ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)
78, *Part* 4, *pp.* 805–827

Robust inference in sample selection models

Mikhail Zhelonkin,

University of Lausanne, Switzerland

Marc G. Genton

King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

and Elvezio Ronchetti

University of Geneva, Switzerland

[Received July 2013. Final revision July 2015]

Summary. The problem of non-random sample selectivity often occurs in practice in many fields. The classical estimators introduced by Heckman 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. 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 Heckman's two-stage estimator, and we demonstrate the non-robustness of this estimator and its estimated variance to small deviations from the model assumed. 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 we 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, i.e. 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 explanatory variables, the spending amount is not independent from the decision to spend, i.e. they are dependent through unobservables. This type of problem arises in many research fields besides economics, including sociology, political science and finance (see, for example, 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 regression system

$$y_{1i}^{*} = x_{1i}^{\mathrm{T}}\beta_1 + e_{1i}, \tag{1}$$

Address for correspondence: Mikhail Zhelonkin, Department of Operations, University of Lausanne, Office 3016, Anthropole Building, CH-1005 Lausanne, Switzerland.
E-mail: Mikhail.Zhelonkin@unil.ch

 1369–7412/16/78805

$$y_{2i}^* = x_{2i}^{\mathrm{T}}\beta_2 + e_{2i}, \tag{2}$$

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 $\sigma_1^2 = 1$ and σ_2^2, and correlation ρ. Note that the variance parameter σ_1^2 is set to be equal to 1 to ensure identifiability. Here equation (1) is the selection equation, defining the observability rule, and equation (2) is the equation of interest. The observed variables are defined by

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

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

where I is the indicator function. This model was 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 is a maximum likelihood estimator (MLE) based on the assumption of bivariate normality of the error terms. The second is a two-stage procedure. Consider the conditional expectation of y_{2i} given x_{2i} and the selection rule

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

Then the conditional expectation of the error term is in general different from 0, which leads to the modified regression

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

where $\beta_\lambda = \rho\sigma_2, \lambda(x_{1i}^{\mathrm{T}}\beta_1) = \phi(x_{1i}^{\mathrm{T}}\beta_1)/\Phi(x_{1i}^{\mathrm{T}}\beta_1)$ 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. 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 equation (5), where the additional variable corrects for the sample selection bias.

Both estimation procedures have advantages and drawbacks, studied extensively in the literature; see for example Stolzenberg and Relles (1997), Puhani (2000), Toomet and Henningsen (2008) and the general reviews by Winship and Mare (1992) and 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 and is well documented by several Monte Carlo studies, where the behaviour of these estimators has been investigated under different distributional assumptions; see Paarsch (1984) and Zuehlke and Zeman (1991). Another important issue is the presence of outlying observations, which is a well-known problem in many classical models and 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 labour 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 long-tailed 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 (2015), a copula-based approach by Smith (2003) and the generalized additive model with location, scale and shape 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 neighbourhoods (in a topological sense) of the central model and do not offer full protection against possible distributional deviations from the central model; see Hampel *et al.* (1986), page 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 researchers have proposed semiparametric (Ahn and Powell, 1993; Newey, 2009; Marra and Radice, 2013) and non-parametric (Das *et al.*, 2003) methods. Gallant and Nychka (1987) proposed a semi-non-parametric 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 non-parametric set-up. 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) neighbourhood of the latter. We then derive estimators and tests which are still reliable in the full neighbourhood. This has the advantage of providing insurance and protection against small but harmful distributional deviations and still to 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 non-parametric methods would be in principle preferable; see the discussion in Hampel *et al.* (1986), page 7. Note, however, that non-parametric methods are not necessarily designed to be robust in the sense mentioned above. Even the arithmetic mean, which is the non-parametric 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), page 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 equations (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 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 robust version can in principle be extended to these situations.

In recent decades, robust estimators and tests have been developed for large classes of models in both the statistical and the econometric literature; see for instance, Huber (1981), Huber and Ronchetti (2009), Hampel *et al.* (1986) and 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 that were mentioned above, the robustness aspects in sample selection models have not received much 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 (IF) and the change-of-variance function (CVF); 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 neighbourhood of the bivariate normal distribution assumed. It turns out that both functions are unbounded and this implies that a small deviation can have a large influence 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 two-stage estimator and test of SSB. They are available in the R package `ssmrob` in the Comprehensive R Archive Network. We study the performance of our estimators in both the presence and the 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). A technical derivation and assumptions are provided in Appendix A. The on-line 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 equations (6) and (7). In particular, we derive its IF and its CVF and discuss the robustness properties of 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 $z_{ji} = (x_{ji}, y_{ji})$, $j = 1, 2$, $i = 1, \ldots, N$, and let F be the distribution function of z_i. 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 (e.g. Zhelonkin *et al.* (2012)) can be expressed as a solution of the empirical counterpart of the system

$$\int \Psi_1\{(x_1, y_1); S(F)\}\, \mathrm{d}F = 0, \tag{6}$$

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

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

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

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

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

2.1. Influence function

For a given functional $T(F)$, the IF was defined by Hampel (1974) as

$$\mathrm{IF}(z; T, F) = \lim_{\epsilon \to 0} [T\{(1-\epsilon)F + \epsilon\Delta_z\} - T(F)]/\epsilon,$$

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 neighbourhood described by the perturbations $F_\epsilon = (1-\epsilon)F + \epsilon G$, where G is some arbitrary distribution function, is approximately

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

Therefore, a condition for (local) robustness is a bounded IF with respect to z, which means that if the $\mathrm{IF}(\cdot;\cdot,\cdot)$ is unbounded then the bias of the estimator can become arbitrarily large.

The following proposition gives the IF of Heckman's two-stage estimator.

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

$$\begin{aligned} \mathrm{IF}(z; T, F) = \left\{ \int \begin{pmatrix} x_2 x_2^{\mathrm{T}} & \lambda x_2 \\ \lambda x_2^{\mathrm{T}} & \lambda^2 \end{pmatrix} y_1 \, \mathrm{d}F \right\}^{-1} \Bigg\{ (y_2 - x_2^{\mathrm{T}}\beta_2 - \lambda\beta_\lambda) \begin{pmatrix} x_2 \\ \lambda \end{pmatrix} y_1 \\ + \int \begin{pmatrix} x_2\beta_\lambda \\ \lambda\beta_\lambda \end{pmatrix} y_1 \lambda' \, \mathrm{d}F \, \mathrm{IF}(z; S, F) \Bigg\}, \end{aligned} \tag{10}$$

where

$$\lambda' = \frac{-\Phi(x_1^{\mathrm{T}}\beta_1)\phi(x_1^{\mathrm{T}})x_1^{\mathrm{T}}\beta_1 - \phi(x_1^{\mathrm{T}}\beta_1)^2}{\Phi(x_1^{\mathrm{T}}\beta_1)^2} x_1^{\mathrm{T}}$$

and

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

The proof is given in the on-line supplementary material.

The first term of equation (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. Note that the function λ is unbounded from the left and tends to 0 from the right. From equation (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 IF.

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 by using the M-estimation framework. Using the result in Hampel *et al.* (1986), the general expression of the asymptotic variance is given by

$$V(T, F) = \int \mathrm{IF}(z; T, F)\, \mathrm{IF}(z; T, F)^{\mathrm{T}} \,\mathrm{d}F(z).$$

Specifically, denote the components of the IF as

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

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

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

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

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

where Δ is a diagonal matrix with elements $\delta_{ii} = \partial\lambda(x_{1i}\beta_1)/\partial(x_{1i}\beta_1)$, the matrix X consists of vectors $\binom{x^{2i}}{\lambda_i}$ and $\mathrm{var}(S, F)$ 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 behaviour 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 $\mathrm{CVF}(z; T, F) = [(\partial/\partial\epsilon) V\{T, (1-\epsilon)F + \epsilon\Delta_z\}]_{\epsilon=0}$, 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

$$V(T, F_\epsilon) \cong V(T, F) \exp\left\{\epsilon \int \frac{\mathrm{CVF}(z; T, F)}{V(T, F)} \,\mathrm{d}G\right\}. \tag{13}$$

If $\mathrm{CVF}(z; T, F)$ is unbounded then the variance can behave unpredictably (arbitrarily large or small); see Hampel *et al.* (1986), page 175. Using expression (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 Heckman's two-stage estimator is given by

$$\begin{aligned}
\mathrm{CVF}(z; S, T, F) = {} & V - M(\Psi_2)^{-1} \left\{ \int D_{\mathrm{H}} \,\mathrm{d}F + \begin{pmatrix} x_2 x_2^{\mathrm{T}} & \lambda x_2 \\ \lambda x_2^{\mathrm{T}} & \lambda^2 \end{pmatrix} y_1 \right\} V \\
& + M(\Psi_2)^{-1} \int \{A_{\mathrm{H}} a^{\mathrm{T}} + A_{\mathrm{H}} b^{\mathrm{T}} + B_{\mathrm{H}} b^{\mathrm{T}}\} \,\mathrm{d}F \, M(\Psi_2)^{-1} \\
& + M(\Psi_2)^{-1} \int \{a A_{\mathrm{H}}^{\mathrm{T}} + b A_{\mathrm{H}}^{\mathrm{T}} + b B_{\mathrm{H}}^{\mathrm{T}}\} \,\mathrm{d}F \, M(\Psi_2)^{-1} \\
& + M(\Psi_2)^{-1} \{a(z)\, a(z)^{\mathrm{T}} + a(z)\, b(z)^{\mathrm{T}} + b(z)\, a(z)^{\mathrm{T}} + b(z)\, b(z)^{\mathrm{T}}\} \, M(\Psi_2)^{-1} \\
& - V \left\{ \int D_{\mathrm{H}} \,\mathrm{d}F + \begin{pmatrix} x_2 x_2^{\mathrm{T}} & \lambda x_2 \\ \lambda x_2^{\mathrm{T}} & \lambda^2 \end{pmatrix} y_1 \right\} M(\Psi_2)^{-1},
\end{aligned} \tag{14}$$

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

The CVF has several sources of unboundedness. The second term of equation (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 have been discussed in general in La Vecchia *et al.* (2012).

2.3. Sample selection bias test

Heckman (1979) proposed to test for the selection bias by 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 for example Vella (1992)), but the simple Heckman test is the most widely used in applications. Here we investigate the effect of contamination on the test statistic $\tau_n = \hat{\beta}_\lambda \sqrt{n} / \sqrt{V(\beta_\lambda, F)}$.

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:

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

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), page 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.

Note, 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 we propose a simple robust alternative to the SSB test.

3.1. Robust two-stage estimator

From the expression of the IF in equation (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 to compute λ, which will transfer potential leverage effects from the first stage to the second, and to use the robust estimator in the second stage, which will correct for the remaining outliers.

Consider the two-stage M-estimation framework that is given by equations (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 score function

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

where $\alpha(\beta_1) = (1/n) \Sigma_{i=1}^n E\{\nu(z_{1i}; \mu_i)\} \omega_1(x_{1i}) \mu_i'$ 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, $\mu_i = \mu_i(z_{1i}, \beta_1) = \Phi(x_{1i}^{\mathrm{T}} \beta_1)$ and $\mu_i' = \partial \mu_i / \partial \beta_1$.

The weight functions are defined by

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

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

$$\psi_{c_1}(r) = \begin{cases} r, & |r| \leqslant c_1, \\ c_1 \operatorname{sgn}(r), & |r| > c_1. \end{cases} \tag{16}$$

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 generalized linear model setting. A simple choice of the weight function $\omega_1(\cdot)$ is $\omega_{1i} = \sqrt{(1 - H_{ii})}$, where H_{ii} is the ith diagonal element of the hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. 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 $\mu_i = \Phi(x_{1i}^{\mathrm{T}}\beta_1)$ and $V(\mu_i) = \Phi(x_{1i}^{\mathrm{T}}\beta_1)\{1 - \Phi(x_{1i}^{\mathrm{T}}\beta_1)\}$ and hence the quasi-likelihood estimating equations are

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

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

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

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 Ψ-function

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

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

$$\omega(x_2, \lambda) = \begin{cases} x_2, & \text{if } d(x_2, \lambda) < c_m, \\ x_2 c_m / d(x_2, \lambda), & \text{if } d(x_2, \lambda) \geqslant c_m, \end{cases} \tag{18}$$

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 the 5% critical level.

The robust estimator that was 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 IMR 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. 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 much research and discussion; see Nelson (1984) and 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 that was 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 expression (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_2s, 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 $(x_2^{(1)}, \ldots, x_2^{(p_2)}, \lambda)$ into two independent components $(x_2^{(1)}, \ldots, x_2^{(q)})$ and $(x_2^{(q+1)}, \ldots, x_2^{(p_2)}, \lambda)$ and compute the robustness weights $\omega(x_2^{(1)}, \ldots, x_2^{(q)})$ and $\omega(x_2^{(q+1)}, \ldots, x_2^{(p_2)}, \lambda)$. Then, we combine these weights as $\omega(x_2, \lambda) = \omega(x_2^{(1)}, \ldots, x_2^{(q)})\,\omega(x_2^{(q+1)}, \ldots, x_2^{(p_2)}, \lambda)$, which guarantee robustness in this case. As a general rule, we suggest grouping 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 Appendix A, Heckman's two-stage estimator defined by equations (15) and (17) is robust, consistent and asymptotically normal, with asymptotic variance given by

$$V(T, F) = M(\Psi_2^{\mathrm{R}})^{-1} \int \{a_{\mathrm{R}}(z)\, a_{\mathrm{R}}(z)^{\mathrm{T}} + b_{\mathrm{R}}(z)\, b_{\mathrm{R}}(z)^{\mathrm{T}}\} \mathrm{d}F\, M(\Psi_2^{\mathrm{R}})^{-1}, \tag{19}$$

where

$$M(\Psi_2^{\mathrm{R}}) = -\int \frac{\partial}{\partial \beta_2} \Psi_2^{\mathrm{R}}(z; \lambda, T)\, \mathrm{d}F,$$

$a_{\mathrm{R}}(z) = \Psi_2^{\mathrm{R}}(z; \lambda, T)$ and

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

The asymptotic variance of the robust estimator has the same structure as that of the classical Heckman 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 by using standard techniques, e.g. the Newton–Raphson procedure. New R (R Development Core Team, 2012) functions for robust estimation and inference in sample selection models are provided in a package `ssmrob`.

Remark 1. In the likelihood framework, the Huber function defines the most efficient estimator, subject to a bounded IF. 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 MLE, 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. $H_0 : \beta_\lambda = 0$ *versus* $H_{\mathrm{A}} : \beta_\lambda \neq 0$, 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 equation (19).

The first term of equation (19), $M(\Psi_2^{\mathrm{R}})^{-1} \int a_{\mathrm{R}}(z)\, a_{\mathrm{R}}(z)^{\mathrm{T}} \mathrm{d}F\, M(\Psi_2^{\mathrm{R}})^{-1}$, is similar to the asymptotic variance of standard linear regression, but with heteroscedasticity. Therefore, we use the

Table 1. Bias, variance and mean-squared error MSE of the classical, robust probit and semiparametric binary regression estimator at the model and under two types of contamination ($N = 1000$)

Parameter	*Results for not contaminated*			*Results for x_1 is contaminated, $y_1 = 1$*			*Results for x_1 is contaminated, $y_1 = 0$*		
	Bias	*Variance*	*MSE*	*Bias*	*Variance*	*MSE*	*Bias*	*Variance*	*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 probit									
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

Eicker (1967)–Huber (1967)–White (1980) heteroscedasticity consistent variance estimator, i.e. we estimate this first term by

$$\hat{M}(\Psi_2^{\mathrm{R}})^{-1} \frac{1}{n} \sum \hat{a}(z_i)\, \hat{a}(z_i)^{\mathrm{T}}\, \hat{M}(\Psi_2^{\mathrm{R}})^{-1},$$

where $\hat{M}(\Psi_2^{\mathrm{R}})$ and $\hat{a}_{\mathrm{R}}(z)$ are the sample versions of M and $a_{\mathrm{R}}(z)$ respectively.

The second term of the asymptotic variance, $M(\Psi_2^{\mathrm{R}})^{-1} \int b_{\mathrm{R}}(z)\, b_{\mathrm{R}}(z)^{\mathrm{T}}\, \mathrm{d}F\, M(\Psi_2^{\mathrm{R}})^{-1}$, 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

$$\hat{M}(\Psi_2^{\mathrm{R}})^{-1} \frac{1}{n} \sum \hat{b}_{\mathrm{R}}(z_i)\, \hat{b}_{\mathrm{R}}(z_i)^{\mathrm{T}}\, \hat{M}(\Psi_2^{\mathrm{R}})^{-1} = \hat{M}(\Psi_2^{\mathrm{R}})^{-1} \frac{1}{n} \times \sum \frac{\partial \Psi_{2i}^{\mathrm{R}}}{\partial \beta_1} \widehat{\mathrm{var}}(S, F) \left(\frac{1}{n} \sum \frac{\partial \Psi_{2i}^{\mathrm{R}}}{\partial \beta_1} \right)^{\mathrm{T}} \hat{M}(\Psi_2^{\mathrm{R}})^{-1},$$

where

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

4. Numerical examples

4.1. Simulation study

We carry out a Monte Carlo study to illustrate the robustness issues in the model that was described in Section 1.1 and compare various estimators. In our experiment we generate $y_{1i}^* = x_{11i} + x_{12i} + 0.75 x_{13i} + e_{1i}$, where $x_{11i} \sim N(0, 1)$, $x_{12i} \sim N(-1, 0.5)$ and $x_{13i} \sim N(1, 1)$. For the

Table 2. Bias, variance and mean-squared error 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 ($N = 1000$)

Parameter	*Results for not contaminated*			*Results for x_1 is contaminated, $y_1 = 1$*			*Results for x_1 is contaminated, $y_1 = 0$*		
	Bias	*Variance*	*MSE*	*Bias*	*Variance*	*MSE*	*Bias*	*Variance*	*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 two stage									
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

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 give $\beta_\lambda = 0.7$. The degree of censoring is controlled by the intercept in the selection equation, which is denoted by β_{10} and set to 0, which corresponds to approximately 45% of censoring. Results (which are 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 $\beta_2 = (1.5, 1, 0.5)^{\mathrm{T}}$. 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). Note that the contaminating point deviates by two standard deviations from the centres of distributions of the explanatory variables, which are very difficult to identify by using standard exploratory analysis. In the on-line supplementary material we report additional simulations when the distribution of the error terms deviates from the normal distribution. 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 behaviour of the estimators is the same. A second simulation design (which is very simple and mostly pedagogical) with one explanatory variable is presented in the on-line supplementary material. In addition, we study there a type of contamination when the outliers emerge only in the selection stage.

Table 3. Bias, variance and mean-squared error MSE of the classical and robust two-stage estimators and the quantile regression estimator at the model and under two types of contamination, when the exclusion restriction is available ($N = 1000$)

Parameter	*Results for not contaminated*			*Results for x_1 is contaminated, $y_1 = 1$*			*Results for x_1 is contaminated, $y_1 = 0$*		
	Bias	*Variance*	*MSE*	*Bias*	*Variance*	*MSE*	*Bias*	*Variance*	*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 two stage									
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
Quantile regression estimator									
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

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 estimator. Moreover, when an exclusion restriction is available, we add the quantile regression estimator. This is the estimator that was 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 by 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 on-line supplementary material.

In Table 1 we first consider only the first stage. We note 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 estimator). However, under contamination only the robust proposal remains nearly unbiased. In Tables 2 and 3 and Figs 1–3 we consider classical and robust two-stage estimators. The quantile regression estimator is included 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 Fig. 2 (with exclusion restriction) and Fig. 3 (without exclusion restriction). When the exclusion restriction is not available the mag-

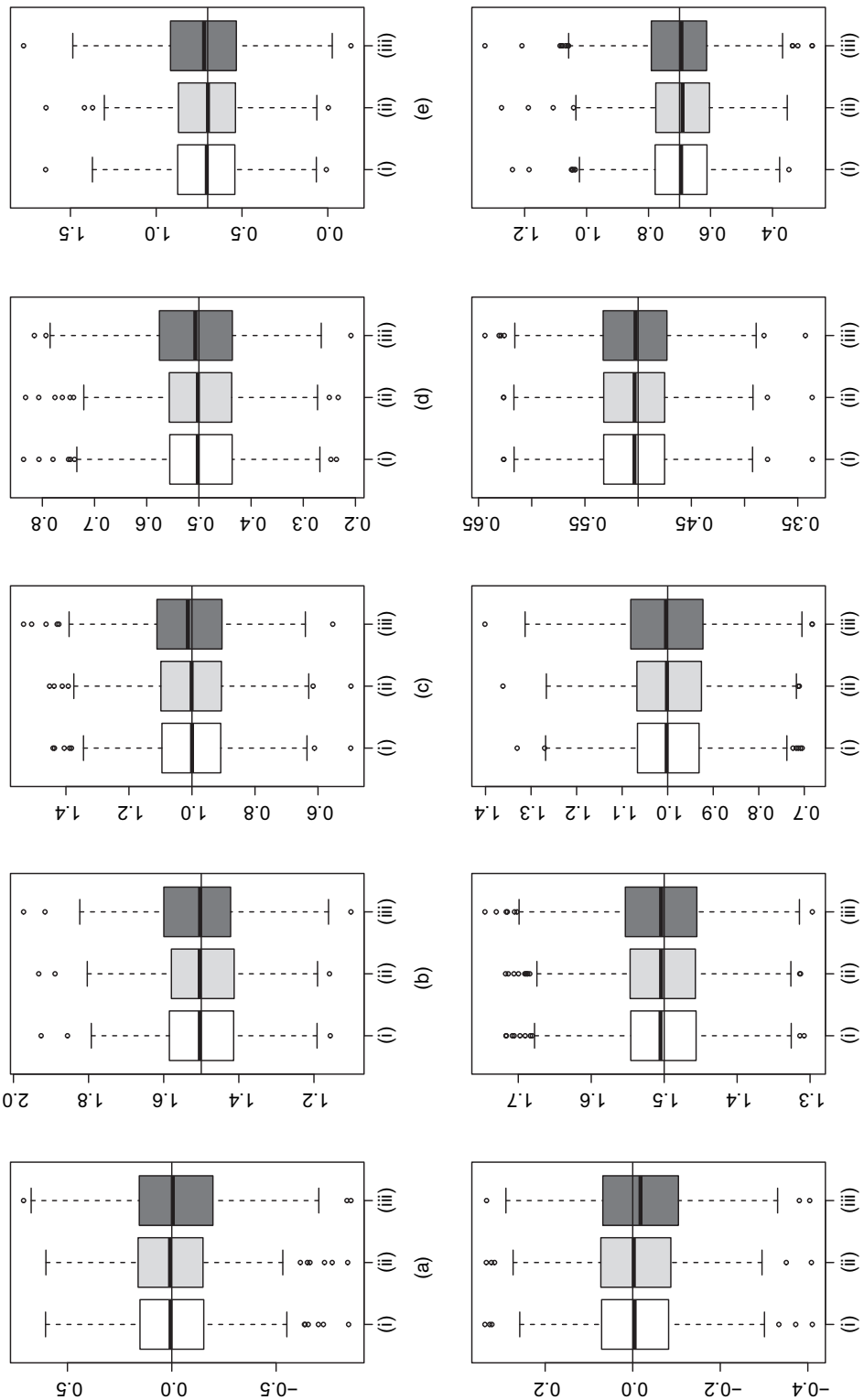

Fig. 1. Comparison of classical and robust two-stage estimators without contamination (stage 2) when the exclusion restriction is (a)–(e) not available and (f)–(j) available (case (i) corresponds to the classical estimator, case (ii) to the robust probit model with OLS in the second stage and (iii) to the robust two-stage estimator; ———, true values of the parameters): (a), (f) β_{20}; (b), (g) β_{21}; (c), (h) β_{22}; (d), (i) β_{23}; (e), (j) β_{λ}

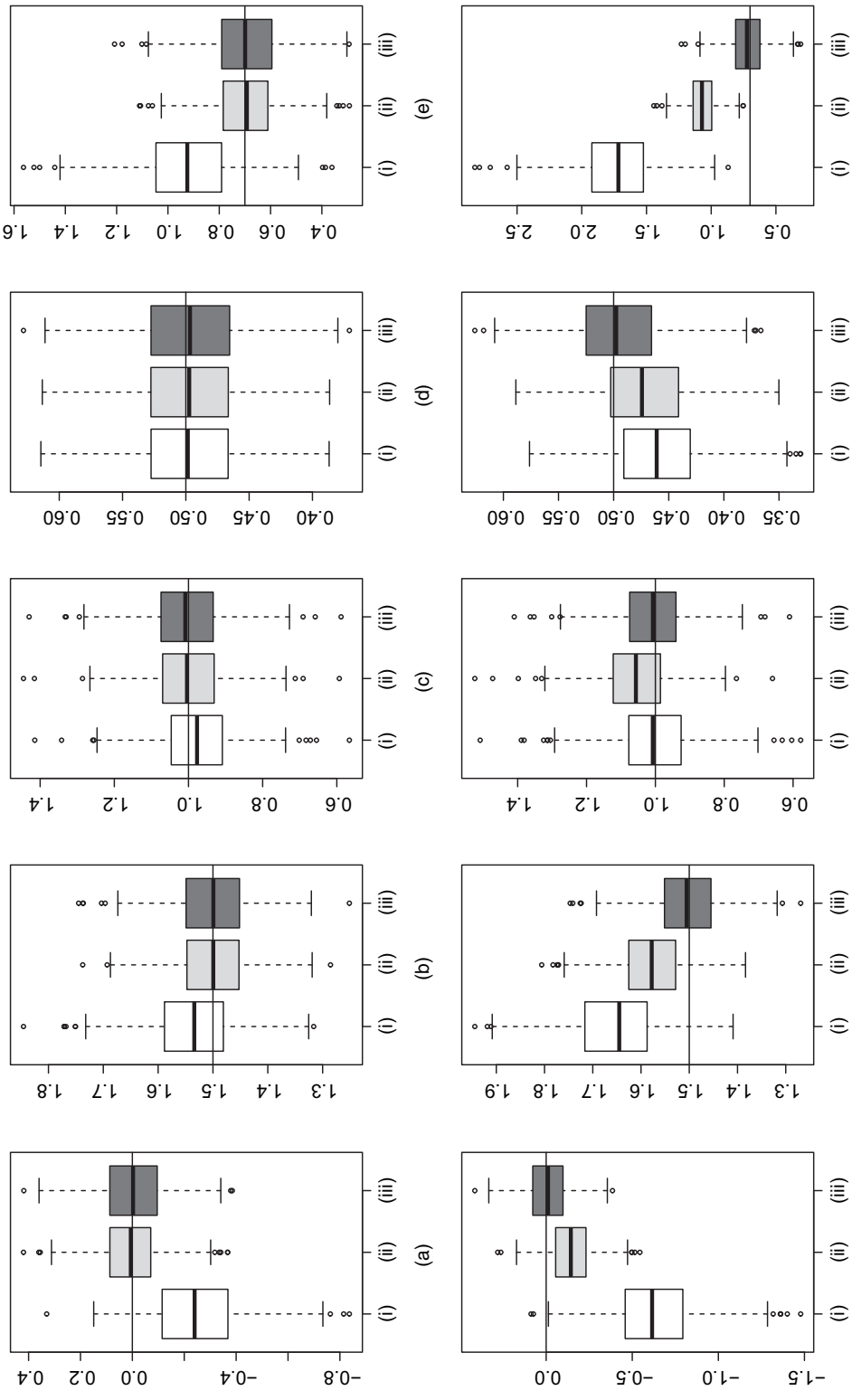

Fig. 2. Comparison of classical and robust two-stage estimators with contamination (stage 2) when the exclusion restriction is available and (a)–(e) $y_1 = 0$ and (f)–(j) $y_1 = 1$ (case (i) corresponds to the classical estimator, case (ii) to the robust probit model with OLS in the second stage and (iii) to the robust two-stage estimator; ———, true values of the parameters): (a), (f) β_{20}; (b), (g) β_{21}; (c), (h) β_{22}; (d), (i) β_{23}; (e), (j) β_λ

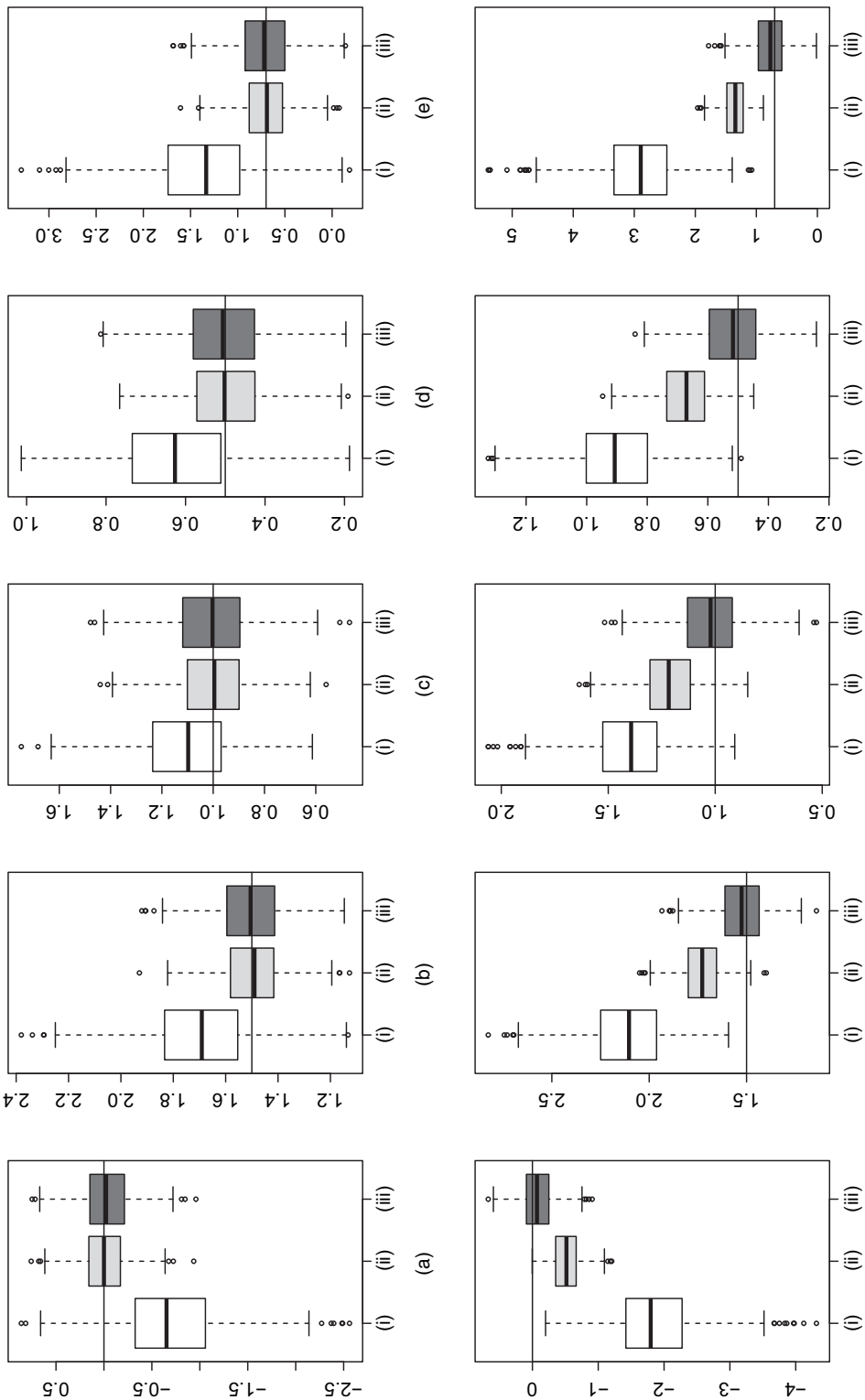

Fig. 3. Comparison of classical and robust two-stage estimators with contamination (stage 2) when the exclusion restriction is not available and (a)–(e) $y_1 = 0$ and (f)–(j) $y_1 = 1$ (case (i) corresponds to the classical estimator, case (ii) to the robust probit model with OLS in the second stage and (iii) to the robust two-stage estimator; ———, true values of the parameters): (a), (f) β_{20}; (b), (g) β_{21}; (c), (h) β_{22}; (d), (i) β_{23}; (e), (j) β_λ

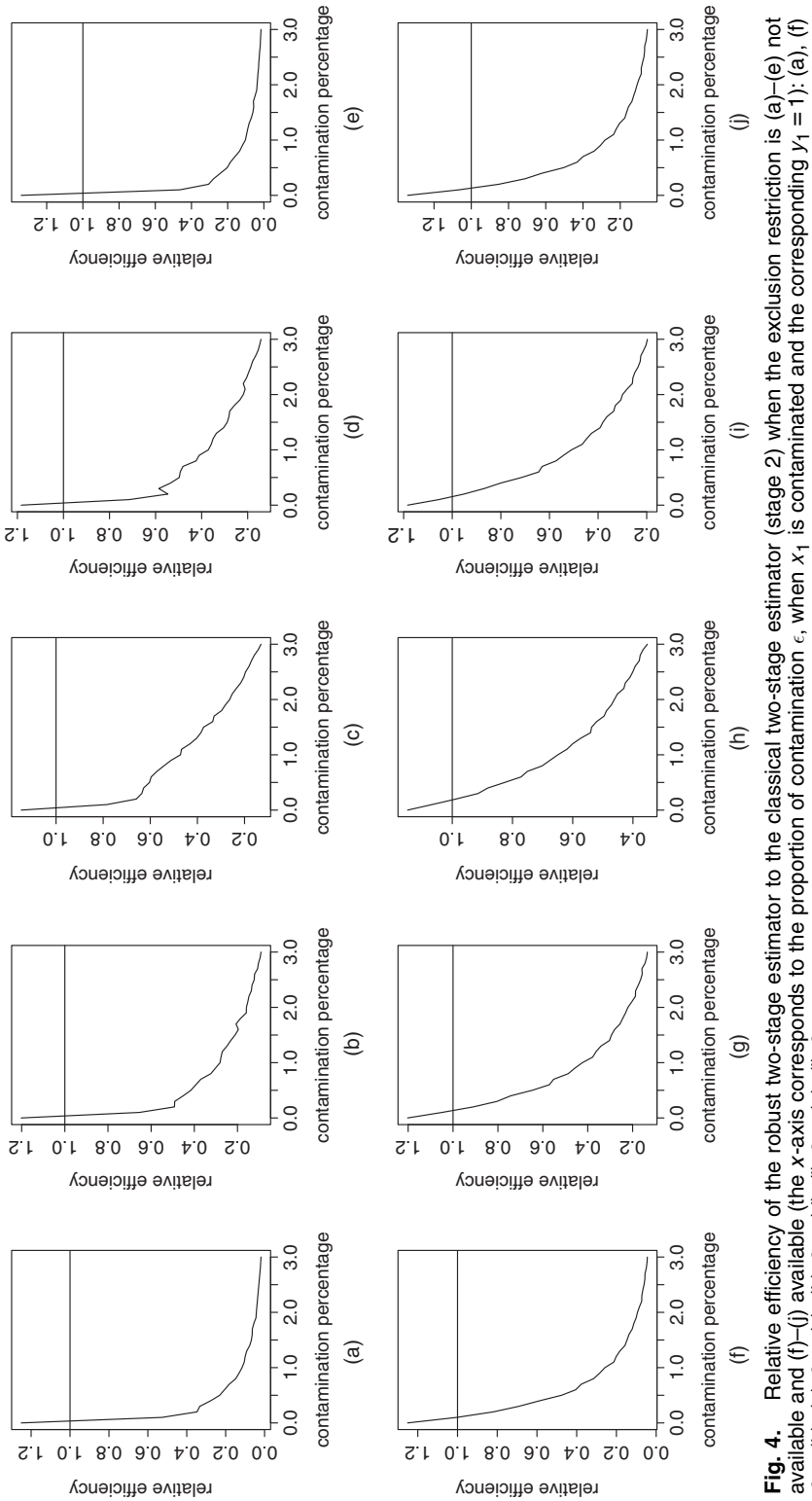

Fig. 4. Relative efficiency of the robust two-stage estimator to the classical two-stage estimator (stage 2) when the exclusion restriction is (a)–(e) not available and (f)–(j) available (the *x*-axis corresponds to the proportion of contamination ϵ, when x_1 is contaminated and the corresponding $y_1 = 1$): (a), (f) β_{20}; (b), (g) β_{21}; (c), (h) β_{22}; (d), (i) β_{23}; (e), (j) β_λ

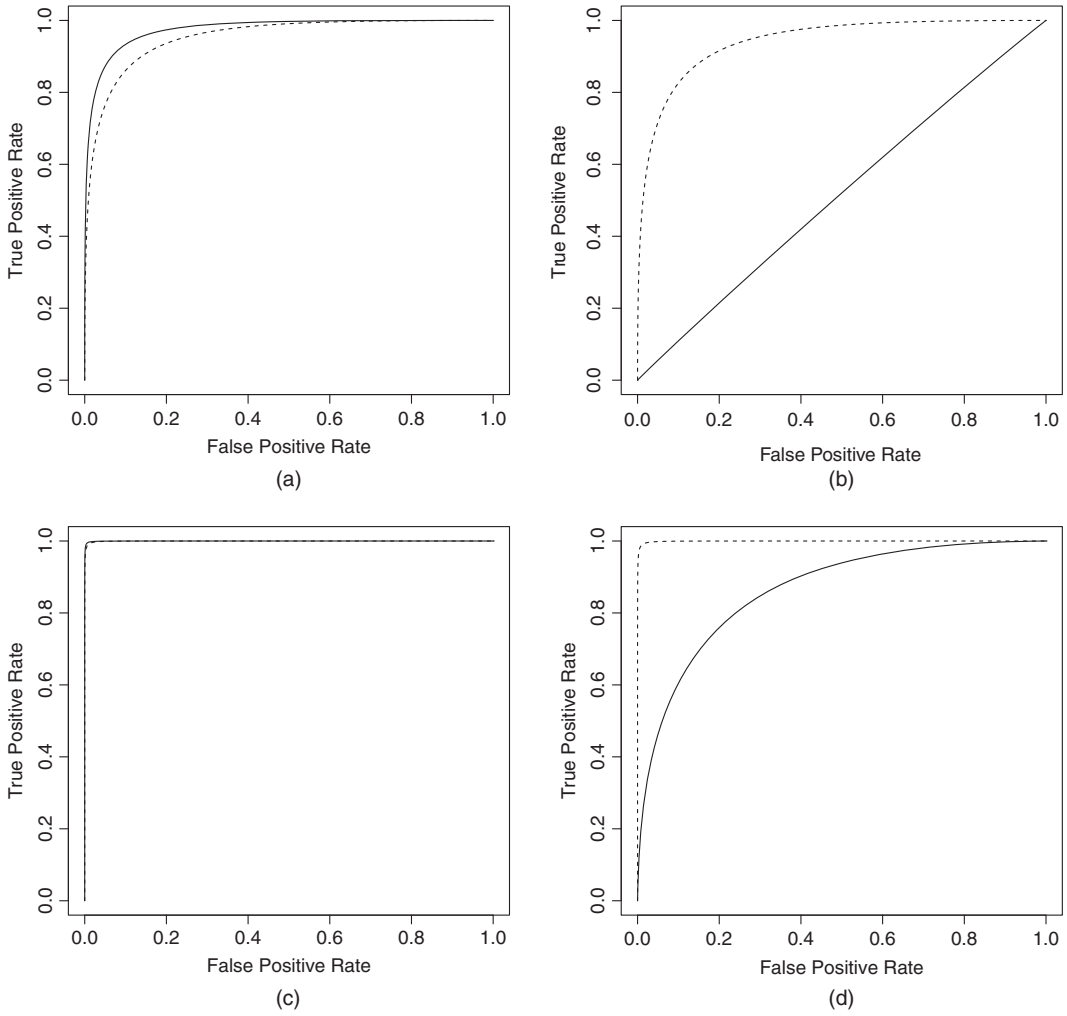

Fig. 5. Receiver operating characteristic curves for the sample selection bias test (———, classical test; - - - - -, robust test): (a) case without exclusion restriction without contamination; (b) case without exclusion restriction with contamination; (c) case with exclusion restriction without contamination; (d) case with exclusion restriction with contamination

nitude 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 quantile regression estimator are robust. However, the estimators of the intercept and β_λ become severely biased. Athough 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, note that the quantile regression estimators of the slopes have larger mean-squared error than those of the robust two-stage estimator.

In the case when the outlier is not transferred to the equation of interest (Figs 2(a)–2(e) and 3(a)–3(e)) it is enough to use a robust probit but, when the outlier emerges in the equation of interest (Figs 2(f)–2(j) and 3(f)–3(j)), a robust estimation of the second stage is necessary. In this case the outliers influence not only both estimation stages directly, but also the effect of contamination is amplified by the influence through λ. The behaviour of the variances of the

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 OLS, with standard errors in parentheses

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

†Level of significance 0.001.
‡Level of significance 0.01.
§Level of significance 0.05.
§§Level of significance 0.1.

robust estimators remains stable, whereas the variance of the classical estimator is seriously affected by the contamination.

In Fig. 4 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 0% 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 on-line supplementary material). As is expected from the theory, the robust estimator is less efficient than the classical estimator, 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 estimator with only 0.1% contamination (Fig. 4(e)). The efficiency loss of the classical estimator concerns not only the IMR parameter, but also the other explanatory variables. Note that the behaviour of the variance of the robust estimator remains stable under contamination (Table 1 and Table 2). Finally, in Fig. 5 we plot the receiver operating characteristic curves. The data-generating process 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 illustrate the behaviour of our new robust methodology further, we consider the data on ambulatory expenditures from the 2001 Medical Expenditure Panel Survey that were analysed by Cameron and Trivedi (2009), page 545. The data consist of 3328 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 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 by 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 by 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 we should use the OLS estimator. However, if the presence of sample selectivity is hidden by the deviation from the model assumed, then 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 by using the sample selection model. Cameron and Trivedi (2009) noted 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 that are similar to those obtained by 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 IMR parameter. The robust estimator returns $\hat{\beta}_{\mathrm{RIMR}} = -0.677$, compared with Heckman's $\hat{\beta}_{\mathrm{IMR}} = -0.480$. 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 $\hat{\beta}_{\mathrm{RIMR}} = -0.690$ 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 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 us to deal with data deviating from the assumed model and to carry out reliable inference even in the presence of small deviations from the normality model assumed. 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 model assumed, at an early exploratory stage one could look for possible large deviations. Robustness against this type of deviation would require the development of high breakdown estimators, which could possibly be obtained by replacing the Huber function in expression (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 on-line supplementary material.

Acknowledgements

The authors thank the Joint 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. Ronchetti's research was partially supported by Swiss National Science Foundation grant 100018–140295.

Appendix A: Assumptions and proof of proposition 3

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

(a) $z_1,\ldots,z_N$ is a sequence of independent identically distributed random vectors with distribution F defined on a space $\mathcal{Z}$;
(b) Θ is a compact subset of $\mathbb{R}^{p_1+p_2}$;
(c) $\int \Psi^{\mathrm{R}}(z;\theta)\,\mathrm{d}F=0$ has a unique solution θ_0 in the interior of Θ;
(d) $\Psi^{\mathrm{R}}(z;\theta)$ and $\partial\Psi^{\mathrm{R}}(z;\theta)/\partial\theta$ are measurable for each θ in Θ and continuous for each z in $\mathcal{Z}$, and there are F-integrable functions ξ_1 and ξ_2 such that, for all $\theta\in\Theta$ and $z\in\mathcal{Z}$, $|\Psi^{\mathrm{R}}(z;\theta)\,\Psi^{\mathrm{R}}(z;\theta)^{\mathrm{T}}|\leqslant\xi_1$ and $|\partial\Psi^{\mathrm{R}}(z;\theta)/\partial\theta|\leqslant\xi_2$;
(e) $\int \Psi^{\mathrm{R}}(z;\theta)\,\Psi^{\mathrm{R}}(z;\theta)^{\mathrm{T}}\,\mathrm{d}F$ is non-singular for each $\theta\in\Theta$;
(f) $\int \partial\Psi^{\mathrm{R}}(z;\theta_0)/\partial\theta\,\mathrm{d}F$ is finite and non-singular.

A.1. 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 $a_{\mathrm{R}}(z)\,b_{\mathrm{R}}(z)^{\mathrm{T}}$ and $b_{\mathrm{R}}(z)\,a_{\mathrm{R}}(z)^{\mathrm{T}}$ vanish after integration, owing to the independence of the error terms.

References

Ahn, H. and Powell, J. L. (1993) Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *J. Econmetr.*, **58**, 3–29.
Amemiya, T. (1984) Tobit models: a survey. *J. Econmetr.*, **24**, 3–61.
Buchinsky, M. (1998) The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *J. Appl. Econmetr.*, **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. *J. Quant. Crimin.*, **23**, 151–178.
Cameron, C. A. and Trivedi, P. K. (2009) *Microeconometrics using Stata.* College Station: Stata Press.
Cantoni, E. and Ronchetti, E. (2001) Robust inference for generalized linear models. *J. Am. Statist. Ass.*, **96**, 1022–1030.
Collier, D. and Mahoney, J. (1996) Insights and pitfalls: selection bias in qualitative research. *Wrld Polit.*, **49**, 56–91.
Copas, J. B. and Li, H. G. (1997) Inference for non-random samples (with discussion). *J. R. Statist. Soc.* B, **59**, 55–95.
Das, M., Newey, W. K. and Vella, F. (2003) Nonparametric estimation of sample selection models. *Rev. Econ. Stud.*, **70**, 33–58.
Duncan, G. M. (1987) A simplified approach to M-estimation with application to two-stage estimators. *J. Econmetr.*, **34**, 373–389.
Eicker, F. (1967) Limit theorems for regression with unequal and dependent errors. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability* (eds L. M. LeCam and J. Neyman), pp. 59–82. Berkeley: University of California Press.
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. *J. Computnl Appl. Math.*, **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. *J. Am. Statist. Ass.*, **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: Wiley.
Hampel, F., Rousseeuw, P. J. and Ronchetti, E. (1981) The change-of-variance curve and optimal redescending M-estimators. *J. Am. Statist. Ass.*, **76**, 643–648.
Heckman, J. (1979) Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.
Huber, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability* (eds L. M. LeCam and J. Neyman), pp. 221–233. Berkeley: University of California Press.
Huber, P. J. (1981) *Robust Statistics*. New York: Wiley.
Huber, M. and Melly, B. (2015) A test of the conditional independence assumption in sample selection models. *J. Appl. Econmetr.*, to be published.
Huber, P. J. and Ronchetti, E. (2009) *Robust Statistics*, 2nd edn. New York: Wiley.
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. *J. Am. Statist. Ass.*, **107**, 1546–1557.
Lennox, C. S., Francis, J. R. and Wang, Z. (2012) Selection models in accounting research. *Accountng Rev.*, **87**, 589–616.
Leung, S. F. and Yu, S. (1996) On the choice between sample selection and two-part models. *J. Econmetr.*, **72**, 197–229.
Leung, S. F. and Yu, S. (2000) Collinearity and two-step estimation of sample selection models: problems, origins, and remedies. *Computnl Econ.*, **15**, 173–199.
Ma, Y., Genton, M. G. and Tsiatis, A. A. (2005) Locally efficient semiparametric estimators for generalized skew-elliptical distributions. *J. Am. Statist. Ass.*, **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. *J. Am. Statist. Ass.*, **108**, 1090–1104.
Marchenko, Y. V. and Genton, M. G. (2012) A Heckman selection-*t* model. *J. Am. Statist. Ass.*, **107**, 304–317.
Maronna, R. A., Martin, G. R. and Yohai, V. J. (2006) *Robust Statistics: Theory and Methods*. Chichester: Wiley.
Marra, G. and Radice, R. (2013) Estimation of a regression spline sample selection model. *Computnl Statist. Data Anal.*, **61**, 158–173.
Melino, A. (1982) Testing for sample selection bias. *Rev. Econ. Stud.*, **49**, 151–153.
von Mises, R. (1947) On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.*, **18**, 309–348.
Montes-Rojas, G. V. (2011) Robust misspecification tests for the Heckman's two-step estimator. *Econmetr. Rev.*, **30**, 154–172.
Nelson, F. D. (1984) Efficiency of the two-step estimator for models with endogenous sample selection. *J. Econmetr.*, **24**, 181–196.
Newey, W. K. (2009) Two-step series estimation of sample selection models. *Economtr. J.*, **12**, S217–S229.

Ogundimu, E. O. and Hutton, J. L. (2015) A sample selection model with skew-normal distribution. *Scand. J. Statist.*, to be published, doi 10.1111/sjos.12171.
Paarsch, H. J. (1984) A Monte Carlo comparison of estimators for censored regression models. *J. Econmetr.*, **24**, 197–213.
Peracchi, F. (1990) Bounded-influence estimators for the Tobit model. *J. Econmetr.*, **44**, 107–126.
Peracchi, F. (1991) Robust M-tests. *Econmetr. Theor.*, **7**, 69–84.
Puhani, P. A. (2000) The Heckman correction for sample selection and its critique. *J. Econ. Surv.*, **14**, 53–68.
R Development Core Team (2012) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.
Ronchetti, E. and Trojani, F. (2001) Robust inference with GMM estimators. *J. Econmetr.*, **101**, 37–69.
Salazar, L. (2008) A robustness study of Heckman's model. *Master's Thesis*. University of Geneva, Geneva.
Smith, M. D. (2003) Modelling sample selection using Archimedean copulas. *Econmetr. J.*, **6**, 99–123.
Stolzenberg, R. M. and Relles, D. A. (1997) Tools for intuition about sample selection bias and its correction. *Am. Sociol. Rev.*, **62**, 494–507.
Toomet, O. and Henningsen, A. (2008) Sample selection models in R: package SampleSelection. *J. Statist. Softwr*, **27**, 1–23.
Vella, F. (1992) Simple tests for sample selection bias in censored and discrete choice models. *J. Appl. Econmetr.*, **7**, 413–421.
Vella, F. (1998) Estimating models with sample selection bias: a survey. *J. Hum. Resour.*, **33**, 127–169.
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. *A. Rev. Sociol.*, **18**, 327–350.
Zhelonkin, M. (2013) Robustness in sample selection models. *PhD Thesis*. University of Geneva, Geneva.
Zhelonkin, M., Genton, M. G. and Ronchetti, E. (2012) On the robustness of two-stage estimators. *Statist. Probab. Lett.*, **82**, 726–732.
Zuehlke, T. W. and Zeman, A. R. (1991) A comparison of two-stage estimators of censored regression models. *Rev. Econ. Statist.*, **73**, 185–188.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Robust inference in sample selection models: supplementary material'.