   6.116 9.62e-10 ***
blhisp      -0.39775    0.06265  -6.349 2.17e-10 ***
totchr       0.83284    0.08028  10.370 3.24e-25 ***
ins          0.18256    0.06371   2.865 4.17e-03 **
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.40154    0.27673  19.520 7.53e-85 ***
age          0.20062    0.02451   8.186 2.70e-16 ***
female       0.25501    0.06993   3.647 2.66e-04 ***
educ         0.01325    0.01162   1.141 2.54e-01
blhisp      -0.15508    0.06507  -2.383 1.72e-02 *
totchr       0.48116    0.03823  12.590 2.52e-36 ***
ins         -0.06707    0.05159  -1.300 1.94e-01
IMR1        -0.67676    0.25928  -2.610 9.05e-03 **
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma  1.317891
```

In this case we see that the robust estimates are considerably different from the classical estimator. The classical estimator returns the inverse Mills ratio coefficient of -0.48 with p value of 0.099, which is significant only at 10% level. Cameron and Trivedi (2009) raised concern about robustness issues in this case (p. 544), and that the conclusion of no selection bias was doubtful. The robust estimator returns $\beta_\lambda = -0.68$ with p value of 0.009, which is significant at 1% level. If we do not reject the hypothesis of no selection bias then the OLS should be preferred. Note that if one uses FIML estimator then the likelihood ratio test has p value of 0.36, which clearly indicates that there is no selection bias. The robust estimator is more reliable and should be preferred in such situations.

We repeat the analysis, but now include the `income` variable as an exclusion restriction.

```
R> selectEq <- dambexp ~ age + female + educ + blhisp + totchr + ins + income
R> summary(selection(selectEq, outcomeEq, data = MEPS2001, method = "2step"))

--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
3328 observations (526 censored and 2802 observed)
18 free parameters (df = 3311)
Probit selection equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.668647   0.194125  -3.444 0.000579 ***
age          0.086815   0.027456   3.162 0.001581 **
female       0.663505   0.060965  10.883  < 2e-16 ***
educ         0.061884   0.012039   5.140 2.90e-07 ***
blhisp      -0.365784   0.061909  -5.908 3.81e-09 ***
```

```
totchr       0.795750   0.071217  11.174  < 2e-16 ***
ins          0.169107   0.062930   2.687 0.007240 **
income       0.002677   0.001310   2.043 0.041131 *
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.28893    0.28852  18.331  < 2e-16 ***
age          0.20247    0.02422   8.359  < 2e-16 ***
female       0.29213    0.07258   4.025 5.82e-05 ***
educ         0.01239    0.01157   1.071  0.28427
blhisp      -0.18287    0.06534  -2.798  0.00516 **
totchr       0.50063    0.04855  10.311  < 2e-16 ***
ins         -0.04651    0.05297  -0.878  0.38002
Multiple R-Squared:0.1926,        Adjusted R-Squared:0.1906
   Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio  -0.4637     0.2826  -1.641    0.101
sigma           1.2914         NA      NA       NA
rho            -0.3591         NA      NA       NA
--------------------------------------------
```

```
R> summary(ssmrob(selectEq, outcomeEq, data = MEPS2001, control = meps.ctrl))

Call:
ssmrob(selection = selectEq, outcome = outcomeEq, data = MEPS2001,
control = meps.ctrl)

Heckman selection model / robust 2-step M-estimation
3328 observations: 526 censored and 2802 observed
Probit selection equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.700434   0.196403  -3.566 3.62e-04 ***
age          0.094589   0.028149   3.360 7.79e-04 ***
female       0.703608   0.062981  11.170 5.61e-29 ***
educ         0.062308   0.012119   5.141 2.73e-07 ***
blhisp      -0.388618   0.062800  -6.188 6.09e-10 ***
totchr       0.834053   0.080226  10.400 2.58e-25 ***
ins          0.172551   0.064032   2.695 7.04e-03 **
income       0.002535   0.001344   1.886 5.93e-02 .
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.40933    0.27291  19.820 1.96e-87 ***
age          0.20029    0.02447   8.185 2.73e-16 ***
female       0.25214    0.06995   3.605 3.13e-04 ***
educ         0.01319    0.01158   1.139 2.55e-01
blhisp      -0.15342    0.06514  -2.355 1.85e-02 *
totchr       0.47956    0.03805  12.600 2.04e-36 ***
ins         -0.06826    0.05174  -1.319 1.87e-01
```

```
IMR1        -0.68995    0.25544  -2.701 6.91e-03 **
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma  1.319788
```

In the case with exclusion restriction we obtain the same pattern as without exclusion restriction. The classical estimator underestimates the IMR coefficient and leads to non-significant test of sample selection (at 10% level).

7. Conclusion

The package **ssmrob** extends a data analyst toolbox by providing the instruments for a robust analysis of specific models with sample selectivity. Robust methods allow to deal with deviations from the assumed model. Given the well-documented sensitivity of the parametric estimators the practitioners using them ideally should also verify that the classical and robust estimators do not diverge considerably. If it happens, then the robust estimators are (typically) more reliable. In any case it should be a signal for a thorough examination of the data in order to prevent misleading conclusions.

The package contains estimators for three very popular models. However the list of models with sample selectivity is larger (see Maddala 1983, for a list of examples). The discussed two-step robust estimation framework provides the necessary background for construction of robust alternatives for other models. Moreover, the estimators discussed in this paper are used as the first steps for evaluation of various treatment effects (Heckman, Tobias, and Vytlacil 2003). The robustness of the treatment effect estimators is investigated by Naghi, Váradi, and Zhelonkin (2021), and the implementation of the robust alternatives is planned for future extensions of the package.

Acknowledgments

The authors thank Andreas Alfons and the editorial team of Journal of Statistical Software for very helpful comments, which improved the original versions of the manuscript and of the package.

References

Ahn H, Powell JL (1993). "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." *Journal of Econometrics*, **58**, 3–29. doi:10.1016/0304-4076(93)90111-h.

Amemiya T (1984). "Tobit Models: A Survey." *Journal of Econometrics*, **24**, 3–61. doi:10.1016/0304-4076(84)90074-5.

Andrews DWK, Schafgans MMA (1998). "Semiparametric Estimation of the Intercept of a Sample Selection Model." *Review of Economic Studies*, **65**, 497–517. doi:10.1111/1467-937x.00055.

Andrews I, Gentzkow M, Shapiro JM (2017). "Measuring the Sensitivity of Parameter Estimates to Estimation Moments." *Quarterly Journal of Economics*, **132**, 1553–1592. doi:10.1093/qje/qjx023.

Arellano M, Bonhomme S (2017). "Quantile Selection Models with an Application to Understanding Changes in Wage Inequality." *Econometrica*, **85**, 1–28. doi:10.3982/ecta14030.

Athey S, Imbens GW (2015). "A Measure of Robustness to Misspecification." *American Economic Review*, **105**, 476–480. doi:10.1257/aer.p20151020.

Athey S, Imbens GW (2017). "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives*, **31**, 3–32. doi:10.1257/jep.31.2.3.

Buchinsky M (1998). "The Dynamics of Changes in the Female Wage Distribution in the USA: A Quantile Regression Approach." *Journal of Applied Econometrics*, **13**, 1–30. doi:10.1002/(sici)1099-1255(199801/02)13:1<1::aid-jae474>3.0.co;2-a.

Cameron CA, Trivedi PK (2009). *Microeconometrics Using Stata*. Stata Press, College Station.

Cantoni E, Ronchetti E (2001). "Robust Inference for Generalized Linear Models." *Journal of the American Statistical Association*, **96**, 1022–1030. doi:10.1198/016214501753209004.

Das M, Newey WK, Vella F (2003). "Nonparametric Estimation of Sample Selection Models." *Review of Economic Studies*, **70**, 33–58. doi:10.1111/1467-937x.00236.

Eicker F (1967). "Limit Theorems for Regression with Unequal and Dependent Errors." In LM LeCam, J Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 59–82. University of California Press, Berkeley.

Gallant RA, Nychka DW (1987). "Semi-Nonparametric Maximum Likelihood Estimation." *Econometrica*, **55**, 363–390. doi:10.2307/1913241.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2020). ***mvtnorm**: Multivariate Normal and t Distributions*. R package version 1.1-1, URL https://CRAN.R-project.org/package=mvtnorm.

Greene WH (1981). "Sample Selection Bias as a Specification Error: Comment." *Econometrica*, **49**, 795–798. doi:10.2307/1911523.

Hampel F (1974). "The Influence Curve and Its Role in Robust Estimation." *Journal of the American Statistical Association*, **69**, 383–393. doi:10.1080/01621459.1974.10482962.

Hampel F, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Heckman J (1979). "Sample Selection Bias as a Specification Error." *Econometrica*, **47**, 153–161. doi:10.2307/1912352.

Heckman J (1990). "Varieties of Selection Bias." *American Economic Review*, **80**, 313–318. doi:10.1007/978-1-349-20570-7_29.

Heckman JJ, Tobias JL, Vytlacil E (2003). "Simple Estimators for Treatment Parameters in a Latent Variable Framework." *Review of Economics and Statistics*, **85**, 748–755. doi:10.1162/003465303322369867.

Heckman JJ, Vytlacil EJ (2007). "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." *Handbook of Econometrics*, **6B**, 4779–4874. doi:10.1016/s1573-4412(07)06070-9.

Huber M, Melly B (2015). "A Test of the Conditional Independence Assumption in Sample Selection Models." *Journal of Applied Econometrics*, **30**, 1144–1168. doi:10.1002/jae.2431.

Huber PJ (1967). "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions." In LM LeCam, N J (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233. University of California Press, Berkeley.

Klein RW, Spady RH (1993). "Efficient Semiparametric Estimator for Binary Response Models." *Econometrica*, **61**, 387–421. doi:10.2307/2951556.

Koenker R (2005). *Quantile Regression*. Cambridge University Press, New York.

Koenker R, Bassett G (1978). "Regression Quantiles." *Econometrica*, **46**, 33–50. doi:10.2307/1913643.

Lee LF (1983). "Generalized Econometric Models with Selectivity." *Econometrica*, **51**, 507–512. doi:10.2307/1912003.

Leung SF, Yu S (2000). "Collinearity and Two-Step Estimation of Sample Selection Models: Problems, Origins, and Remedies." *Computational Economics*, **15**, 173–199. doi:10.1023/a:1008749011772.

Maddala GS (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York.

Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Conceicao ELT, Anna di Palma M (2021). ***robustbase**: Basic Robust Statistics*. R package version 0.93-8, URL https://CRAN.R-project.org/package=robustbase.

Manski CF (1989). "Anatomy of the Selection Problem." *The Journal of Human Resources*, **24**, 343–360. doi:10.2307/145818.

Marchenko YV, Genton MG (2012). "A Heckman Selection-t Model." *Journal of the American Statistical Association*, **107**, 304–317. doi:10.1080/01621459.2012.656011.

Naghi AA, Váradi M, Zhelonkin M (2021). "Robust Estimation of Treatment Effects in a Latent-Variable Framework." Working Paper.

Newey WK (2009). "Two-Step Series Estimation of Sample Selection Models." *Econometrics Journal*, **12**, S217–S229. doi:10.1111/j.1368-423x.2008.00263.x.

Ogundimu EO, Hutton JL (2016). "A Sample Selection Model with Skew-Normal Distribution." *Scandinavian Journal of Statistics*, **43**, 172–190. doi:10.1111/sjos.12171.

Pagan A, Ullah A (1999). *Nonparametric Econometrics*. Cambridge University Press.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rousseeuw PJ (1985). "Multivariate Estimation with High Breakdown Point." In W Grossmann, G Pflug, I Vincze, W Wertz (eds.), *Mathematical Statistics and Applications*, pp. 283–297. Reidel, Dordrecht.

Rousseeuw PJ, Van Driessen K (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator." *Technometrics*, **41**, 212–223. doi:10.1080/00401706.1999.10485670.

SAS Institute Inc (2014). *SAS/STAT Software, Version 13.2*. Cary. URL https://www.sas.com/.

Smith MD (2003). "Modelling Sample Selection Using Archimedean Copulas." *Econometrics Journal*, **6**, 99–123. doi:10.1111/1368-423x.00101.

Stapleton DC, Young DJ (1984). "Censored Normal Regression with Measurement Error on the Dependent Variable." *Econometrica*, **52**, 737–760. doi:10.2307/1913474.

StataCorp (2017). *Stata Statistical Software: Release 15*. StataCorp LLC, College Station. URL http://www.stata.com/.

Todorov V, Filzmoser P (2009). "An Object-Oriented Framework for Robust Multivariate Analysis." *Journal of Statistical Software*, **32**(3), 1–47. doi:10.18637/jss.v032.i03.

Toomet O, Henningsen A (2008). "Sample Selection Models in R: Package **SampleSelection**." *Journal of Statistical Software*, **27**, 1–23. doi:10.18637/jss.v027.i07.

Vella F (1998). "Estimating Models with Sample Selection Bias: A Survey." *The Journal of Human Resources*, **33**, 127–169. doi:10.2307/146317.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag. doi:10.1007/978-0-387-21706-2.

White H (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, **48**, 817–838. doi:10.2307/1912934.

Wojtyś M, Marra G, Radice R (2016). "Copula Regression Spline Sample Selection Models: The R Package **SemiParSampleSel**." *Journal of Statistical Software*, **71**, 1–66. doi:10.18637/jss.v071.i06.

Wooldridge JM (2002). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge.

Zhelonkin M, Genton MG, Ronchetti E (2012). "On the Robustness of Two-Stage Estimators." *Statistics & Probability Letters*, **82**, 726–732. doi:10.1016/j.spl.2011.12.014.

Zhelonkin M, Genton MG, Ronchetti E (2016). "Robust Inference in Sample Selection Models." *Journal of the Royal Statistical Society B*, **78**, 805–827. doi:10.1111/rssb.12136.

Zhelonkin M, Genton MG, Ronchetti E (2021). **ssmrob**: *Robust Estimation and Inference in Sample Selection Models*. R package version 1.0, URL https://CRAN.R-project.org/package=ssmrob.

A. Endogenous treatment model: IF of the two-step estimator

We derive the IF of the classical two-step estimator (probit MLE/OLS) of the endogenous treatment model. The first estimation step is the probit MLE. Let S be its estimating functional. Then, its IF is given by

$$IF(z; S, F) = \left[\int \frac{\phi(x_1^\top \beta_1)^2 x_1 x_1^\top}{\Phi(x_1^\top \beta_1)\{1 - \Phi(x_1^\top \beta_1)\}} dF \right]^{-1} \{y_1 - \Phi(x_1^\top \beta_1)\} \frac{\phi(x_1^\top \beta_1)}{\Phi(x_1^\top \beta_1)\{1 - \Phi(x_1^\top \beta_1)\}} x_1.$$

The inverse Mills ratio for the complete sample is defined by

$$\lambda^C\{z_1; S(F)\} = y_1 \left\{ \frac{\phi(x_1^\top \beta_1)}{\Phi(x_1^\top \beta_1)} \right\} + (1 - y_1) \left\{ \frac{-\phi(x_1^\top \beta_1)}{1 - \Phi(x_1^\top \beta_1)} \right\}.$$

The OLS score function is

$$\Psi_2[z_2; \lambda^C\{z_1; S(F)\}, T(F)] = (y_2 - x_2^\top \beta_2 - \alpha y_1 - \beta_\lambda \lambda^C) \begin{pmatrix} x_2 \\ y_1 \\ \lambda^C \end{pmatrix}.$$

Using the result for the two-step M-estimators (Zhelonkin, Genton, and Ronchetti 2012) we get the IF of the second step

$$IF(z; T, F) = \left[\int \begin{Bmatrix} x_2 x_2^\top & x_2 y_1 & x_2 \lambda^C \\ x_2^\top y_1 & y_1^2 & y_1 \lambda^C \\ x_2^\top \lambda^C & y_1 \lambda^C & (\lambda^C)^2 \end{Bmatrix} dF \right]^{-1} \left\{ (y_2 - x_2^\top \beta_2 - \alpha y_1 - \beta_\lambda \lambda^C) \begin{pmatrix} x_2 \\ y_1 \\ \lambda^C \end{pmatrix} \int \begin{pmatrix} x_2 \beta_\lambda \\ y_1 \beta_\lambda \\ \lambda^C \beta_\lambda \end{pmatrix} (\lambda^C)' dF \cdot IF(z; S, F) \right\},$$

where

$$(\lambda^C)' = y_1 \left\{ \frac{-\phi(x_1^\top \beta_1)\Phi(x_1^\top \beta_1) x_1^\top \beta_1 - \phi(x_1^\top \beta_1)^2}{\Phi(x_1^\top \beta_1)^2} x_1^\top \right\} + (1 - y_1) \left[\frac{\{1 - \Phi(x_1^\top \beta_1)\}\phi(x_1^\top \beta_1) x_1^\top \beta_1 - \phi(x_1^\top \beta_1)^2}{\{1 - \Phi(x_1^\top \beta_1)\}^2} x_1^\top \right].$$

B. Endogenous treatment model: Simulation study

We illustrate the performance of the proposed estimator in a simulation study. For the data generating process we use a modification of the setup used in Section 6.1. We generate $y_{1i} = I(x_{1i}^\top (1, 1, 0.75)^\top + e_{1i} > 0)$, where $x_{1i} \sim N\{(0, -1, 1)^\top; diag(1, 0.25, 1)\}$. The outcome equation is $y_{2i} = x_{2i}^\top (1.5, 1, 0.5)^\top + 1.25 y_{1i} + e_{2i}$, where the x's are the same $x_2 = x_1$ if the exclusion restriction is not available, and if it is available, then the last explanatory variable of x_2 is generated independently of x_1 from the same distribution. The errors e_1 and e_2 follow a zero mean bivariate normal distribution with $\sigma_1 = \sigma_2 = 1$ and $\rho = -0.7$. The sample size

	With exclusion restriction					
	Not contaminated		$y_1 = 0$		$y_1 = 1$	
Estimator	Bias	MSE	Bias	MSE	Bias	MSE
FIML	0.000	0.013	−0.890	0.822	−0.999	1.025
Classical two-step	0.001	0.023	−1.022	1.137	−1.321	1.894
Robust two-step	0.025	0.028	0.004	0.027	−0.022	0.027

	Without exclusion restriction					
	Not contaminated		$y_1 = 0$		$y_1 = 1$	
Estimator	Bias	MSE	Bias	MSE	Bias	MSE
FIML	−0.012	0.024	−1.436	2.253	−2.164	8.357
Classical two-step	−0.022	0.088	−3.720	15.125	−5.111	28.763
Robust two-step	0.022	0.116	−0.038	0.115	−0.110	0.120

Table 1: Bias and mean-squared error (MSE) of the FIML, classical 2-step and robust 2-step estimators of treatment effect parameter α at the model and under two types of contamination, when x is contaminated and corresponding $y_1 = 0$ (columns 4 and 5) and $y_1 = 1$ (columns 6 and 7).

is 1000 and we repeat the experiment 500 times. For other sample sizes (500 or 5000) the results are similar.

We estimate the model without contamination and with two types of contamination. In the first scenario we have outliers in the control group, i.e., x_1 is contaminated when $y_1 = 0$. To generate the outliers, we replace the original observations with probability 0.01 by a point mass at $(y_1, y_2) = (0, 1)$ and $x_1 = x_2 = (2, 0, 3)$. In this case the observation is unlikely to be in the control group but appears there. In the second scenario we have outliers in the treatment group, i.e., $y_1 = 1$. The mechanism of contamination is the same, but the point mass is at $(y_1, y_2) = (1, 0)$ and $x_1 = x_2 = (-2, -2, -1)$. In this case the observation should be in the control group but appears in the treatment group.

We compare the classical two-step estimator, full information maximum likelihood (FIML) and our robust two-step estimator. In Table 1 and Figure 1 we report the estimation's results of the treatment effect parameter α. All three estimators perform well at the model. The biases are close to zero, FIML is the most efficient and the robust two-step estimator is a bit less efficient than the classical two-step estimator. When the contamination is added, both FIML and classical two-step break down. They are outperformed by the robust estimator in terms of bias and efficiency. Notice that the biases are negative, which in the context of treatment effects can lead to a non-significant effect when it is present. The robust estimator allows some bias (it is clearly visible when the exclusion restriction is not available), but it is small relatively to the classical estimators. In Table 2 (with exclusion restriction) and Table 3 (without exclusion restriction) we report the estimation's results of the other parameters. Similarly, all estimators perform well without contamination. FIML is the most efficient, then the classical two-step, then the robust two-step, as expected from the theory. Under contamination both FIML and the classical two-step are biased with inflated variances, while the robust estimator remains stable.

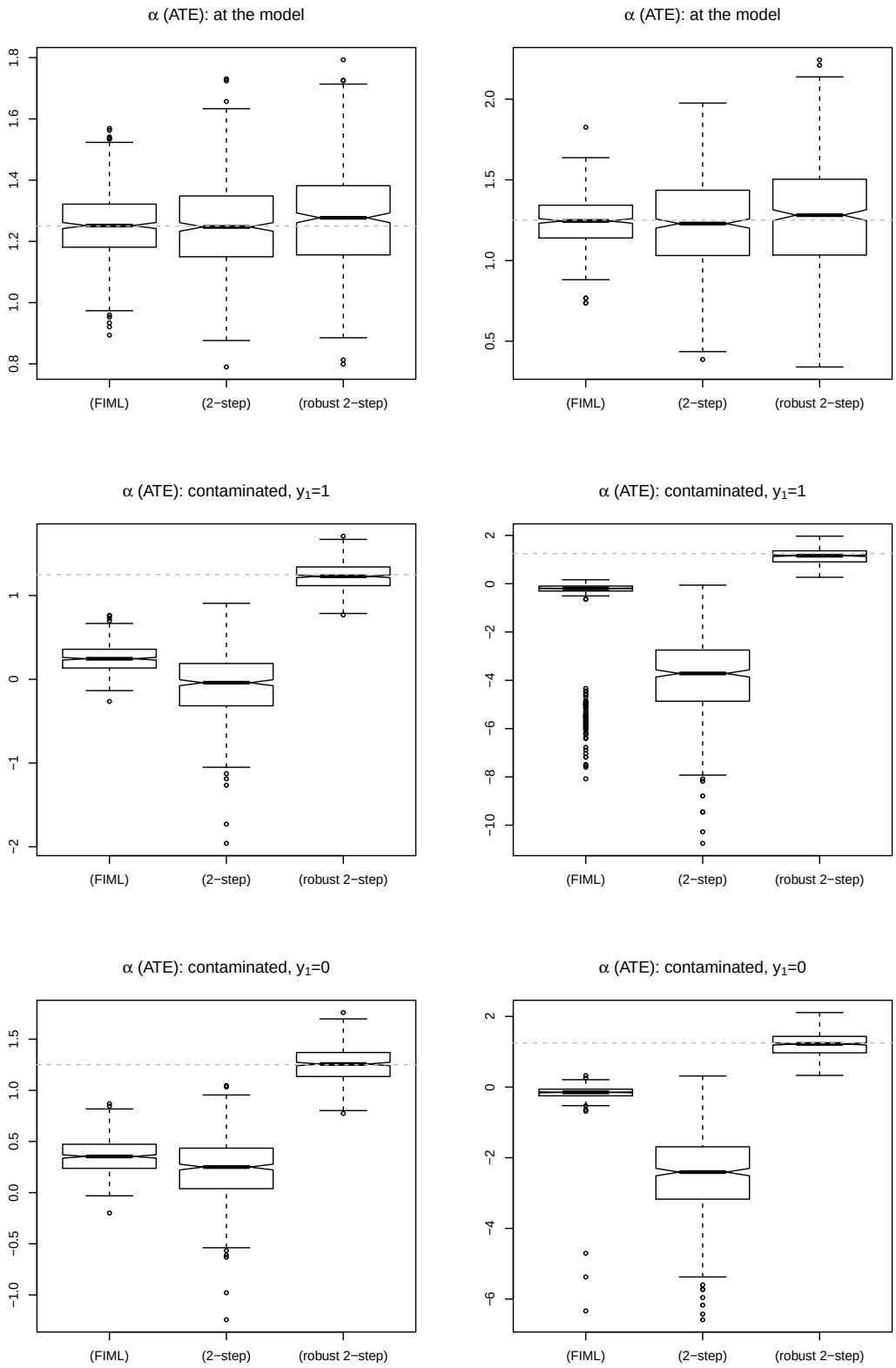

Figure 1: Boxplots of the estimated average treatment effect (ATE), parameter α in (4). Left panels correspond to the case with exclusion restriction, right panels correspond to the case without exclusion restriction.

	Not contaminated		$y_1 = 0$		$y_1 = 1$	
Parameter	Bias	MSE	Bias	MSE	Bias	MSE
FIML						
β_{20}	−0.001	0.012	0.470	0.233	0.609	0.384
β_{21}	0.000	0.002	0.130	0.018	0.137	0.020
β_{22}	−0.001	0.005	0.045	0.007	0.035	0.007
β_{23}	0.001	0.001	−0.072	0.006	−0.082	0.008
Classical two-step						
β_{20}	−0.002	0.017	0.545	0.326	0.803	0.697
β_{21}	0.000	0.002	0.157	0.028	0.204	0.046
β_{22}	−0.001	0.005	0.068	0.010	0.092	0.014
β_{23}	0.001	0.001	−0.068	0.006	−0.074	0.006
β_{λ}	0.001	0.009	0.928	0.918	1.175	1.469
Robust two-step						
β_{20}	−0.014	0.020	−0.003	0.019	0.013	0.019
β_{21}	−0.002	0.002	0.002	0.003	0.006	0.003
β_{22}	−0.002	0.006	0.001	0.006	0.004	0.006
β_{23}	0.002	0.001	0.001	0.001	−0.000	0.001
β_{λ}	−0.004	0.012	0.018	0.012	0.043	0.014

Table 2: Bias and mean-squared error (MSE) of the FIML, classical two-step and robust two-step estimators of endogenous treatment model parameters at the model and under two types of contamination, when x is contaminated and the corresponding $y_1 = 0$ (columns 4 and 5) and $y_1 = 1$ (columns 6 and 7). The exclusion restriction is available.

C. Tobit-5 model: Example

Similarly to the Tobit-2 model, we generate the data using the same algorithm.

```
R> set.seed(2)
R> N <- 5000
R> beta1 <- c(0, 1.0, 1.0, 0.75)
R> beta21 <- c(0, 1.5, 1.0, 0.5)
R> beta22 <- c(1, -1.5, 1.0, 0.5)
R> covm <- diag(3)
R> covm[lower.tri(covm)] <- c(0.75, 0.5, 0.25)
R> covm[upper.tri(covm)] <- covm[lower.tri(covm)]
R> eps <- rmvnorm(N, rep(0, 3), covm)
R> x1 <- rmvnorm(N, mean = c(0, -1, 1), sigma = diag(c(1, 0.5, 1)))
R> x21 <- x1
R> x22 <- x1
R> x21[, 3] <- rnorm(N, 1, 1)
R> x22[, 3] <- rnorm(N, 1, 1)
R> y1 <- ifelse(cbind(1, x1) %*% beta1 + eps[, 1] > 0, 1, 0)
R> y2 <- ifelse(y1 > 0.5, cbind(1, x21) %*% beta21 + eps[, 2],
+    cbind(1, x22) %*% beta22 + eps[, 3])
```

The DGP is similar to the Tobit-2 case, but with minor modifications. We generate two

	Not contaminated		$y_1 = 0$		$y_1 = 1$	
Parameter	Bias	MSE	Bias	MSE	Bias	MSE
FIML						
β_{20}	0.007	0.012	0.569	0.367	1.025	1.963
β_{21}	0.000	0.003	0.236	0.064	0.363	0.262
β_{22}	0.003	0.005	0.138	0.029	0.228	0.141
β_{23}	0.003	0.002	0.153	0.028	0.235	0.114
Classical two-step						
β_{20}	0.012	0.029	1.607	2.803	2.438	6.530
β_{21}	0.003	0.006	0.706	0.538	0.969	1.013
β_{22}	0.006	0.010	0.539	0.315	0.734	0.581
β_{23}	0.005	0.004	0.488	0.258	0.667	0.480
β_{λ}	0.013	0.029	2.493	6.742	3.370	12.408
Robust two-step						
β_{20}	−0.010	0.036	0.018	0.035	0.054	0.036
β_{21}	−0.003	0.008	0.010	0.008	0.025	0.008
β_{22}	−0.001	0.011	0.012	0.011	0.027	0.012
β_{23}	0.001	0.005	0.010	0.005	0.021	0.005
β_{λ}	−0.002	0.042	0.043	0.043	0.094	0.048

Table 3: Bias and mean-squared error (MSE) of the FIML, classical two-step and robust two-step estimators of endogenous treatment model parameters at the model and under two types of contamination, when x is contaminated and the corresponding $y_1 = 0$ (columns 4 and 5) and $y_1 = 1$ (columns 6 and 7). The exclusion restriction is not available.

explanatory variables for the outcome equation, namely `x21` and `x22` for the first and second regimes respectively. The error terms follow a multivariate normal distribution (3). The response variable (`y2`) in the outcome equation has two regimes, depending on the selection variable `y1`. Without contamination we have the following output:

```
R> summary(ssmrob(y1 ~ x1, list(y2 ~ x21, y2 ~ x22), control = rob.ctrl))

Call:
ssmrob(selection = y1 ~ x1, outcome = list(y2 ~ x21, y2 ~ x22),
control = rob.ctrl)

Switching regression model / robust 2-step M-estimation
5000 observations: 2266 in regime 1 and 2734 in regime 2
Probit selection equation:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.01606    0.04685  0.3429  7.32e-01
x11          0.92441    0.03130 29.5300 1.12e-191 ***
x12          0.97669    0.03993 24.4600 3.98e-132 ***
x13          0.75138    0.02952 25.4600 6.04e-143 ***
Outcome equation, regime 1:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -0.007297    0.04452 -0.1639  8.70e-01
```

```
x211         1.534097    0.02866 53.5200  0.00e+00 ***
x212         0.992513    0.03807 26.0700 8.14e-150 ***
x213         0.477753    0.01863 25.6500 4.78e-145 ***
IMR1         0.771236    0.06026 12.8000  1.67e-37 ***
Outcome equation, regime 2:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   0.9769    0.07628  12.810  1.50e-37 ***
x221         -1.5003    0.02948 -50.890  0.00e+00 ***
x222          0.9823    0.03473  28.290 5.19e-176 ***
x223          0.4798    0.01922  24.960 1.72e-137 ***
IMR2          0.4798    0.06602   7.268  3.66e-13 ***
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma: 0.9988461 in regime 1 and 0.9970672 in regime 2
```

The estimates are close to the true values of the parameters. Next, we introduce the contamination. With probability 0.01 we introduce the leverage outliers in the selection equations, such that they appear in the equation of interest in the second regime.

```
R> uni <- runif(N, 0, 1)
R> for(i in 1:N)
+    if(uni[i] < 0.01)
+    {
+    x1[i,] <- c(-2, -2, -1)
+    x21[i,] <- c(-2, -2, -1)
+    y1[i] <- 1
+    y2[i] <- 0
+    }
```

We estimate the contaminated sample and obtain the following output:

```
R> summary(ssmrob(y1 ~ x1, list(y2 ~ x21, y2 ~ x22), control = rob.ctrl))

Call:
ssmrob(selection = y1 ~ x1, outcome = list(y2 ~ x21, y2 ~ x22),
control = rob.ctrl)

Switching regression model / robust 2-step M-estimation
5000 observations: 2283 in regime 1 and 2717 in regime 2
Probit selection equation:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.01993    0.04695  0.4245  6.71e-01
x11          0.92290    0.03137 29.4200 3.52e-190 ***
x12          0.97700    0.04002 24.4100 1.22e-131 ***
x13          0.74967    0.02956 25.3600 6.24e-142 ***
Outcome equation, regime 1:
```

```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -0.04405    0.04576 -0.9626  3.36e-01
x211         1.55147    0.02933 52.8900  0.00e+00 ***
x212         1.01207    0.03908 25.9000 6.78e-148 ***
x213         0.47300    0.01882 25.1300 2.33e-139 ***
IMR1         0.86789    0.06432 13.4900  1.69e-41 ***
Outcome equation, regime 2:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   0.9769    0.07686  12.710  5.26e-37 ***
x221         -1.4968    0.02971 -50.390  0.00e+00 ***
x222          0.9816    0.03499  28.050 3.76e-173 ***
x223          0.4795    0.01931  24.830 3.97e-136 ***
IMR2          0.4827    0.06632   7.279  3.36e-13 ***
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma: 1.045347 in regime 1 and 0.9984809 in regime 2
```

The estimator is stable. Of course, it is affected by the contamination, but the bias is controlled and can be slightly reduced. The estimator of the first regime remains the same. To compare, below we give the output of the classical estimator obtained by our package.

```
R> clas.ctrl <- heckitrob.control(tcc = 1000, t.c = 1000)
R> summary(ssmrob(y1 ~ x1, list(y2 ~ x21, y2 ~ x22), control = clas.ctrl))

Call:
ssmrob(selection = y1 ~ x1, outcome = list(y2 ~ x21, y2 ~ x22),
control = clas.ctrl)

Switching regression model / robust 2-step M-estimation
5000 observations: 2283 in regime 1 and 2717 in regime 2
Probit selection equation:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.06967    0.04152   1.678  9.34e-02 .
x11          0.71948    0.02429  29.620 7.42e-193 ***
x12          0.74956    0.03219  23.290 5.91e-120 ***
x13          0.54595    0.02326  23.470 7.42e-122 ***
Outcome equation, regime 1:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -0.2977    0.05622  -5.295  1.19e-07 ***
x211          1.6346    0.03379  48.380  0.00e+00 ***
x212          1.1074    0.04308  25.710 9.58e-146 ***
x213          0.4561    0.01855  24.590 1.66e-133 ***
IMR1          1.4404    0.07788  18.500  2.25e-76 ***
Outcome equation, regime 2:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   1.0791    0.08887  12.140  6.24e-34 ***
```

```
x221         -1.4876    0.03128 -47.560  0.00e+00 ***
x222          1.0042    0.03648  27.530 7.53e-167 ***
x223          0.4841    0.01866  25.940 2.54e-148 ***
IMR2          0.5934    0.08127   7.302  2.84e-13 ***
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma: 1.320736 in regime 1 and 1.008994 in regime 2
```

The estimators of both regimes are seriously affected by the contamination.

D. Endogenous treatment model: Example

We generate the data using the data generating process from Appendix B. We only increased the sample size to 5000. The true value of the treatment effect is $\alpha = 1.25$

```
R> set.seed(2)
R> N <- 5000
R> beta1 <- c(0, 1.0, 1.0, 0.75)
R> beta2 <- c(0, 1.5, 1.0, 0.5)
R> alpha <- 1.25
R> x1 <- rmvnorm(N, mean = c(0, -1, 1), sigma = diag(c(1, 0.5, 1)))
R> x2 <- x1
R> x2[, 3] <- rnorm(N, 1, 1)
R> covmtrx <- matrix(c(1, -0.7, -0.7, 1), 2, 2)
R> eps <- rmvnorm(N, mean = rep(0, 2), sigma = covmtrx)
R> y1 <- ifelse(cbind(1, x1) %*% beta1 + eps[, 1] > 0, 1, 0)
R> y2 <- cbind(1, x2) %*% beta2 + alpha * y1 + eps[,2]
```

First we estimate the model without contamination:

```
R> summary(etregrob(y1 ~ x1, y2 ~ x2, control = rob.ctrl))

Call:
etregrob(selection = y1 ~ x1, outcome = y2 ~ x2, control = rob.ctrl)

Endogenous treatment model / Robust 2-step M-estimation
5000 observations
Probit selection equation:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -0.008046    0.04797 -0.1677  8.67e-01
x11          1.034998    0.03467 29.8500 7.68e-196 ***
x12          1.088900    0.04342 25.0800 8.11e-139 ***
x13          0.798620    0.03056 26.1300 1.46e-150 ***
Outcome equation:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.01441    0.05930   0.243  8.08e-01
```

```
x21          1.49460    0.02313  64.630  0.00e+00 ***
x22          1.00690    0.02893  34.800 2.31e-265 ***
x23          0.49567    0.01545  32.080 7.57e-226 ***
YS           1.26905    0.07573  16.760  4.96e-63 ***
CIMR        -0.73426    0.05500 -13.350  1.17e-40 ***
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma  1.010713
```

The estimated treatment effect $\hat{\alpha}$ is 1.269, the estimated parameter for the control function (CIMR - complete inverse Mills ratio) is -0.734 with a true value equal to -0.7. Now we add the contamination:

```
R> uni <- runif(N,0,1)
R> for(i in 1:N)
+    if(uni[i] < 0.01)
+    {
+    x1[i,] <- c(-2, -2, -1)
+    x2[i,] <- c(-2, -2, -1)
+    y1[i] <- 1
+    y2[i] <- 0
+    }
```

The output of the robust estimator is as follows:

```
R> summary(etregrob(y1 ~ x1, y2 ~ x2, control = rob.ctrl))

Call:
etregrob(selection = y1 ~ x1, outcome = y2 ~ x2, control = rob.ctrl)

Endogenous treatment model / Robust 2-step M-estimation
5000 observations
Probit selection equation:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -0.0110    0.04819 -0.2283  8.19e-01
x11           1.0351    0.03485 29.7100 6.61e-194 ***
x12           1.0938    0.04367 25.0400 2.06e-138 ***
x13           0.7999    0.03072 26.0400 1.77e-149 ***
Outcome equation:
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  0.03458    0.05924   0.5838  5.59e-01
x21          1.50006    0.02302  65.1600  0.00e+00 ***
x22          1.01150    0.02895  34.9500 1.51e-267 ***
x23          0.49390    0.01550  31.8700 6.13e-223 ***
YS           1.24208    0.07576  16.3900  2.09e-60 ***
CIMR        -0.70329    0.05507 -12.7700  2.42e-37 ***
```

```
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma  1.206593
```

The output of the robust estimator is very close to the one without contamination. The output of the classical estimator (obtained by `etregrob()` with increased tuning constants):

```
R> summary(etregrob(y1 ~ x1, y2 ~ x2, control = clas.ctrl))

Call:
etregrob(selection = y1 ~ x1, outcome = y2 ~ x2, control = clas.ctrl)

Endogenous treatment model / Robust 2-step M-estimation
5000 observations
Probit selection equation:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.05371    0.04091   1.313  1.89e-01
x11          0.71322    0.02431  29.330 3.72e-189 ***
x12          0.75762    0.03237  23.410 3.72e-121 ***
x13          0.53371    0.02259  23.620 2.36e-123 ***
Outcome equation:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   0.7807    0.10387  7.5160  5.63e-14 ***
x21           1.6564    0.02764 59.9200  0.00e+00 ***
x22           1.1891    0.03279 36.2600 5.87e-288 ***
x23           0.4386    0.01492 29.4000 5.08e-190 ***
YS            0.1460    0.14961  0.9756  3.29e-01
CIMR          0.2960    0.12164  2.4340  1.49e-02 *
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

sigma  1.041848
```

The effect of contamination is clearly pronounced. The treatment effect parameter (0.146) is biased towards zero and is not significant. The other parameters in both equations are also biased.

Affiliation:

Mikhail Zhelonkin
Econometric Institute
Erasmus University Rotterdam
Burg. Oudlaan 50
3062PA Rotterdam, The Netherlands
E-mail: Zhelonkin@ese.eur.nl

Elvezio Ronchetti
Research Center for Statistics and GSEM
University of Geneva
Blv. Pont d'Arve 40
CH-1211 Geneva, Switzerland
E-mail: Elvezio.Ronchetti@unige.ch

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
August 2021, Volume 99, Issue 4 *Submitted:* 2019-05-16
doi:10.18637/jss.v099.i04 *Accepted:* 2020-07-23