Robust Inference in Sample Selection Models

Mikhail Zhelonkin*, Marc G. Genton† and Elvezio Ronchetti‡

June 11, 2015

Abstract

The problem of non-random sample selectivity often occurs in practice in many different fields. The classical estimators introduced by Heckman (1979) are the backbone of the standard statistical analysis of these models. However, these estimators are very sensitive to small deviations from the distributional assumptions which are often not satisfied in practice. In this paper we develop a general framework to study the robustness properties of estimators and tests in sample selection models. We derive the influence function and the change-of-variance function of the Heckman's two-stage estimator, and demonstrate the non-robustness of this estimator and its estimated variance to small deviations from the assumed model. We propose a procedure for robustifying the estimator, prove its asymptotic normality and give its asymptotic variance. Both cases with and without an exclusion restriction are covered. This allows us to construct a simple robust alternative to the sample selection bias test. We illustrate the use of our new methodology in an analysis of ambulatory expenditures and compare the performance of the classical and robust methods in a Monte Carlo simulation study.

Keywords: Change-of-variance function; Heckman model; Influence function; M-estimator; Robust estimation; Robust inference; Sample selection; Two-stage estimator.

1 Introduction

A sample selectivity problem occurs when an investigator observes a non-random sample of a population, that is, when the observations are present according to some selection rule. Consider, for instance, the analysis of consumer expenditures, where typically the spending amount is related to the decision to spend. More specifically, the selection bias arises if controlling for

*Department of Operations, University of Lausanne, UNIL-Dorigny Anthropole, CH-1005 Lausanne, Switzerland. E-mail: Mikhail.Zhelonkin@unil.ch.

†CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. E-mail: Marc.Genton@kaust.edu.sa

‡Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Switzerland. E-mail: Elvezio.Ronchetti@unige.ch

explanatory variables, the spending amount is not independent from the decision to spend, i.e. they are dependent through unobservables. This type of problems arise in many research fields besides economics, including sociology, political science, finance, and many others (see e.g., Winship and Mare 1992, Collier and Mahoney 1996, Bushway et al. 2007, Lennox et al. 2012, and references therein for various applications).

A sample selection model can be represented by the following regression system

$$y_{1i}^* = x_{1i}^T \beta_1 + e_{1i}, \quad (1)$$

$$y_{2i}^* = x_{2i}^T \beta_2 + e_{2i}, \quad (2)$$

where the responses y_{1i}^* and y_{2i}^* are unobserved latent variables, x_{ji} is a vector of explanatory variables, β_j is a $p_j \times 1$ vector of parameters, $j = 1, 2$, and the error terms follow a bivariate normal distribution with variances $\sigma_1^2 = 1$, σ_2^2 , and correlation ρ . Notice that 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. The observed variables are defined by

$$y_{1i} = I(y_{1i}^* > 0), \quad (3)$$

$$y_{2i} = y_{2i}^* I(y_{1i}^* > 0), \quad (4)$$

where I is the indicator function. This model is classified by Amemiya (1984) as a ‘‘Tobit type-2 model’’ or ‘‘Heckman model’’ originally discussed by Heckman (1979) in his seminal paper.

If all the data were available, or the data were missing at random, i.e. no selection mechanism was involved, we could estimate the model by Ordinary Least Squares (OLS). But in general these conditions are not satisfied and the OLS estimator is biased and inconsistent.

Heckman (1979) proposed two estimation procedures for this model. The first one is a Maximum Likelihood Estimator (MLE) based on the assumption of bivariate normality of the error terms. The second one is a two-stage procedure. Consider the conditional expectation

of y_{2i} given x_{2i} and the selection rule

$$E(y_{2i}|x_{2i}, y_{1i}^* > 0) = x_{2i}^T \beta_2 + E(e_{2i}|e_{1i} > -x_{1i}^T \beta_1).$$

Then the conditional expectation of the error term is in general different from zero, which leads to the following modified regression

$$y_{2i} = x_{2i}^T \beta_2 + \beta_\lambda \lambda(x_{1i}^T \beta_1) + v_i, \quad (5)$$

where $\beta_\lambda = \rho\sigma_2$, $\lambda(x_{1i}^T \beta_1) = \phi(x_{1i}^T \beta_1)/\Phi(x_{1i}^T \beta_1)$ is the Inverse Mills Ratio (IMR), v_i is the error term with zero expectation, and $\phi(\cdot)$ denotes the density and $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution, respectively. Heckman (1979) proposed then to estimate β_1 in the first stage by probit MLE and to compute estimated values of λ , and in a second stage to use OLS in (5), where the additional variable corrects for the sample selection bias.

Both estimation procedures have advantages and drawbacks, studied extensively in the literature; see e.g. Stolzenberg and Relles (1997), Puhani (2000), Toomet and Henningsen (2008), and the general reviews by Winship and Mare (1992), Vella (1998) and references therein. An important criticism is their sensitivity to the normality assumption of the error terms, which is often violated in practice, which is well documented by several Monte Carlo studies, where the behavior of these estimators have been investigated under different distributional assumptions; see Paarsch (1984) and Zuehlke and Zeman (1991). Another important issue is the presence of outlying observations, a well-known problem in many classical models, which is often encountered in practice. For instance, an individual having several chronic diseases with zero medical expenditures, or a woman with several young children present in the labor force (see the sensitivity analysis of wage offer data in Zhelonkin 2013) would generate observations with high leverage. Outliers can be gross errors or legitimate extreme observations (perhaps coming from a longtailed distribution). In both cases it is of interest to identify them and this can be difficult by using only classical estimators.

A first simple strategy to tackle this problem is to develop (robust) misspecification tests for normality as in Montes-Rojas (2011). This can lead to a useful diagnostic strategy, but it does not provide new, more resistant, estimators and tests for the parameters of the model.

More preventive strategies aiming at relaxing the distributional assumptions have been proposed. They include more flexible parametric methods, such as a sample selection model based on the t distribution by Marchenko and Genton (2012), an extension for the skew-normal distribution by Ogundimu and Hutton (2012), a copula-based approach by Smith (2003), and the `gamlss` approach by Rigby and Stasinopoulos (2005). Although these models are more flexible than the standard normal model in that they contain additional parameters that can be used to accommodate skewness, kurtosis and possible outliers, they do not generate full neighborhoods (in a topological sense) of the central model and do not offer a full protection against possible distributional deviations from the latter; see Hampel et al. (1986, p. 10). Moreover, the introduction of additional parameters to fit a few possible outliers might be justifiable only when these observations contain important information about the problem under investigation.

Finally, several authors have proposed semiparametric (Ahn and Powell 1993, Newey 2009, Marra and Radice 2013), and nonparametric (Das et al. 2003) methods. Gallant and Nychka (1987) proposed a semi-nonparametric estimator based on Hermite series. Genton et al. (2012) and Ma et al. (2005, 2013) investigated semiparametric methods in the case of skew-symmetric distributions when no covariates are available.

In this paper we take a middle way between the classical strict parametric model and the fully nonparametric setup. We still assume the classical normal model as the central model, but we believe that it is only *approximate* in the sense that the true distribution of the error terms lies in a (small) neighborhood of the latter. We then derive estimators and tests which are still reliable in the full neighborhood. This has the advantage to provide insurance and protection against small but harmful distributional deviations and to still benefit from the parametric

structure, e.g., its computational simplicity and interpretability. From a data analytic point of view, our robust procedure fits the majority of the data and identifies outliers and possible substructures for further special treatment.

Of course, in situations when we are completely uncertain about the underlying distribution, the use of nonparametric methods would be in principle preferable; see the discussion in Hampel et al. (1986, p. 7). Notice however that nonparametric methods are not necessarily designed to be robust in the sense mentioned above. Even the arithmetic mean which is the nonparametric estimator of the expectation (if it exists) of any underlying distribution, is very sensitive to outliers and is not robust. For a detailed discussion see Huber (1981, p. 6).

A robustification of the MLE for this model could be carried out by applying standard tools of robust statistics (Salazar 2008). However, the resulting robust estimator would be computationally complex and this would be a clear disadvantage in applications. Indeed simulation techniques would be required to compute the Fisher consistency correction for the robust (truncated) score function. Therefore, we focus here on the robustness analysis of Heckman's two-stage procedure for the model specified by (1)-(4). It is structurally simpler, has a straightforward interpretation, and leads to a robust estimator, which is computationally simple. Moreover, there are numerous extensions of the classical Heckman's model, including switching regressions, simultaneous equations with selectivity, and models with self-selectivity, to mention a few, where the construction of the joint likelihood becomes cumbersome, whereas Heckman's estimator can be easily computed. Our robustified version can in principle be extended to these situations.

In the past decades, robust estimators and tests have been developed for large classes of models both in the statistical and econometric literature; see for instance, the books by Huber (1981, 2nd edition by Huber and Ronchetti 2009), Hampel et al. (1986), Maronna et al. (2006) in the statistical literature and Peracchi (1990, 1991) and Ronchetti and Trojani (2001) in the econometric literature. In particular, the quantile regression approach (Koenker 2005)

has proved fruitful as a specific way to robustify classical procedures and in the framework of sample selection models, it has been proposed by Buchinsky (1998) and Huber and Melly (2015). Details about this approach are provided in Section 4. However, except for a few specific contributions mentioned above, the robustness aspects in sample selection models have not received a lot of attention and no general theory is available.

In this paper we fill this gap by providing the following contributions to the literature. In Section 2 we investigate the robustness properties of Heckman’s two-stage estimator and its estimated variance by deriving the influence function and the change-of-variance function; see Hampel et al. (1986). These functions are used to quantify the bias of the estimator and its estimated variance respectively, when deviations from the assumed bivariate normal model are present and the true data generating process lies in a neighborhood of the assumed bivariate normal distribution. It turns out that both functions are unbounded and this implies that a small deviation can have a large impact on the bias of the estimator and on its variance. The latter in turn has a large effect on the corresponding confidence intervals. Moreover, by means of these functions, we provide a von Mises expansion of the test statistic for the test on sample selection bias (SSB) which demonstrates its non-robustness.

Since the classical estimation and testing procedures are not robust with respect to deviations from the assumed underlying stochastic model, we propose in Section 3 new robust estimators and a new robust test for SSB, which are the natural robust counterparts of the classical Heckman’s two-stage estimator and test of SSB. They are available in the R package `ssmrob` on CRAN. We study the performance of our estimators both in the presence and absence of exclusion restrictions. Section 4 reports the finite sample performance of the new estimators and test in a simulation setting with several types of contamination, degrees of selection bias and severity of censoring. A comparison with the quantile regression approach is included. Moreover, we compare a classical and robust analysis on a real data set (ambulatory expenditures data). The technical derivations and assumptions are provided in the Appendix.

The Supplementary Material contains the derivations of the IF and the CVF of Heckman's estimator, additional simulation results, more details about the quantile regression approach, and the use of our results for two important extensions of the basic model, namely switching regression models and simultaneous equations models with selectivity.

2 Robustness Issues with Heckman's Two-Stage Estimator

In this section we present our main results concerning the two-stage estimator in the general framework defined by (6) and (7). In particular, we derive its influence function (IF) and its change-of-variance function (CVF) and discuss the robustness properties of the Heckman's estimator. Moreover, we explain the connection between its IF and the asymptotic variance. Finally, we explore the robustness properties of the SSB test. This provides a theoretical framework for the analysis of the robustness properties of Heckman's estimator and SSB test.

We consider a parametric sample selection model $\{F_\theta\}$, where $\theta = (\beta_1, \beta_2, \sigma_2, \rho)$ lies in Θ , a compact subset of $\mathbb{R}^{p_1+p_2+2}$. Let F_N be the empirical distribution function putting mass $1/N$ at each observation $z_i = (z_{1i}, z_{2i})$, where $z_{ji} = (x_{ji}, y_{ji})$, $j = 1, 2$, $i = 1, \dots, N$, and let F be the distribution function of z_i . The Heckman's estimator can be represented as a two-stage M-estimator, with probit MLE in the first stage and OLS in the second stage. Define two statistical functionals S and T corresponding to the estimators of the first and second stage, respectively. The domain of S is a class of probability distributions on $\mathbb{R}^{p_1}$ and its range is a vector in $\mathbb{R}^{p_1}$. The domain of T is a class of probability distributions on $\mathbb{R}^{p_1+p_2+2}$ and its range is a vector in $\mathbb{R}^{p_2+2}$.

The two-stage estimator can be expressed as a solution of the empirical counterpart of the system:

$$\int \Psi_1\{(x_1, y_1); S(F)\}dF = 0, \quad (6)$$

$$\int \Psi_2[(x_2, y_2); \lambda\{(x_1, y_1); S(F)\}, T(F)]dF = 0, \quad (7)$$

where $\Psi_1(\cdot; \cdot)$ and $\Psi_2(\cdot; \cdot, \cdot)$ are the score functions of the first and second stage estimators, respectively. In the classical case $\Psi_1(\cdot; \cdot)$ and $\Psi_2(\cdot; \cdot, \cdot)$ are given by

$$\Psi_1\{(x_1, y_1); S(F)\} = \{y_1 - \Phi(x_1^T \beta_1)\} \frac{\phi(x_1^T \beta_1)}{\Phi(x_1^T \beta_1)\{1 - \Phi(x_1^T \beta_1)\}} x_1, \quad (8)$$

$$\Psi_2[(x_2, y_2); \lambda\{(x_1, y_1); S(F)\}, T(F)] = (y_2 - x_2^T \beta_2 - \lambda \beta_\lambda) \begin{pmatrix} x_2 \\ \lambda \end{pmatrix} y_1. \quad (9)$$

Here $\lambda\{(x_1, y_1); S(F)\}$ denotes the dependence of λ on $S(F) = \beta_1$, while $T(F)$ depends directly on F and indirectly on F through $S(F)$.

2.1 Influence Function

For a given functional $T(F)$, the influence function (IF) is defined by Hampel (1974) as $IF(z; T, F) = \lim_{\epsilon \rightarrow 0} [T\{(1 - \epsilon)F + \epsilon \Delta_z\} - T(F)]/\epsilon$, where Δ_z is the probability measure which puts mass 1 at the point z . In our case $(1 - \epsilon)F + \epsilon \Delta_z$ is a contamination of the joint distribution of z_i , but marginal contaminations on the components of z_i can also be considered; see the comments below. The IF describes the standardized asymptotic bias on the estimator due to a small amount of contamination ϵ at the point z . Moreover, using a von Mises (1947) expansion, the maximum bias over the neighborhood described by the perturbations $F_\epsilon = (1 - \epsilon)F + \epsilon G$, where G is some arbitrary distribution function, is approximately

$$\sup_G \|T(F_\epsilon) - T(F)\| \cong \epsilon \sup_z \|IF(z; T, F)\|.$$

Therefore, a condition for (local) robustness is a bounded IF with respect to z , which means that if the $IF(\cdot; \cdot, \cdot)$ is unbounded then the bias of the estimator can become arbitrarily large. The next proposition gives the influence function of Heckman's two-stage estimator.

Proposition 1. *For the model (1)-(4), the IF of the Heckman's two-stage estimator is*

$$IF(z; T, F) = \left\{ \int \begin{pmatrix} x_2 x_2^T & \lambda x_2 \\ \lambda x_2^T & \lambda^2 \end{pmatrix} y_1 dF \right\}^{-1} \left\{ (y_2 - x_2^T \beta_2 - \lambda \beta_\lambda) \begin{pmatrix} x_2 \\ \lambda \end{pmatrix} y_1 \right. \\ \left. + \int \begin{pmatrix} x_2 \beta_\lambda \\ \lambda \beta_\lambda \end{pmatrix} y_1 \lambda' dF \cdot IF(z; S, F) \right\}, \quad (10)$$

where

$$IF(z; S, F) = \left(\int \left[\frac{\phi(x_1^T \beta_1)^2 x_1 x_1^T}{\Phi(x_1^T \beta_1) \{1 - \Phi(x_1^T \beta_1)\}} \right] dF \right)^{-1} \{y_1 - \Phi(x_1^T \beta_1)\} \frac{\phi(x_1^T \beta_1) x_1}{\Phi(x_1^T \beta_1) \{1 - \Phi(x_1^T \beta_1)\}}. \quad (11)$$

The proof is given in the Supplementary Material.

The first term of (10) is the score function of the second stage and it corresponds to the IF of a standard OLS regression. The second term contains the IF of the first stage estimator. Clearly, the first term is unbounded with respect to y_2 , x_2 and λ . Notice that the function λ is unbounded from the left, and tends to zero from the right. From (11) we can see that the second term is also unbounded, which means that there is a second source of unboundedness arising from the selection stage. Therefore, the estimator fails to be locally robust. A small amount of contamination is enough for the estimator to become arbitrarily biased. In Section 3 we present a way to construct a two-stage estimator with a bounded influence function.

2.2 Asymptotic Variance and Change-of-Variance Function

The expression of the asymptotic variance for the two-stage estimator has been derived by Heckman (1979), and later corrected by Greene (1981). Duncan (1987) suggested another approach to derive the asymptotic variance using the M-estimation framework. Using the result in Hampel et al. (1986), the general expression of the asymptotic variance is given by

$$V(T, F) = \int IF(z; T, F) \cdot IF(z; T, F)^T dF(z).$$

Specifically, denote the components of the IF as follows:

$$\begin{aligned} a(z) &= (y_2 - x_2^T \beta_2 - \lambda \beta_\lambda) \begin{pmatrix} x_2 \\ \lambda \end{pmatrix} y_1, \\ b(z) &= \left\{ \int \begin{pmatrix} x_2 \beta_\lambda \\ \lambda \beta_\lambda \end{pmatrix} y_1 \lambda' dF \right\} \cdot IF(z; S, F), \\ M(\Psi_2) &= \int \begin{pmatrix} x_2 x_2^T & \lambda x_2 \\ \lambda x_2^T & \lambda^2 \end{pmatrix} y_1 dF. \end{aligned} \quad (12)$$

Then the expression of the asymptotic variance of Heckman's two-stage estimator is

$$V(T, F) = M(\Psi_2)^{-1} \int \{a(z)a(z)^T + a(z)b(z)^T + b(z)a(z)^T + b(z)b(z)^T\} dF(z) M(\Psi_2)^{-1}.$$

After integration and some simplifications we obtain the asymptotic variance of the classical estimator

$$V \left\{ \begin{pmatrix} \beta_2 \\ \beta_\lambda \end{pmatrix}, F \right\} = (X^T X)^{-1} \left[\sigma_2^2 \left\{ X^T \left(I - \frac{\beta_\lambda^2}{\sigma_2^2} \Delta \right) X \right\} + \beta_\lambda^2 X^T \Delta X_1 \text{Var}(S, F) X_1^T \Delta X \right] (X^T X)^{-1},$$

where Δ is a diagonal matrix with elements $\delta_{ii} = \frac{\partial \lambda(x_{1i}\beta_1)}{\partial (x_{1i}\beta_1)}$, the matrix X consists of vectors $x_i = \begin{pmatrix} x_{2i} \\ \lambda_i \end{pmatrix}$, and $\text{Var}(S, F)$ denotes the asymptotic variance of the probit MLE.

Robustness issues are not limited to the bias of the estimator, but concern also the stability of the asymptotic variance. Indeed, the latter is used to construct confidence intervals for the parameters and we want the influence of small deviations from the underlying distribution on their coverage probability and length to be bounded. Therefore, we investigate the behavior of the asymptotic variance of the estimator under a contaminated distribution F_ϵ and derive the CVF, which reflects the influence of a small amount of contamination on the asymptotic variance of the estimator. These results are used in Section 2.3 to investigate the robustness properties of the SSB test.

The CVF of an M-estimator T at a distribution F is defined by the matrix $CVF(z; T, F) = \left[(\partial/\partial\epsilon)V\{T, (1-\epsilon)F + \epsilon\Delta_z\} \right]_{\epsilon=0}$, for all z where this expression exists; see Hampel et al. (1981) and Genton and Rousseeuw (1995). Again a von Mises (1947) expansion of $\log V(T, F_\epsilon)$ at F gives

$$V(T, F_\epsilon) \cong V(T, F) \exp \left\{ \epsilon \int \frac{CVF(z; T, F)}{V(T, F)} dG \right\}. \quad (13)$$

If the $CVF(z; T, F)$ is unbounded then the variance can behave unpredictably (arbitrarily large or small); see Hampel et al. (1986, p. 175). Using (13) one can obtain the numerical approximation of the variance of the estimator T at a given underlying distribution $(1-\epsilon)F + \epsilon G$ for a given G .

Proposition 2. *The CVF of the Heckman's two-stage estimator is given by*

$$\begin{aligned}
CVF(z; S, T, F) = & V - M(\Psi_2)^{-1} \left\{ \int D_H dF + \begin{pmatrix} x_2 x_2^T & \lambda x_2 \\ \lambda x_2^T & \lambda^2 \end{pmatrix} y_1 \right\} V \\
& + M(\Psi_2)^{-1} \int \{A_H a^T + A_H b^T + B_H b^T\} dF M(\Psi_2)^{-1} \\
& + M(\Psi_2)^{-1} \int \{a A_H^T + b A_H^T + b B_H^T\} dF M(\Psi_2)^{-1} \\
& + M(\Psi_2)^{-1} \{a(z) a(z)^T + a(z) b(z)^T + b(z) a(z)^T + b(z) b(z)^T\} M(\Psi_2)^{-1} \\
& - V \left\{ \int D_H dF + \begin{pmatrix} x_2 x_2^T & \lambda x_2 \\ \lambda x_2^T & \lambda^2 \end{pmatrix} y_1 \right\} M(\Psi_2)^{-1}, \tag{14}
\end{aligned}$$

where V denotes the asymptotic variance of the Heckman (1979) two-stage estimator, $a(z)$, $b(z)$, and $M(\Psi_2)$ are defined by (12). Explicit expressions for these terms are given in the *Supplementary Material*.

The CVF has several sources of unboundedness. The second term of (14) contains the derivative of the score function $\Psi_2(\cdot; \cdot, \cdot)$ with respect to the parameter which is unbounded. The same holds for the last term. Finally, in the fifth term there are two components depending on the score functions of two estimators which are unbounded. Clearly, the CVF is unbounded, which means that the variance can become arbitrarily large. Taking into account that the two-stage estimator by definition is not efficient, we can observe a combined effect of inefficiency with non-robustness of the variance estimator. These problems can lead to misleading p -values and incorrect confidence intervals. Second-order effects in the von Mises expansion are discussed in general in La Vecchia et al. (2012).

2.3 Sample Selection Bias Test

Heckman (1979) proposed to test for the selection bias using the standard t -test of the coefficient β_λ . Melino (1982) showed that this test is equivalent to a Lagrange multiplier test and has desirable asymptotic properties. Several other proposals are available in the literature, see e.g. Vella (1992), but the simple Heckman's test is the most widely used in applications. Here we investigate the effect of contamination on the test statistic $\tau_n = \sqrt{n} \hat{\beta}_\lambda / \sqrt{V(\beta_\lambda, F)}$.

Using the expressions for the IF and CVF of the estimator and its asymptotic variance, we obtain the von Mises expansion of the test statistic:

$$\frac{T(F_\epsilon)}{\sqrt{V(F_\epsilon)/n}} = \frac{T(F)}{\sqrt{V(F)/n}} + \epsilon \left[\frac{IF(z; T, F)}{\sqrt{V(F)/n}} + \frac{1}{2n} T(F) \frac{CVF(z, T, F)}{\{V(F)/n\}^{5/2}} \right] + o(\epsilon),$$

which provides an approximation of the bias of the test statistic under contamination. It is clear that the IF of the test depends on the IF and CVF of the estimator. Hence, the IF of the test statistic is also unbounded. Since, according to Hampel et al. (1986, p. 199), the IFs of the level and of the power of the test are proportional to the IF of the test statistic, the test is not robust. Moreover, because Heckman's two-stage estimator suffers from a lack of efficiency, small deviations from the model can enhance this effect and increase the probability of type I and type II errors of the SSB test.

Notice however that the term containing the CVF is of higher order, which means that the influence of the contamination on the test statistic is mostly explained by the IF of the corresponding estimator. Hence, for practical purposes we need to have at least a robust estimator with a bounded IF with an additional bonus if the CVF is bounded as well.

3 Robust Estimation and Inference

In this section we suggest how to robustify the two-stage estimator and propose a simple robust alternative to the SSB test.

3.1 Robust Two-Stage Estimator

From the expression of the IF in (10), it is natural to construct a robust two-stage estimator by robustifying the estimators in both stages. The idea is to obtain an estimator with bounded bias in the first stage, then compute λ , which will transfer potential leverage effects from the first stage to the second, and use the robust estimator in the second stage, which will correct for the remaining outliers.

Consider the two-stage M-estimation framework given by (6) and (7). We can obtain a

robust estimator by bounding both score functions. In the first stage, we construct a robust probit estimator. We use a general class of M-estimators of Mallows' type, where the influence of deviations on y_1 and x_1 are bounded separately; see Cantoni and Ronchetti (2001). The estimator is defined by the following score function:

$$\Psi_1^R\{z_1; S(F)\} = \nu(z_1; \mu)\omega_1(x_1)\mu' - \alpha(\beta_1), \quad (15)$$

where $\alpha(\beta_1) = \frac{1}{n} \sum_{i=1}^n E\{\nu(z_{1i}; \mu_i)\omega_1(x_{1i})\mu'_i\}$ is a term to ensure the unbiasedness of the estimating function with the expectation taken with respect to the conditional distribution of $y|x$, $\nu(\cdot|\cdot)$, $\omega_1(x_1)$ are weight functions defined below, $\mu_i = \mu_i(z_{1i}, \beta_1) = \Phi(x_{1i}^T \beta_1)$, and $\mu'_i = \frac{\partial}{\partial \beta_1} \mu_i$.

The weight functions are defined by

$$\nu(z_{1i}; \mu_i) = \psi_{c_1}(r_i) \frac{1}{V^{1/2}(\mu_i)},$$

where $r_i = \frac{y_{1i} - \mu_i}{V^{1/2}(\mu_i)}$ are Pearson residuals and ψ_{c_1} is the Huber function defined by

$$\psi_{c_1}(r) = \begin{cases} r, & |r| \leq c_1, \\ c_1 \text{sign}(r), & |r| > c_1. \end{cases} \quad (16)$$

The tuning constant c_1 is chosen to ensure a given level of asymptotic efficiency at the normal model. A typical value is 1.345, as advocated by Cantoni and Ronchetti (2001) in the GLM setting. A simple choice of the weight function $\omega_1(\cdot)$ is $\omega_{1i} = \sqrt{1 - H_{ii}}$, where H_{ii} is the i th diagonal element of the hat matrix $H = X(X^T X)^{-1} X^T$. More sophisticated choices for ω_1 are available, e.g. the inverse of the robust Mahalanobis distance based on high breakdown robust estimators of location and scatter of the x_{1i} . For the probit case we have that $\mu_i = \Phi(x_{1i}^T \beta_1)$, $V(\mu_i) = \Phi(x_{1i}^T \beta_1)\{1 - \Phi(x_{1i}^T \beta_1)\}$ and hence the quasi-likelihood estimating equations are

$$\sum_{i=1}^n \left\{ \psi_{c_1}(r_i) \omega_1(x_{1i}) \frac{1}{[\Phi(x_{1i}^T \beta_1)\{1 - \Phi(x_{1i}^T \beta_1)\}]^{1/2}} \phi(x_{1i}^T \beta_1) x_{1i} - \alpha(\beta_1) \right\} = 0,$$

and $E\{\psi_{c_1}(r_i)\}$ in the $\alpha(\beta_1)$ term is equal to

$$E \left[\psi_{c_1} \left\{ \frac{y_{1i} - \mu_i}{V^{1/2}(\mu_i)} \right\} \right] = \psi_{c_1} \left\{ \frac{-\mu_i}{V^{1/2}(\mu_i)} \right\} \{1 - \Phi(x_{1i}^T \beta_1)\} + \psi_{c_1} \left\{ \frac{1 - \mu_i}{V^{1/2}(\mu_i)} \right\} \Phi(x_{1i}^T \beta_1).$$

This estimator has a bounded IF and ensures robustness of the first estimation stage.

To obtain a robust estimator for the equation of interest (second stage), we propose to use an M-estimator of Mallows-type with the following Ψ -function:

$$\Psi_2^R(z_2; \lambda, T) = \Psi_{c_2}(y_2 - x_2^T \beta_2 - \lambda \beta_\lambda) \omega(x_2, \lambda) y_1, \quad (17)$$

where $\Psi_{c_2}(\cdot)$ is the classical Huber function defined by (16), but with possibly a different tuning constant c_2 , $\omega(\cdot)$ is a weight function on the x 's, which can also be based on the robust Mahalanobis distance $d(x_2, \lambda)$, e.g.

$$\omega(x_2, \lambda) = \begin{cases} x_2, & \text{if } d(x_2, \lambda) < c_m, \\ \frac{x_2 c_m}{d(x_2, \lambda)}, & \text{if } d(x_2, \lambda) \geq c_m, \end{cases} \quad (18)$$

where c_m is chosen according to the level of tolerance, given that the squared Mahalanobis distance follows a χ^2 -distribution. The choices of c_2 , $\omega(\cdot)$, and c_m come from the results in the theory of robust linear regression; see Hampel et al. (1986). In our numerical applications, we use $c_2 = 1.345$ and c_m corresponding to 5% critical level.

The robust estimator derived above assumes implicitly the presence of exclusion restrictions, i.e. $x_1 \neq x_2$, but often in practice the sets of explanatory variables are the same for both selection and outcome equations, i.e. $x_1 = x_2$. This issue can lead to multicollinearity because of quasi-linearity of the inverse Mills ratio in a substantial range of its support. In practice it is recommended that there should be a predictor which explains y_1 and is not significant for y_2 , although it might not be easy to find such a variable. Notice that, the lack of exclusion restriction is not a peculiar problem of the two-stage estimator, but the MLE can also suffer from it (Leung and Yu 2000). This topic has generated a lot of research and discussion; see Nelson (1984), Leung and Yu (1996) among others, and a review by Leung and Yu (2000) for a general discussion. From a robustness perspective, we would like our estimator to be still reliable also when the exclusion restriction is not available. Therefore, we now propose a slight modification of the robust estimator developed above to cover this situation.

In the presence of a high degree of correlation between the explanatory variables, the Mahalanobis distance can become inflated. This leads to an increase in the number of zero weights

in (18) and, hence, to an additional loss of efficiency. Given that the source of the multicollinearity is known, i.e. λ can be (approximately) expressed as a linear combination of x_2 's, a simple solution is to split the design space (x_2, λ) while computing the robustness weights $\omega(x_2, \lambda)$. We split the (approximately) linearly dependent components $(x_2^{(1)}, \dots, x_2^{(p_2)}, \lambda)$ into two independent components $(x_2^{(1)}, \dots, x_2^{(q)})$ and $(x_2^{(q+1)}, \dots, x_2^{(p_2)}, \lambda)$ and compute the robustness weights $\omega(x_2^{(1)}, \dots, x_2^{(q)})$ and $\omega(x_2^{(q+1)}, \dots, x_2^{(p_2)}, \lambda)$. Then, we combine these weights as $\omega(x_2, \lambda) = \omega(x_2^{(1)}, \dots, x_2^{(q)}) \omega(x_2^{(q+1)}, \dots, x_2^{(p_2)}, \lambda)$, which guarantee robustness in this case. As a general rule, we suggest to group λ with variable(s) having the smallest correlation with it. The question of an optimal split of the design space for the entire class of the Mallows-type estimators is beyond the scope of this paper and is left for future research.

We summarize the properties of the proposed estimator in the following proposition.

Proposition 3. *Under the assumptions stated in the Appendix, Heckman's two-stage estimator defined by (15) and (17) is robust, consistent, and asymptotically normal, with asymptotic variance given by:*

$$V(T, F) = M (\Psi_2^R)^{-1} \int \{a_R(z)a_R(z)^T + b_R(z)b_R(z)^T\} dF M (\Psi_2^R)^{-1}, \quad (19)$$

where $M (\Psi_2^R) = - \int \frac{\partial}{\partial \beta_2} \Psi_2^R(z; \lambda, T) dF$, $a_R(z) = \Psi_2^R(z; \lambda, T)$, and

$$b_R(z) = \int \frac{\partial}{\partial \lambda} \Psi_2^R(z; \lambda, T) \frac{\partial}{\partial \beta_1} \lambda dF \left\{ \int \frac{\partial}{\partial \beta_1} \Psi_1^R(z; S) dF \right\}^{-1} \Psi_1^R(z; S).$$

The asymptotic variance of the robust estimator has the same structure as that of the classical Heckman's estimator. Its computation can become complicated, depending on the choice of the score function, but for simple cases, e.g. the Huber function, it is relatively simple. The estimator can be obtained numerically using standard techniques, e.g. Newton-Raphson procedure. New R (R Core Team 2012) functions for robust estimation and inference in sample selection models are provided in a package **ssmrob**.

Remark In the likelihood framework, the Huber function defines the most efficient estimator, subject to a bounded influence function. Therefore, in addition to its computational

simplicity, it seems natural to use this function in our case. Of course, in principle other bounded score functions could be used, such as that defining the maximum likelihood estimator under a t_ν distribution; see in a more restricted setting Marchenko and Genton (2012).

3.2 Robust Sample Selection Bias Test

To test SSB, i.e. $H_0 : \beta_\lambda = 0$ vs $H_A : \beta_\lambda \neq 0$, we simply propose to use a t -test based on the robust estimator of β_λ and the corresponding estimator of its standard error derived in Section 3.1, where the latter is obtained by estimating (19).

The first term of (19), $M(\Psi_2^R)^{-1} \int a_R(z) a_R(z)^T dF M(\Psi_2^R)^{-1}$, is similar to the asymptotic variance of standard linear regression, but with heteroscedasticity. Therefore, we use the Eicker (1967) - Huber (1967) - White (1980) heteroscedasticity-consistent variance estimator, i.e. we estimate this first term by $\hat{M}(\Psi_2^R)^{-1} \frac{1}{n} \sum \hat{a}(z_i) \hat{a}(z_i)^T \hat{M}(\Psi_2^R)^{-1}$, where $\hat{M}(\Psi_2^R)$ and $\hat{a}_R(z)$ are the sample versions of M and $a_R(z)$, respectively.

The second term of the asymptotic variance, $M(\Psi_2^R)^{-1} \int b_R(z) b_R(z)^T dF M(\Psi_2^R)^{-1}$, is the asymptotic variance of the probit MLE pre- and post-multiplied by the constant matrix, which depends on the form of the score function of the second stage. Thus, a consistent estimator is

$$\hat{M}(\Psi_2^R)^{-1} \frac{1}{n} \sum \hat{b}_R(z_i) \hat{b}_R(z_i)^T \hat{M}(\Psi_2^R)^{-1} = \hat{M}(\Psi_2^R)^{-1} \frac{1}{n} \sum \frac{\partial \Psi_{2i}^R}{\partial \beta_1} \hat{\text{Var}}(S, F) \left(\frac{1}{n} \sum \frac{\partial \Psi_{2i}^R}{\partial \beta_1} \right)^T \hat{M}(\Psi_2^R)^{-1},$$

where $\frac{\partial \Psi_{2i}^R}{\partial \beta_1} = \frac{\partial \Psi_2(z_{2i}; \lambda, T(F))}{\partial \lambda} \frac{\partial \lambda(z_{1i}; S(F))}{\partial \beta_1}$.

4 Numerical Examples

4.1 Simulation Study

We carry out a Monte Carlo study to illustrate the robustness issues in the model described in Section 1.1 and compare different estimators. In our experiment we generate $y_{1i}^* = x_{11i} + x_{12i} + 0.75x_{13i} + e_{1i}$, where $x_{11i} \sim N(0, 1)$, $x_{12i} \sim N(-1, 0.5)$, and $x_{13i} \sim N(1, 1)$. For the equation of interest when the exclusion restriction is not available, we use the same explanatory variables $x_2 = x_1$. When it is available, the variable x_{23i} is generated independently from x_{13i} , and

follows the same distribution. The errors e_1 and e_2 are from a bivariate normal distribution with expectation 0, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.7$, which gives $\beta_\lambda = 0.7$. The degree of censoring is controlled by the intercept in the selection equation, denoted by β_{10} and set to 0, which corresponds to approximately 45% of censoring. Results (not shown here) are similar for other censoring proportions such as 75% and 25%. In the equation of interest the intercept is $\beta_{20} = 0$ and the slope coefficients are $\beta_2 = (1.5, 1, 0.5)^T$. We find the estimates of β_1 and β_2 without contamination and with two types of contamination. In the first scenario we contaminate x_1 when the corresponding $y_1 = 0$. We generate observations from the model described above and replace them with probability $\epsilon = 0.01$ by a point mass at $(x_{11}, x_{12}, x_{13}, y_1, y_2) = (2, 0, 3, 0, 1)$. In this case we study the effect of leverage outliers when they are not transferred to the main equation. In the second scenario we contaminate x_1 when the corresponding $y_1 = 1$. We use the same type of contamination as in the first scenario, but the point mass is at $(-2, -2, -1, 1, 0)$. Notice that the contaminating point deviates by two standard deviations from the centers of distributions of the explanatory variables, which is very hard to identify using standard exploratory analysis. In the Supplementary Material we report additional simulations when the distribution of the error terms deviates from the normal. The sample size is $N = 1000$ and we repeat the experiment 500 times. For other sample sizes, e.g. $N = 2000$ or $N = 500$, the behavior of the estimators is the same. A second simulation design (very simple and mostly pedagogical) with one explanatory variable is presented in the Supplementary Material. In addition, we study there a type of contamination when the outliers emerge only in the selection stage.

We compare Heckman's estimator with two robust versions derived in Section 3, i.e. the robust probit with OLS and the robust two-stage (2S) estimator. Moreover, when an exclusion restriction is available, we add the quantile regression estimator (QRE). This is the estimator proposed by Buchinsky (1998) and extended by Huber and Melly (2015), which is a combination of a semiparametric binary regression as in Klein and Spady (1993) in the first stage and

quantile regression in the second stage. It is computed using the code kindly provided to us by M. Huber. More details and a discussion of the robustness properties of this estimator can be found in the Supplementary Material.

In Table 1 we first consider only the first stage. We notice that the three estimators perform well at the model (with very small efficiency losses for our robust proposal and for semiparametric binary regression with respect to the classical one). However, under contamination only the robust proposal remains nearly unbiased. In Tables 2-3 and Figures 1 to 3 we consider classical and robust two-stage estimators. The QRE is included in the table only when an exclusion restriction is available.

Again all estimators perform well without contamination. As expected, under contamination Heckman's estimator breaks down. This effect can be seen in Figure 2 (with exclusion restriction) and Figure 3 (without exclusion restriction). When the exclusion restriction is not available the magnitude of the bias of the classical estimator is considerably higher than that when the exclusion restriction is available. The estimation of the slope coefficients by QRE are robust. However, the estimators of the intercept and β_λ become severely biased. While it is true that often one is mostly interested only in the slopes, the non-robustness with respect to β_λ affects the subsequent test for selectivity. Finally, notice that the QRE of the slopes have larger MSE than those of the robust two-stage.

In the case when the outlier is not transferred to the equation of interest (Figures 2-3 top panels) it is enough to use a robust probit, but when the outlier emerges in the equation of interest (Figures 2-3 bottom panels), a robust estimation of the second stage is necessary. In this case the outliers influence not only both estimation stages directly, but the effect of contamination is amplified by the influence through λ . The behavior of the variances of the robust estimators remains stable, while the variance of the classical estimator is seriously affected by the contamination.

In Figure 5 we study the efficiency of the estimators. We present the plots of the relative

efficiency of the robust two-stage estimator versus the classical estimator, depending on the amount of contamination ϵ , which varies from zero to 3%. We show the figures for the case when the contaminated observations emerge at both stages ($y_{1i} = 1$, the case when $y_{1i} = 0$ is presented in the Supplementary Material). As it is expected from the theory, the robust estimator is less efficient than the classical one, when the distributional assumptions hold exactly. However, when a small amount of contamination is introduced, the situation changes completely. For instance, when the exclusion restriction is not available and the contaminated observations emerge in the second stage, the classical estimator of β_λ becomes less efficient than the robust one with only 0.1% contamination (top panel of Figure 5). The efficiency loss of the classical estimator concerns not only the IMR parameter, but also the other explanatory variables. Note that the behavior of the variance of the robust estimator remains stable under contamination (Table 1, Table 2). Finally, in Figure 4 we plot the Receiver Operating Characteristic (ROC) curves. The data generating process (DGP) is as discussed above except for $\rho = -0.7$, and the contamination is very mild ($\epsilon = 0.001$). We study the case when $y_1 = 1$ and put the contaminating point mass at $(-1.5, -1.75, -0.5, 1, 0)$. Without contamination the curves are close, however when the data are slightly contaminated the classical test loses its power.

4.2 Ambulatory Expenditures Data

To further illustrate the behavior of our new robust methodology, we consider the data on ambulatory expenditures from the 2001 Medical Expenditure Panel Survey analyzed by Cameron and Trivedi (2009, p. 545). The data consist of 3,328 observations, where 526 (15.8%) correspond to zero expenditures. The distribution of the expenditures is skewed, so the log scale is used. The selection equation includes such explanatory variables as *age*, gender (*female*), education status (*educ*), ethnicity (*blhisp*), number of chronic diseases (*totchr*), and the insurance status (*ins*). The outcome equation holds the same variables. The exclusion restriction can be introduced by means of the income variable. We explore both cases, with and without

exclusion restriction (Sections 16.6.5 and 16.6.4 in Cameron and Trivedi 2009, respectively).

The results of the estimation obtained using the R package `sampleSelection` are reported in Table 4. Using the classical estimator, both with and without exclusion restriction, all the variables are significant for the decision to spend, and all except education status and insurance status are significant for the spending amount. The p-values of the SSB t -test are 0.099 and 0.101 which are not significant at the 5% level. The possible conclusion of no selection bias seems to be doubtful. The estimation of this model using the joint MLE returns the p-value of the Wald test equal to 0.380 and 0.395 with and without exclusion restriction, respectively. Theoretically, when the SSB test is not significant, the conclusion of absence of the sample selectivity is made, and one should use the OLS estimator. However, if the presence of sample selectivity is hidden by the deviation from the assumed model, then the OLS can produce biased estimates. The last column of Table 4 reports the results of the estimation by OLS. The values of the parameters and the significance of the variables are different from those obtained using the sample selection model. Cameron and Trivedi (2009) note that the conclusions about the absence of sample selectivity obtained by the classical tests should be treated with caution because of lack of robustness.

Using the robust two-stage estimator, we obtained results similar to those obtained using the classical estimators but with an important difference. The output is reported in Table 4. For all the variables the differences (regarding estimates and standard errors) are not dramatic, except for the inverse Mills ratio (IMR) parameter. The robust estimator returns $\hat{\beta}_{IMR} = -0.677$, compared to the Heckman's $\hat{\beta}_{IMR} = -0.480$. The robust SSB test is highly significant with a p-value $p = 0.009$. If the exclusion restriction is used, then the results are similar both for the estimators and for the tests. Using the robust estimator we obtain $\hat{\beta}_{IMR} = -0.690$ with a p-value $p = 0.006$. The robust analysis indicates that sample selection is present and that the two parts of the model are not independent. The lack of agreement between the classical and robust analysis is an important diagnostic information for the analyst. Our simulations

have shown that even small contaminations can bias the parameters and change the inference, which is likely to be the reason in this example.

5 Discussion

We introduced a framework for robust estimation and testing for sample selection models. These methods allow to deal with data deviating from the assumed model and to carry out reliable inference even in the presence of small deviations from the assumed normality model. Monte Carlo simulations demonstrated the good performance of the robust estimators under the model and with different types of contamination. Although at the inference stage one is concerned only with potential small deviations from the assumed model, at an early exploratory stage one could look for possible large deviations. Robustness against this type of deviations would require the development of high-breakdown estimators, which could possibly be obtained by replacing the Huber function in (16) by a redescending score function. However, this is the subject of future research. Although we focused on the basic sample selection model, our methodology can be easily extended to more general frameworks, including for instance Copas and Li (1997). Moreover, our techniques can be adapted to models beyond simple regression. The Switching Regression Model and the Simultaneous Equations Model with Selectivity are briefly discussed in the Supplementary Material.

A Assumptions and Proof of Proposition 3

Denote $\Psi^R(z; \theta) = \{\Psi_1^R(z; \beta_1)^T, \Psi_2^R(z; \beta_1, \beta_2)^T\}$. Assume the following conditions, which have been adapted from Duncan (1987):

- (i) $z_1, \dots, z_N$ is a sequence of independent identically distributed random vectors with distribution F defined on a space $\mathcal{Z}$;
- (ii) Θ is a compact subset of $\mathbb{R}^{p_1+p_2}$;
- (iii) $\int \Psi^R(z; \theta) dF = 0$ has a unique solution, θ_0 , in the interior of Θ ;

- (iv) $\Psi^R(z; \theta)$ and $\frac{\partial}{\partial \theta} \Psi^R(z; \theta)$ are measurable for each θ in Θ , continuous for each z in $\mathcal{Z}$, and there exist F -integrable functions ξ_1 and ξ_2 such that for all $\theta \in \Theta$ and $z \in \mathcal{Z}$

$$|\Psi^R(z; \theta) \Psi^R(z; \theta)^T| \leq \xi_1 \text{ and } |\frac{\partial}{\partial \theta} \Psi^R(z; \theta)| \leq \xi_2;$$
- (v) $\int \Psi^R(z; \theta) \Psi^R(z; \theta)^T dF$ is non-singular for each $\theta \in \Theta$;
- (vi) $\int \frac{\partial}{\partial \theta} \Psi^R(z; \theta_0) dF$ is finite and non-singular.

Proof of Proposition 3. Consistency and asymptotic normality follow directly from Theorems 1-4 in Duncan (1987). The asymptotic variance consists of two terms, because $a_R(z)b_R(z)^T$ and $b_R(z)a_R(z)^T$ vanish after integration, due to the independence of the error terms. $\square$

Acknowledgements The authors would like to thank the Editor, the Associate Editor, three Referees, Francis Vella, Stefan Sperlich, and Maria-Pia Victoria-Feser for very helpful comments which improved the original version of the manuscript. The work of Genton was supported by King Abdullah University of Science and Technology (KAUST). Ronchetti's research was partially supported by a Swiss National Science Foundation grant 100018–140295.

References

- Ahn, H. and Powell, J. L. (1993), "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.
- Amemiya, T. (1984), "Tobit Models: a Survey," *Journal of Econometrics*, 24, 3–61.
- 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.
- Bushway, S., Johnson, B. D., and Slocum, L. (2007), "Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology," *Journal of Quantitative Criminology*, 23, 151–178.
- Cameron, C. A. and Trivedi, P. K. (2009), *Microeconometrics Using Stata*, College Station, TX: Stata Press.
- Cantoni, E. and Ronchetti, E. (2001), "Robust Inference for Generalized Linear Models," *Journal of the American Statistical Association*, 96, 1022–1030.

- Collier, D. and Mahoney, J. (1996), “Insights and Pitfalls: Selection Bias in Qualitative Research,” *World Politics*, 49, 56–91.
- Copas, J. B. and Li, H. G. (1997), “Inference for Non-random Samples,” *Journal of the Royal Statistical Society Series B*, 59, 55–95.
- Das, M., Newey, W. K., and Vella, F. (2003), “Nonparametric Estimation of Sample Selection Models,” *Review of Economic Studies*, 70, 33–58.
- Duncan, G. M. (1987), “A Simplified Approach to M-Estimation with Application to Two-Stage Estimators,” *Journal of Econometrics*, 34, 373–389.
- Eicker, F. (1967), “Limit Theorems for Regression with Unequal and Dependent Errors,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L.M. LeCam, J. Neyman (Eds.), Berkeley: University of California Press, 59–82.
- Gallant, R. A. and Nychka, D. W. (1987), “Semi-Nonparametric Maximum Likelihood Estimation,” *Econometrica*, 55, 363–390.
- Genton, M. G., Kim, M., and Ma, Y. (2012), “Semiparametric Location Estimation under Non-random Sampling,” *Stat*, 1, 1–11.
- Genton, M. G. and Rousseeuw, P. J. (1995), “The Change-of-Variance Function of M-Estimators of Scale under General Contamination,” *Journal of Computational and Applied Mathematics*, 64, 69–80.
- Greene, W. H. (1981), “Sample Selection Bias as a Specification Error: Comment,” *Econometrica*, 49, 795–798.
- Hampel, F. (1974), “The Influence Curve and Its Role in Robust Estimation,” *Journal of the American Statistical Association*, 69, 383–393.
- Hampel, F., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley and Sons.
- Hampel, F., Rousseeuw, P. J., and Ronchetti, E. (1981), “The Change-of-Variance Curve and Optimal Redescending M-Estimators,” *Journal of the American Statistical Association*, 76, 643–648.
- Heckman, J. (1979), “Sample Selection Bias as a Specification Error,” *Econometrica*, 47, 153–161.
- Huber, M. and Melly, B. (2015), “A Test of the Conditional Independence Assumption in Sample Selection Models,” *Journal of Applied Econometrics*, To Appear.
- Huber, P. J. (1967), “The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L.M. LeCam, J. Neyman (Eds.), Berkeley: University of California Press, 221–233.
- (1981), *Robust Statistics*, New York: John Wiley and Sons.

- Huber, P. J. and Ronchetti, E. (2009), *Robust Statistics*, New York: John Wiley and Sons, 2nd ed.
- Klein, R. W. and Spady, R. H. (1993), “Efficient Semiparametric Estimator for Binary Response Models,” *Econometrica*, 61, 387–421.
- Koenker, R. (2005), *Quantile Regression*, New York: Cambridge University Press.
- La Vecchia, D., Ronchetti, E., and Trojani, F. (2012), “Higher-Order Infinitesimal Robustness,” *Journal of the American Statistical Association*, 107, 1546–1557.
- Lennox, C. S., Francis, J. R., and Wang, Z. (2012), “Selection Models in Accounting Research,” *The Accounting Review*, 87, 589–616.
- Leung, S. F. and Yu, S. (1996), “On the Choice Between Sample Selection and Two-part Models,” *Journal of Econometrics*, 72, 197–229.
- (2000), “Collinearity and Two-Step Estimation of Sample Selection Models: Problems, Origins, and Remedies,” *Computational Economics*, 15, 173–199.
- Ma, Y., Genton, M. G., and Tsiatis, A. A. (2005), “Locally Efficient Semiparametric Estimators for Generalized Skew-elliptical Distributions,” *Journal of the American Statistical Association*, 100, 980–989.
- Ma, Y., Kim, M., and Genton, M. G. (2013), “Semiparametric Efficient and Robust Estimation of an Unknown Symmetric Population under Arbitrary Sample Selection Bias,” *Journal of the American Statistical Association*, 108, 1090–1104.
- Marchenko, Y. V. and Genton, M. G. (2012), “A Heckman Selection- t Model,” *Journal of the American Statistical Association*, 107, 304–317.
- Maronna, R. A., Martin, G. R., and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, Chichester: John Wiley and Sons.
- Marra, G. and Radice, R. (2013), “Estimation of a Regression Spline Sample Selection Model,” *Computational Statistics and Data Analysis*, 61, 158–173.
- Melino, A. (1982), “Testing for Sample Selection Bias,” *Review of Economic Studies*, XLIX, 151–153.
- Montes-Rojas, G. V. (2011), “Robust Misspecification Tests for the Heckman’s Two-step Estimator,” *Econometric Reviews*, 30, 154–172.
- Nelson, F. D. (1984), “Efficiency of the Two-Step Estimator for Models with Endogenous Sample Selection,” *Journal of Econometrics*, 24, 181–196.
- Newey, W. K. (2009), “Two-Step Series Estimation of Sample Selection Models,” *Econometrica Journal*, 12, S217–S229.
- Ogundimu, E. O. and Hutton, J. L. (2012), “A Sample Selection Model with Skew-Normal Distribution,” *CRiSM report 12-05*.

- Paarsch, H. J. (1984), “A Monte Carlo Comparison of Estimators for Censored Regression Models,” *Journal of Econometrics*, 24, 197–213.
- Peracchi, F. (1990), “Bounded-Influence Estimators for the Tobit Model,” *Journal of Econometrics*, 44, 107–126.
- (1991), “Robust M-Tests,” *Econometric Theory*, 7, 69–84.
- Puhani, P. A. (2000), “The Heckman Correction for Sample Selection and Its Critique,” *Journal of Economic Surveys*, 14, 53–68.
- R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Rigby, R. and Stasinopoulos, D. (2005), “Generalized Additive Models for Location, Scale and Shape,” *Journal of the Royal Statistical Society Series C*, 54, 507–554.
- Ronchetti, E. and Trojani, F. (2001), “Robust Inference with GMM Estimators,” *Journal of Econometrics*, 101, 37–69.
- Salazar, L. (2008), “A Robustness Study of Heckman’s Model,” *Master’s Thesis, University of Geneva, Switzerland*.
- Smith, M. D. (2003), “Modelling Sample Selection Using Archimedean Copulas,” *Econometrics Journal*, 6, 99–123.
- Stolzenberg, R. M. and Relles, D. A. (1997), “Tools for Intuition about Sample Selection Bias and Its Correction,” *American Sociological Review*, 62, 494–507.
- Toomet, O. and Henningsen, A. (2008), “Sample Selection Models in R: Package SampleSelection,” *Journal of Statistical Software*, 27, 1–23.
- Vella, F. (1992), “Simple Tests for Sample Selection Bias in Censored and Discrete Choice Models,” *Journal of Applied Econometrics*, 7, 413–421.
- (1998), “Estimating Models with Sample Selection Bias: A Survey,” *The Journal of Human Resources*, 33, 127–169.
- von Mises, R. (1947), “On the Asymptotic Distribution of Differentiable Statistical Functions,” *The Annals of Mathematical Statistics*, 18, 309–348.
- White, H. (1980), “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica*, 48, 817–838.
- Winship, C. and Mare, R. D. (1992), “Models for Sample Selection Bias,” *Annual Review of Sociology*, 18, 327–350.
- Zhelonkin, M. (2013), “Robustness in Sample Selection Models,” *PhD Thesis, University of Geneva, Switzerland*.
- Zuehlke, T. W. and Zeman, A. R. (1991), “A Comparison of Two-Stage Estimators of Censored Regression Models,” *The Review of Economics and Statistics*, 73, 185–188.

Table 1: Bias, Variance and MSE of the classical, robust probit, and semiparametric binary regression estimator at the model and under two types of contamination.

$N = 1000$	Not contaminated			x_1 is contaminated, $y_1 = 1$			x_1 is contaminated, $y_1 = 0$		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
Classical									
β_{10}	-0.013	0.015	0.015	-0.074	0.011	0.016	-0.194	0.012	0.050
β_{11}	0.007	0.005	0.005	-0.323	0.005	0.110	-0.291	0.005	0.089
β_{12}	0.002	0.011	0.011	-0.456	0.013	0.221	-0.419	0.013	0.189
β_{13}	0.008	0.004	0.004	-0.274	0.004	0.079	-0.247	0.004	0.065
Robust									
β_{10}	-0.011	0.016	0.016	-0.011	0.016	0.016	-0.013	0.016	0.016
β_{11}	0.008	0.006	0.006	0.006	0.006	0.006	0.003	0.006	0.006
β_{12}	0.004	0.013	0.013	0.002	0.013	0.013	-0.003	0.013	0.013
β_{13}	0.009	0.004	0.004	0.007	0.004	0.004	0.004	0.005	0.005
Klein-Spady+Probit									
β_{10}	-0.014	0.020	0.020	0.024	0.010	0.011	-0.102	0.011	0.021
β_{11}	0.005	0.005	0.005	-0.364	0.006	0.138	-0.330	0.005	0.114
β_{12}	-0.001	0.015	0.015	-0.352	0.011	0.135	-0.320	0.011	0.113
β_{13}	0.007	0.005	0.005	-0.267	0.004	0.075	-0.240	0.004	0.062

Table 2: Bias, Variance and MSE of the classical and robust two-stage estimators at the model and under two types of contamination, when the exclusion restriction is not available.

$N = 1000$	Not contaminated			x_1 is contaminated, $y_1 = 1$			x_1 is contaminated, $y_1 = 0$		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
Classical									
β_{20}	0.000	0.064	0.064	-1.872	0.445	3.947	-0.695	0.339	0.822
β_{21}	-0.004	0.016	0.016	0.615	0.044	0.422	0.197	0.046	0.085
β_{22}	0.000	0.023	0.023	0.406	0.040	0.205	0.111	0.041	0.053
β_{23}	0.001	0.011	0.011	0.411	0.022	0.191	0.129	0.025	0.041
β_λ	-0.003	0.073	0.073	2.237	0.491	5.497	0.682	0.350	0.815
Robust probit + OLS									
β_{20}	0.001	0.064	0.064	-0.520	0.051	0.322	-0.004	0.065	0.065
β_{21}	-0.005	0.016	0.016	0.229	0.012	0.064	-0.003	0.016	0.016
β_{22}	-0.001	0.024	0.024	0.217	0.021	0.068	0.001	0.024	0.024
β_{23}	-0.001	0.011	0.011	0.172	0.008	0.038	0.002	0.011	0.011
β_λ	-0.005	0.073	0.073	0.653	0.040	0.466	0.001	0.074	0.074
Robust 2S									
β_{20}	-0.027	0.080	0.081	-0.072	0.075	0.080	-0.030	0.081	0.082
β_{21}	-0.005	0.020	0.020	0.025	0.018	0.019	0.006	0.020	0.020
β_{22}	0.009	0.027	0.027	0.028	0.026	0.027	0.008	0.028	0.028
β_{23}	0.008	0.013	0.013	0.022	0.012	0.013	0.008	0.013	0.013
β_λ	0.019	0.099	0.099	0.078	0.088	0.094	0.021	0.100	0.100

Table 3: Bias, Variance and MSE of the classical, robust two-stage estimators, and QRE at the model and under two types of contamination, when the exclusion restriction is available.

$N = 1000$	Not contaminated			x_1 is contaminated, $y_1 = 1$			x_1 is contaminated, $y_1 = 0$		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
Classical									
β_{20}	0.006	0.015	0.015	-0.638	0.062	0.469	-0.249	0.032	0.094
β_{21}	-0.000	0.005	0.005	0.153	0.009	0.032	0.036	0.006	0.008
β_{22}	0.002	0.011	0.011	0.002	0.015	0.015	-0.024	0.012	0.012
β_{23}	-0.004	0.002	0.002	-0.041	0.002	0.004	-0.004	0.002	0.002
β_λ	-0.003	0.016	0.016	1.018	0.090	1.127	0.225	0.038	0.089
Robust probit + OLS									
β_{20}	0.007	0.015	0.015	-0.150	0.017	0.039	0.005	0.015	0.015
β_{21}	-0.000	0.005	0.005	0.078	0.005	0.011	0.000	0.005	0.005
β_{22}	0.002	0.011	0.011	0.056	0.012	0.016	0.002	0.011	0.011
β_{23}	-0.004	0.002	0.002	-0.028	0.002	0.003	-0.004	0.002	0.002
β_λ	-0.003	0.017	0.016	0.368	0.012	0.148	-0.001	0.017	0.017
Robust 2S									
β_{20}	-0.001	0.017	0.017	-0.011	0.017	0.017	-0.003	0.017	0.017
β_{21}	-0.001	0.005	0.005	0.003	0.005	0.005	-0.001	0.005	0.005
β_{22}	0.004	0.012	0.012	0.008	0.012	0.012	0.004	0.012	0.012
β_{23}	-0.004	0.002	0.002	-0.005	0.002	0.002	-0.004	0.002	0.002
β_λ	-0.001	0.021	0.021	0.023	0.021	0.021	0.001	0.022	0.022
QRE									
β_{20}	0.020	0.168	0.168	-0.029	5.998	5.999	0.099	3.299	3.310
β_{21}	-0.009	0.010	0.010	0.011	0.009	0.009	-0.011	0.009	0.009
β_{22}	-0.001	0.021	0.021	0.005	0.021	0.021	0.002	0.021	0.021
β_{23}	-0.005	0.003	0.003	-0.005	0.003	0.003	-0.006	0.003	0.003
β_λ	-0.234	11.370	11.370	-0.575	338.820	339.150	-1.306	174.335	176.040

Table 4: Estimation results of the Medical Expenditures data by the classical estimator, by the robust two-stage estimator from Section 3.1, and by the OLS. The standard errors are given in parentheses. Significance codes: “****” 0.001, “**” 0.01, “*” 0.05, “.” 0.1.

	Without Ex.Res.		With Ex.Res.		
	Classical	Robust	Classical	Robust	OLS
Selection					
intercept	−0.71771 (0.19247)***	−0.74914 (0.19507)***	−0.66865 (0.19413)***	−0.70043 (0.19640)***	
age	0.09732 (0.02702)***	0.10541 (0.02770)***	0.08682 (0.02746)**	0.09459 (0.02814)***	
female	0.64421 (0.06015)***	0.68741 (0.06226)***	0.66351 (0.06097)***	0.70361 (0.06298)***	
educ	0.07017 (0.01134)***	0.07012 (0.01147)***	0.06188 (0.01204)***	0.06231 (0.01212)***	
blhisp	−0.37449 (0.06175)***	−0.39775 (0.06265)***	−0.36578 (0.06191)***	−0.38861 (0.06280)***	
totchr	0.79352 (0.07112)***	0.83284 (0.08028)***	0.79575 (0.07122)***	0.83405 (0.08023)***	
ins	0.18124 (0.06259)**	0.18256 (0.06371)**	0.16911 (0.06293)**	0.17255 (0.06403)**	
income			0.00268 (0.00131)*	0.00253 (0.00134)·	
Outcome					
intercept	5.30257 (0.29414)***	5.40154 (0.27673)***	5.28893 (0.28852)***	5.40933 (0.27291)***	4.90783 (0.16815)***
age	0.20212 (0.02430)***	0.20062 (0.02451)***	0.20247 (0.02422)***	0.20029 (0.02447)***	0.21723 (0.02222)***
female	0.28916 (0.07369)***	0.25501 (0.06992)***	0.29213 (0.07258)***	0.25214 (0.06994)***	0.37938 (0.04858)***
educ	0.01199 (0.01168)	0.01325 (0.01162)	0.01239 (0.01157)	0.01318 (0.01158)	0.02223 (0.00976)*
blhisp	−0.18106 (0.06585)**	−0.15508 (0.06507)*	−0.18287 (0.06534)**	−0.15342 (0.06514)*	−0.23853 (0.05519)***
totchr	0.49833 (0.04947)***	0.48116 (0.03822)***	0.50063 (0.04855)***	0.47956 (0.03805)***	0.56182 (0.03051)***
ins	−0.04740 (0.05315)	−0.06707 (0.05159)	−0.04651 (0.05297)	−0.06825 (0.05174)	−0.02082 (0.05001)
inverse Mills ratio	−0.4802 (0.2907)·	−0.67676 (0.25928)**	−0.4637 (0.2826)	−0.68995 (0.25544)**

Figure 1: Comparison of classical and robust two-stage estimators without contamination, when the exclusion restriction is not available (top panel), and when it is available (bottom panel). Case (a) corresponds to the classical estimator, (b) corresponds to robust probit with OLS in the second stage, and (c) to robust two-stage. Horizontal lines mark the true values of the parameters.

Figure 2: Comparison of classical and robust two-stage estimators with contamination, when the exclusion restriction is available. Top panel corresponds to $y_1 = 0$, and bottom panel corresponds to $y_1 = 1$. Case (a) corresponds to the classical estimator, (b) corresponds to robust probit with OLS in the second stage, and (c) to robust two-stage. Horizontal lines mark the true values of the parameters.

Figure 3: Comparison of classical and robust two-stage estimators with contamination, when the exclusion restriction is not available. Top panel corresponds to $y_1 = 0$, and bottom panel corresponds to $y_1 = 1$. Case (a) corresponds to the classical estimator, (b) corresponds to robust probit with OLS in the second stage, and (c) to robust two-stage. Horizontal lines mark the true values of the parameters.

Figure 4: ROC curves for the sample selection bias test. Top left plot corresponds to the case without exclusion restriction without contamination, top right plot corresponds to the case without exclusion restriction with contamination, bottom left plot corresponds to the case with exclusion restriction without contamination, and bottom right plot corresponds to the case with exclusion restriction with contamination. The solid line corresponds to the classical test and the dashed line corresponds to the robust test.

Figure 5: Relative efficiency of the robust two-stage estimator to the classical two-stage estimator. The top and bottom panels correspond to the cases when the exclusion restriction is not available and when it is available, respectively. The x axis corresponds to the proportion of contamination ϵ , where x_1 is contaminated and the corresponding $y_1 = 1$.