Robustness in sample selection models

Zhelonkin, Mikhail

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES

Robustness in Sample Selection Models

Thèse
présentée à la Faculté des sciences économiques et sociales
de l'Université de Genève
par

Mikhail Zhelonkin

sous la direction de
Prof. Marc G. Genton et
Prof. Elvezio Ronchetti

pour l'obtention du grade de
Docteur ès sciences économiques et sociales
mention statistique

Membres du jury de thèse:

Prof. Marc G. GENTON, King Abdullah University of Science and Technology, Saudi Arabia
Prof. Elvezio RONCHETTI, Université de Genève
Prof. Stephan SPERLICH, Université de Genève
Prof. Francis VELLA, Georgetown University, USA
Prof. Maria-Pia VICTORIA-FESER, Présidente du jury, Université de Genève

Thèse N° 801
Genève, le 17 05 2013

La Faculté des sciences économiques et sociales, sur préavis du jury, a autorisé l'impression de la présente thèse, sans entendre, par là, n'émettre aucune opinion sur les propositions qui s'y trouvent énoncées et qui n'engagent que la responsabilité de leur auteur.

Genève, le 17 05 2013

Le doyen
Bernard MORARD

Impression d'après le manuscrit de l'auteur

AMS 2010 subject classification. Primary 62F35; secondary 62P20.
JEL subject classification. C24; C34; C65; C87.

To my family and friends.

Contents

Résumé

Le problème de la sélection d'échantillons non aléatoires apparait souvent en pratique dans de nombreux domaines différents. En présence de sélection de l'échantillon, les observations apparaissent dans l'échantillon selon une règle de sélection. Par exemple, les personnes choisissent d'entrer sur le marché du travail si leur salaire est supérieur à leur salaire de réserve, c'est à dire que si leur salaire sur le marché est inférieur à une certaine limite, ils ne participent pas au marché du travail et ils sont exclus de l'échantillon. Dans ces cas, les outils standard construits pour les échantillons complets, par exemple les moindres carrés ordinaires, produisent des résultats biaisés, et par conséquent, des méthodes de correction de ce biais sont nécessaires. Dans son ouvrage fondamental, Heckman (1976, 1979) a proposé deux estimateurs pour résoudre ce problème. Ces estimateurs sont devenus l'épine dorsale de l'analyse statistique standard des modèles avec une sélection d'échantillon. Toutefois, ces estimateurs sont basés sur l'hypothèse de normalité et ils sont très sensibles aux deviations par rapport aux hypothèses de la distribution, qui ne sont souvent pas satisfaites en pratique. Dans cette thèse, nous développons un système général pour étudier les propriétés de robustesse des estimateurs et des tests dans les modèles avec une sélection de l'échantillon. Nous utilisons une approche infinitésimale (Hampel et al. 1986), qui nous permet d'explorer le problème de robustesse et de construire des estimateurs et des tests robustes. Nous commençons par l'étude des propriétés de robustesse de la classe générale des estimateurs en deux étapes. Nous dérivons la fonction d'influence, la fonction de changement de la variance et la variance asymptotique d'un M-estimateur général en deux étapes, et fournissons leurs interprétations. Nous illustrons nos résultats dans le cas de l'estimateur maximum de vraisemblance en deux étapes, l'estimateur de "two-stage least squares", et l'estimation des séries chronologiques. En utilisant les résultats généraux pour un M-estimateur en deux étapes, nous dérivons la fonction d'influence et la fonction de changement de la variance pour l'estimateur de Heckman en deux étapes,

et démontrons la non-robustesse de cet estimateur et de sa variance estimée par rapport aux petites déviations du modèle hypothétique. Nous proposons une procédure pour robustifier l'estimateur, prouvons sa normalité asymptotique et calculons sa variance asymptotique. Cela nous permet de construire une alternative simple et robuste pour le test de biais de sélection d'échantillon. Nous illustrons l'utilisation de notre nouvelle méthodologie dans l'analyse des dépences médicales, et comparons les performances des méthodes classiques et robustes dans une étude de simulation de Monte Carlo. De plus, nous étendons nos résultats aux modèles d'équations simultanées avec sélectivité. Nous explorons les propriétés de robustesse de ces modèles, et proposons une alternative robuste aux estimateurs classiques. L'analyse de sensibilité des données sur la force de travail et l'étude de simulation de Monte Carlo démontrent l'utilité de la méthodologie robuste. Pour faciliter l'utilisation des méthodes développées, nous fournissons un package `ssmrob` pour le logiciel statistique `R`.

Abstract

The problem of non-random sample selectivity often occurs in practice in many different fields. In presence of sample selection, the data appears in the sample according to some selection rule. For instance, people choose to enter the labor force if their wage is greater than their reservation wage, i.e. if their wage on the market is below the certain limit, they do not participate in the labor force and are excluded from the sample. In these cases, the standard tools designed for complete samples, e.g. ordinary least squares, produce biased results, and hence, methods correcting this bias are needed. In his seminal work, Heckman (1976, 1979) proposed two estimators to solve this problem. These estimators became the backbone of the standard statistical analysis of sample selection models. However, these estimators are based on the assumption of normality and are very sensitive to small deviations from the distributional assumptions which are often not satisfied in practice. In this thesis we develop a general framework to study the robustness properties of estimators and tests in sample selection models. We use an infinitesimal approach (Hampel et al. 1986), which allows us to explore the robustness issues and to construct robust estimators and tests. We start by investigating the robustness properties of the general class of two-stage estimators. We derive the influence function, the change-of-variance function, and the asymptotic variance of a general two-stage M-estimator, and provide their interpretations. We illustrate our results in the case of the two-stage maximum likelihood estimator, the two-stage least squares estimator, and the estimation of time series. Using the general results for two-stage M-estimators 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. 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. Furthermore, we extend our results to the simultaneous equations models with selectivity. We explore the robustness properties of these models and propose the robust alternative to the classical estimators. The sensitivity analysis of the labor force participation data and the Monte Carlo simulation study demonstrate the usefulness of the robust methodology. To facilitate the use of the developed methods, we provide an `R` package `ssmrob`.

Acknowledgements

I would like to express my gratitude to my thesis advisors, Prof. Marc G. Genton and Prof. Elvezio Ronchetti, for their invaluable guidance, help, and support during this research. This thesis would have been impossible without belief and confidence they have had in me for carrying out this research.

I would also like to thank the other members of my thesis committee, Prof. Stephan Sperlich, Prof. Francis Vella, and Prof. Maria-Pia Victoria-Feser for their questions and useful comments, which allowed to improve the quality of the manuscript and to extend my personal knowledge.

I would like to express my thanks to Prof. Grigory Belyavskiy, who introduced me to the world of probability and statistics and established my interest in research.

My thanks also go to my friends and colleagues from the University of Geneva, William Aeberhard, Marco Avella Medina, Carlos De Porres, Daniel Flores, and Irina Irincheeva for stimulating discussions and their help in various aspects of this work.

Finally, I would like to express my gratitude to my parents, my family, and my friends for their encouragement and support over the years.

Chapter 1

Introduction

Randomization of the sample is one of the crucial principles in statistics. When the observations in the sample are selected in a non random way, the sample selection mechanism can lead to wrong inferences about the underlying process. In economic and social sciences this problem is encountered very often and constitutes an important issue. In 2000 the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel, was awarded to James Joseph Heckman "for his development of theory and methods for analyzing selective samples."

The problem first appeared in the analysis of female labor market behavior and the determinants of wages (Heckman 1974, Gronau 1974). Suppose we are interested to identify the factors influencing the wages of women. If we consider the sample with working women, we exclude those who could work but decided not to do so, because their market wages are lower than their home wages. This truncation will lead to biased inferences. On the other hand, if we consider the random sample both with working and non working women, we will have a subsample of zeroes. The simple ordinary least squares regression on the complete sample will be biased due to this subsample of zeroes. The relation between the decision to work and the wage is the source of the selection bias. In order to obtain unbiased results the information about the decision to work must be included in the structure of the model.

Another example, which we will consider in this thesis, comes from the analysis of the ambulatory expenditures (Cameron and Trivedi 2009). It is natural to expect that the amount of money spent on the medical services is linked with the decision to spend. It means that if we estimate the data with only positive expenditures we will not take into account

the information about zero expenditures, and hence we will obtain biased estimates and wrong inference. On the contrary, if we estimate the regression with all the data, then we will obtain biased results due to the concentration of zeroes. The solution will be to introduce the selection rule, which explains the decision to spend, and to use it to model the amount of expenditures.

In spite of the fact that the sample selection problem first appeared in economics, this methodology is widely used in many other fields, including sociology, finance, political science, and computer science to mention a few. The general review of the problem with discussion of various methodological aspects is given by Vella (1998). For examples in sociology see Berk (1983). A lot of other examples in social sciences can be found in a review by Winship and Mare (1992). Banasik et al. (2003) investigated the sample selection problem in credit scoring models. Collier and Mahoney (1996) discussed the sample selection issues in political science. For the example in machine learning, see Zadrozny (2004). For the review of the problem in criminology see Bushway et al. (2007), and references therein.

The examples discussed above suggest the model with a system of two regressions. The first regression defines the selection mechanism (decision to work or to spend), and the second is the regression of interest (the size of the wage or the amount of expenditures). For the estimation of this model first the joint maximum likelihood estimator was used. Then, Heckman (1976, 1979) proposed an easy-to-implement two-stage correction for the estimation procedure, which treats the selection problem as an omitted variable problem. It consists of estimation of the selection equation by probit maximum likelihood, and the normal linear regression with additional variable correcting the selection bias for the equation of interest. Very often the researcher is also interested in the inference about the selection equation, which makes the two-step procedure very appealing. This method became a standard tool for solutions of such problems due to its simplicity and straightforward interpretation.

Classical statistics relies largely on parametric models, and the sample selection models are not exceptions. It often happens in practice that the assumed distribution does not hold exactly. It can hold for the majority of the observations, with other observations having another pattern, or instead of assumed distribution the data follows another distribution in the neighborhood. Many classical procedures are well-known for not being robust. These procedures are optimal when the assumed model holds

exactly, but they are biased and/or inefficient when even small deviations from the model occur. The statistical results obtained from standard classical procedures on real data applications can therefore be misleading. Robust statistics deals with deviations from the stochastic assumptions and their dangers for classical estimators and tests and develops statistical procedures which are still reliable and reasonably efficient in the presence of such deviations. It can be also viewed as a statistical theory dealing with approximate parametric models by providing a reasonable compromise between the rigidity of a strict parametric approach and the potential difficulties of interpretation of a fully nonparametric analysis.

In the past decades, robust estimators and tests have been developed for large classes of models both in the statistical and econometric literature, and were used in various applications in different fields of science, including biology, computer science, psychology, political science and many others. Standard general books are Huber (1981, 2nd edition by Huber and Ronchetti 2009), Hampel et al. (1986), Maronna et al. (2006), and more recently Heritier et al. (2009), and Jureckova and Picek (2006). For the history of the development of robust statistics see Stigler (1973, 2010).

The classical estimation procedures for sample selection models are based on the maximum likelihood and ordinary least squares. These estimators are well known to be very sensitive to the distributional assumptions (Hampel et al. 1986). The crucial assumption in selection models is the assumption of joint normality of errors. Theoretically it is an approximation of reality, but in practice, the assumption of normality is often violated, and the approximation is very far from the real data. The model has been widely criticized for this sensitivity in literature (see more detailed discussion in Chapter 2), but the general robustness theory for this class of models is still missing.

In this thesis we try to fill this gap. We focus on the robustness analysis of two-stage estimation procedures. Although a robustness investigation of the MLE for this model could be carried out applying standard tools of robust statistics, we feel that it is more important to perform such an analysis for two-stage procedures. In fact the two-stage estimators are structurally simpler, have a straightforward interpretation, and are much less computationally intensive than the joint MLE which explains its success in applied economic analysis. Moreover, there are numerous extensions of the classical Heckman's selection model, including switching regressions (see Chapter 6), simultaneous equations with selectivity (see Chapter 5), and models with self-selectivity, to mention a few, where

the construction of the joint likelihood becomes difficult and cumbersome whereas Heckman-type estimator can be easily computed.

The rest of the document is organized as follows. In Chapter 2 we present the sample selection models. We revise the estimation methods and testing for selectivity bias. Also we point out the robustness problem and show the possible dangers of non-robustness. In Chapter 3 we present the general framework of two-stage estimators, that is needed for the robust analysis of sample selection models. The Heckman's two-stage estimator is a member of this class. Chapter 3 is a more detailed version of Zhelonkin et al. (2012) paper. We derive robustness properties of two-stage M-estimators and illustrate this methodology on three examples. In Chapter 4 we present our new methodology for robust estimation and testing for the classical Heckman selection model (Zhelonkin et al. 2013). We illustrate its use in a simulation study and in the analysis of the ambulatory expenditures data. Chapter 5 presents the extension of our methodology for the simultaneous equations models with selectivity. In Chapter 6 we present the new `R` package `ssmrob` for robust estimation and inference in sample selection models, and show the extension for the case of switching regressions model. Chapter 7 offers some concluding remarks, discussion, and an outlook for future research.

Chapter 2

Sample Selection Models

In this chapter we present the sample selection models. In this field there has been an enormous amount of research in past decades. We only make a brief survey of estimation methods, testing issue, and problems highlighted in literature. In Section 2.1 we present the basic selection problem. Section 2.2 presents an overview of the estimation methods and their critique. Section 2.3 deals with the testing for sample selection bias. Some extensions of the model are given in Section 2.4. The robustness problem is illustrated in Section 2.5.

2.1 Model

The conventional form of a sample selection model can be represented by the following regression system

$$y_{1i}^* = x_{1i}^T \beta_1 + e_{1i}, \tag{2.1}$$
$$y_{2i}^* = x_{2i}^T \beta_2 + e_{2i}, \tag{2.2}$$

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

$$\begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_2 \\ \rho\sigma_2 & \sigma_2^2 \end{pmatrix} \right\}. \tag{2.3}$$

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 (2.1)

is the selection equation, defining the observability rule, and (2.2) is the equation of interest. The observed variables are defined by

$$y_{1i} = \begin{cases} 1, & \text{if } y_{1i}^* > 0, \\ 0, & \text{if } y_{1i}^* \leq 0, \end{cases} \tag{2.4}$$

$$y_{2i} = \begin{cases} y_{2i}^*, & \text{if } y_{1i} = 1, \\ 0, & \text{if } y_{1i} = 0. \end{cases} \tag{2.5}$$

Note, that Heckman (1979) used different indexation, in his original paper, "2" indexed the selection equation and "1" indexed the equation of interest. We define indexes to correspond to the number of the estimation stage.

This model is classified by Amemiya (1984) as a "Tobit type-2 model" or "Heckman model" due to the estimator proposed by Heckman (1976, 1979). In statistics literature the selection models are sometimes referred as models with missing not at random data or nonignorable missing-data models (see Little and Rubin 2002, Ch.11).

If all the data were available, i.e. the regressor matrix was of full rank, 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.

2.2 Estimation

In order to estimate the model (2.1)-(2.5) there are two popular parametric estimation procedures. First solution is the Maximum Likelihood Estimator (MLE) based on joint likelihood function (see Heckman 1974).

Using the assumption of bivariate normality, the likelihood function is given by

$$\begin{aligned} l(\theta|z_i) = & \sum \log \left[\Phi \left\{ \frac{x_{1i}^T \beta_1 + \frac{\rho}{\sigma_2}(y_{2i} - x_{2i}^T \beta_2)}{\sqrt{1-\rho^2}} \right\} \right] \\ + & \sum \log \left\{ \phi \left(\frac{y_{2i} - x_{2i}^T \beta_2}{\sigma_2} \right) \right\} + \sum \log(\Phi(-x_{1i}^T \beta_1)), \end{aligned} \tag{2.6}$$

where θ is the vector of parameters, ϕ and Φ denote the probability density function (pdf) and cumulative distribution function (cdf) respectively.

This estimator is consistent, asymptotically normal and efficient, but it has several drawbacks. First of all it is non-linear and obviously requires iterative numerical methods. It is very expensive from the computational point of view. Of course, given a modern computational power, it is no longer a problem, but except for non-linearity, the likelihood function also has local maxima (Olsen 1982), which requires a good starting point for the numerical algorithm.

The second estimation procedure, and probably the most popular one, is the two-stage procedure proposed by Heckman (1976, 1979). The idea of this method is based on the fact that

$$E(e_{2i}|e_{1i} > -x_{1i}^T\beta_1) = \rho\sigma_2 \frac{\phi(x_{1i}^T\beta_1)}{\Phi(x_{1i}^T\beta_1)}. \tag{2.7}$$

Using (2.7) we can rewrite the equation of interest as:

$$y_{2i} = x_{2i}^T\beta_2 + \lambda_i\beta_\lambda + v_i, \tag{2.8}$$

where $\lambda_i = \phi(x_{1i}^T\beta_1)/\Phi(x_{1i}^T\beta_1)$, $\beta_\lambda = \rho\sigma_2$, and v_i is a zero mean error term. The selection equation can be estimated by probit MLE, then we estimate λ's, and finally we estimate (2.8) by ordinary least squares (OLS) using the observed subsample.

Both methods are criticized in literature, and have advantages and drawbacks which will be discussed below.

The main advantage of the two-stage estimator is its simplicity. It is easy to compute, it does not require any complicated algorithms, and the interpretation of this estimator is straightforward. It can be also used to compute the initial values for the ML estimator. At the same time it has been also criticized for several issues.

The use of the inverse Mills ratio λ can lead to possible problem of multicollinearity. In fact λ is quasi-linear (see Figure 2.1) and can be approximated by the linear function of $x_1^T\beta_1$. If the sets of explanatory variables x_1 and x_2 overlap then the multicollinearity can be encountered. For more insight about this problem see Stolzenberg and Relles (1997). The possible treatment is to introduce the exclusion restrictions, i.e. to make the sets of parameters in two estimation stages different by excluding one or several variables from one of the equations.

Sometimes, the estimated value of ρ lies outside the interval $[-1; 1]$, which happens due to the fact that in the second estimation stage the estimated values of λ are used, and the true value of ρ is close to ± 1

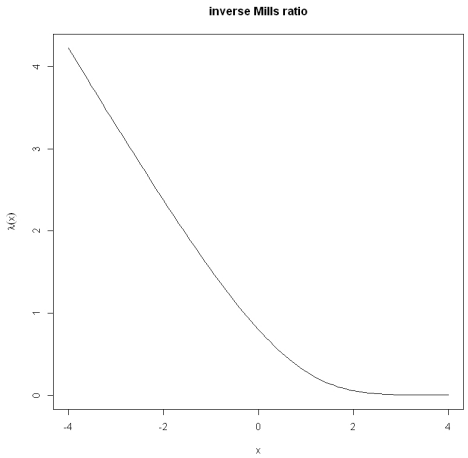

Figure 2.1: *Inverse Mills ratio.*

(Toomet and Henningsen 2008). But the MLE also holds this drawback, because of the convergence problem, as reported in Nawata (1994).

Davidson and MacKinnon (1993) point out the problem of heteroskedasticity of the residuals in the second stage. The OLS covariance matrix is valid only when $\rho = 0$. Otherwise the heteroskedasticity consistent estimator must be used. But this problem was solved in the original paper by Heckman (1979), and the computation of the corrected matrix is not difficult.

But the major direction for critique, as in the case of MLE, was of course the sensitivity to the assumption of normality. The two-stage estimator is a bit less sensitive than the MLE, but it is far from being stable in presence of deviation from the distributional assumptions. Several Monte Carlo studies have investigated the behavior of the estimators under completely different distributional assumptions; see, for instance, Paarsch (1984), Zuehlke and Zeman (1991), and a survey by Puhani (2000). All these studies stated the poor performance of the classical estimators in non-standard conditions.

2.3 Inference

The problem of testing for Sample Selection Bias (SSB) often arises for applied researchers. The question of the presence or absence of the SSB

is important for the choice of the model and therefore for the estimation and further inference. In the case of absence of SSB it is possible to use normal linear regression methodology, while its presence leads to more complicated estimation procedures discussed above.

The simplest and the most widely used test was proposed by Heckman (1979). Essentially, it is a simple regression t-test of $H_0 : \beta_\lambda = 0$. It follows directly from the two-stage estimator. Even if it is argued that the MLE is preferred to the two-stage estimator because of its higher efficiency, the t-test is still suggested to be used for practical purposes (Davidson and MacKinnon 1993). Its popularity raises not only because of its simplicity. Melino (1982) proved that Heckman's test is equivalent to the Lagrange multiplier test, which allows to deduce that it has desirable large sample properties.

Other options, especially if the MLE is used, are the likelihood ratio test of independent equations and the Wald test of $H_0 : \rho = 0$ (or, equivalently, $H_0 : \text{arctanh}\rho = 0$).

2.4 Robustness Problem

The problem of sensitivity of the estimators to the misspecification in the data has arisen many times in literature. The research on alternative estimators is ongoing. Various methods 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), and extension for skew-normal distribution by Ogundimu and Hutton (2012), semiparametric (Ahn and Powell 1993, Newey 2009), and nonparametric (Das et al. 2003) methods. Gallant and Nychka (1987) proposed a semi-nonparametric estimator based on Hermite series. However, the general robustness theory for this class of models is still missing.

In some sources nonparametric methods are classified as robust or together with robust, but these notions have little overlap (for extensive discussion see Huber 1981, page 6). Indeed, the sample mean is a nonparametric estimator of the population mean, but it is not robust, one outlying observation is enough to break it down. The relaxation of the distributional assumptions does not necessarily guarantee the robustness of the method in the statistical sense. The definition of the quantitative robustness is given by Hampel (1971) and it requires continuity with

respect to the topology of weak convergence, or in other words, the estimator must have finite sensitivity to small deviations from the underlying model. The classical nonparametric estimators are not designed to be robust[1] from this point of view. Nevertheless it is not an argument against their use. In spite of the standard difficulties concerned with the nonparametric estimators, like hard interpretability, often complicated bandwidth selection, and curse of dimensionality, in situations when we are completely uncertain about the undelying distribution the use of nonparametric methods is preferable, and these estimators outperform the classical and robust estimators. However, if there is evidence to use the parametric model, then the robust estimators provide insurance and protection against distributional deviations and allow to benefit from the parametric structure, e.g. computational simplicity and interpretability.

The goal of robust statistics in general and this thesis in particular is to develop statistical procedures, which provide reliable results not only at the model, but also in the neighborhood of it. We assume that the data generating process lies in the neighborhood[2] of some parametric model F_θ (it encompasses both the structural and stochastic parts). There are several ways to formalize the concept of neighborhood (see Huber 1981, Chapter 2), but probably the most convenient way to do it is the so-called Gross Error Model:

$$F_\epsilon = (1-\epsilon)F_\theta + \epsilon G.$$

The model parameters are contained in vector θ, which contains both structural and stochastic parameters. Distribution G is some arbitrary distribution (an important particular case is when G is a point mass), and $0 \leq \epsilon \leq 1$. Usually we expect ϵ to be between 0 and 0.5, which defines the contamination, i.e. proportion of the data from the arbitrary distribution G. If one truly believes to know the exact composition of F_ϵ, he can use the classical methods if $\epsilon = 0$, or to use mixtures if $\epsilon \geq 0$. In the latter case one must adequately choose the distribution G. In general, practitioner cannot know exactly the data generating process, and it is hard or even impossible to define G. Our strategy is to develop methods which would be reliable for the majority of observations generated from F_θ and non-sensitive to the contamination G.

In order to illustrate the robustness problem in selection models we use two examples. First example is the mixture of distributions and the

[1] Sometimes the term "resistant" is used.

[2] Note, that the word "neighborhood" is not exactly a neighborhood in the topological sense, the idea is to assume the set of distributions around the true distribution.

second example is the sensitivity to outliers. Let us first explore the mixture.

Consider the case, when the error term follows a contaminated normal distribution:

$$\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \sim (1-\epsilon)N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix};\ \Sigma_1 \right\} + \epsilon N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix};\ \Sigma_2 \right\}, \quad (2.9)$$

where

$$\Sigma_1 = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} \nu_1^2 & \tau\nu_1\nu_2 \\ \tau\nu_1\nu_2 & \nu_2^2 \end{pmatrix}. \quad (2.10)$$

We can explicitly compute the conditional expectation of $E(e_1|e_2 > -x_1^T\beta_1)$, which characterizes the selection bias, and allows to see its behavior in presence of contaminated observations. It is given by

$$\begin{aligned} E(e_1|e_2 > -d) &= \left\{(1-\epsilon)\Phi(x_1^T\beta_1) + \epsilon\Phi\left(\tfrac{x_1^T\beta_1}{\nu_2}\right)\right\}^{-1} \\ &\quad \times \left\{(1-\epsilon)\rho\sigma_1\phi(x_1^T\beta_1) + \epsilon\tau\nu_1\phi\left(\tfrac{x_1^T\beta_1}{\nu_2}\right)\right\}. \end{aligned} \quad (2.11)$$

Note that if $\epsilon = 0$ in (2.11), the usual selection bias under the normal distribution is recovered. The derivation is given in the Appendix B.

Figure 2.2 depicts various plots of (2.11) as a function of linear predictor $x_1^T\beta_1$ with $\rho = 0.5$ and $\sigma_1 = 1$. In the top left panel, we set $\nu_1 = \nu_2 = 1$ and let ϵ vary. This means that a fraction ϵ of points come from an uncorrelated bivariate normal distribution, hence the selection bias decreases when ϵ increases, as one would expect. In the top right panel, we set $\epsilon = 0.1$, $\nu_2 = 1$, and let ν_1 increase. The effect is that the selection bias increases too. In the bottom left panel, we set $\epsilon = 0.1$, $\nu_1 = 1$, and let ν_2 increase. The effect is that the selection bias is reduced. This will certainly have an impact on the estimators and tests for selection bias. In our framework, when the variance of the error term in the selection equation is equal to 1, this contamination corresponds to the increased probability of leverage outliers. If the variability of errors increases, then for large negative x_1 we can obtain a large positive error, which will set the corresponding $y_1 = 1$, generating an outlier. The bottom right panel combines the two previous cases.

This example is purely illustrative. It would be very naive to expect such a simple contaminating distribution, and even if we exactly know that there is a mixture of distributions, we cannot be sure that it is not a

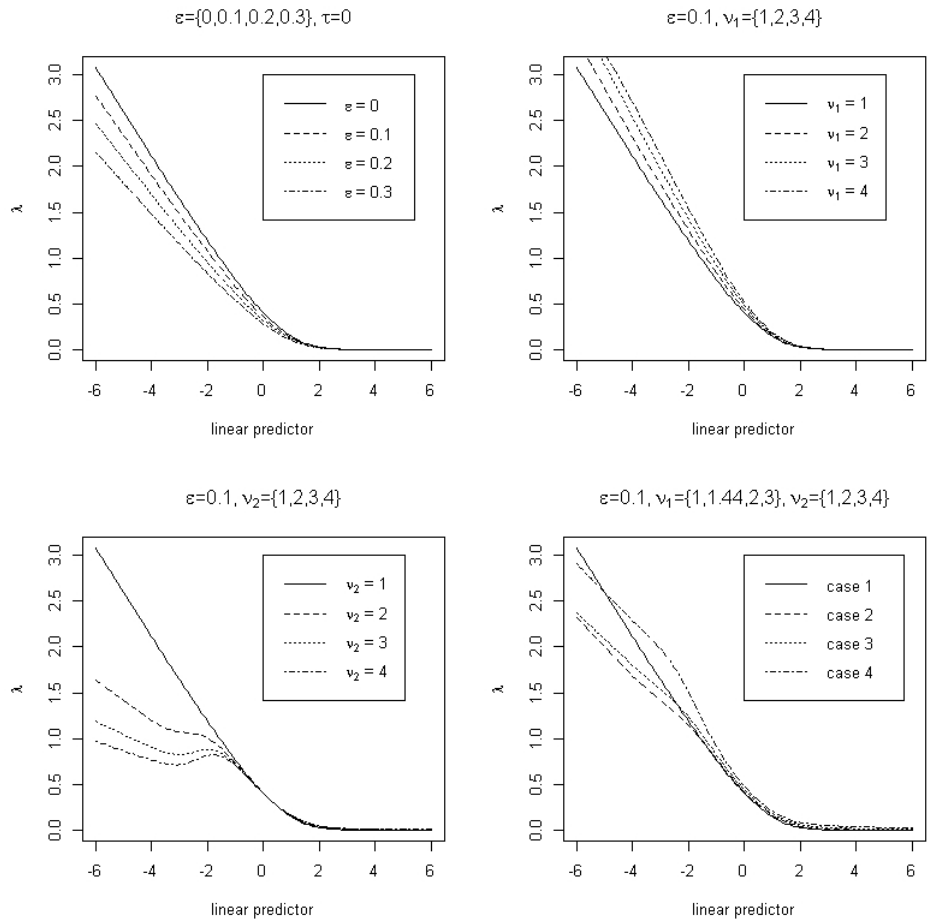

Figure 2.2: *Selection bias under contaminated normal distribution.*

mixture with some arbitrary distribution in the neighborhood, and not a normal with different variance. Of course in real data there can be much more complicated situations, and the consequences for the estimation and testing can be more dramatic.

In the second example we investigate the sensitivity of the classical sample selection bias (SSB) test (Heckman 1979). We carry out a simple Monte Carlo simulation study. The explanatory variables x_1 and x_2 for the selection and outcome equations respectively are generated independently from the standard normal distribution. The error terms are generated from the bivariate normal with zero means, correlation $\rho = 0.75$,

and $\sigma_2 = 1$. In order to examine the sensitivity of the test we move a single observation x_{11} from 0 (the mean value under the normal model) to -6 in a sample of size 200.

The boxplots of the test statistic, the $\log_{10}(p\text{-values})$ and the empirical power of the test are shown in Figure 2.3. As x_{11} moves away from 0, the test statistic decreases and eventually becomes smaller than the threshold of 1.96 reversing the test decision. As a corollary, we see a substantial increase of the p-values and a drastic reduction of the empirical power of the test. Therefore, a single observation out of 200 can change completely the decision about the presence of sample selection bias.

2.5 Extensions and Related Models

The class of sample selection models is very broad. The basic sample selection model was extended in a variety of directions. Here we review the most important contributions.

The formulation (2.1)-(2.5) gives a convenient framework with clear structure of the model, which can be simplified or complicated if necessary. To some extent, the sample selection model is an extension of the famous Tobit model (Tobin 1958), or rather Tobit model is a particular case of sample selection models. It is defined by

$$y_i^* = x_i^T\beta + e_i, \tag{2.12}$$

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* > 0, \\ 0, & \text{if } y_i^* \leq 0, \end{cases} \tag{2.13}$$

where y_i^* is an unobserved response variable, y_i is an observed response, x_i is a vector of explanatory variables, and e_i is an error term. Equation (2.13) defines the censoring rule (the model is also known as censored regression model). The connection between Tobit model (2.12)-(2.13) and the Heckman's selection model (2.1)-(2.5) is obvious. We only need to replace $y_{1i} = 1$ and $y_{1i} = 0$ in (2.5) by $y_i^* > 0$ and $y_i^* \leq 0$ respectively. The model is usually estimated by MLE, but the Heckman's two-stage estimator is also applicable. The robustness problem for this model has been studied by Peracchi (1990), where he proposed a robustified version of the MLE.

A natural extension of the basic selection model is to consider the model with two regimes, i.e. instead of truncation and non-observed data we have the data following another regime. It is classified by Amemiya

(1984) as a "Tobit type-5 model". There are numerous applications, see Greene (2008), Amemiya (1984), and references therein. We explore the robustness issues for this model and propose a robust estimation procedure in Chapter 6.

Another important issue is the possible endogeneity of regressors. The simplest example is the estimation of supply and demand, which influence one another. The problem is usually treated using the simultaneous equations models; see any general econometrics textbook, e.g. Davidson and MacKinnon (1993), Greene (2008), Wooldridge (2002). The models with non-random sample selection often have endogenous regressors. The issue has been studied extensively in literature both in methodology and in applications; see Lee et al. (1980) and Maddala (1983). We discuss the robustness issues for the models with endogeneity and selectivity in Chapter 5.

The model given by (2.1)-(2.5) uses the selection rule of probit type. In some cases, for instance Kenny et al. (1979), the selection rule of Tobit type is used. The difference between probit and Tobit selection rules is that in case of Tobit the variable in selection equation is truncated and not binary, i.e. y_{1i} in (2.4) is given by $y_{1i} = I(y_{1i}^* > 0)y_{1i}^*$, where I is the indicator function. This model can be estimated by the two-stage estimator, and our new methodology of robust estimation is applicable in this case too.

In this thesis we study cross-sectional data, but of course panel data often have the selectivity issue too. Several estimation procedures have been proposed by Wooldridge (1995) and Kiriazidou (1997). The problem of sample selectivity together with endogeneity was studied by Vella and Verbeek (1999) and Semykina and Wooldridge (2010). Of course the robustness issues in this context are also of interest, and leaving it beyond the scope of this thesis, we expect to explore it in future research.

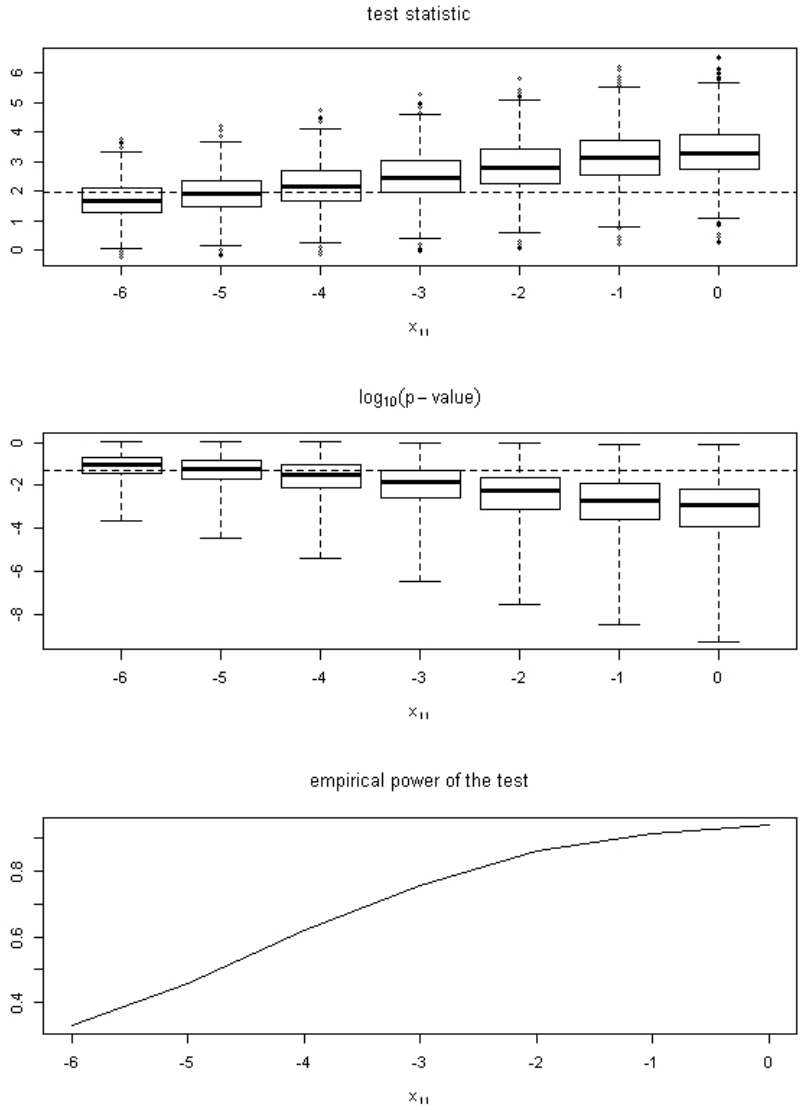

Figure 2.3: *Influence of a single observation* x_{11} *on the sample selection bias (SSB) test. The data generating process is described in Section 4.1 with the only difference that here* $\rho = 0.75$*. The sample size is* 200 *and we generate* 1000 *replicates. The top panel represents the SSB test statistic, the middle panel its* $\log_{10}(p\text{-values})$*, and the bottom panel represents its empirical power, i.e. the fraction of rejections of the null hypothesis among the* 1000 *replicates. On the horizontal axis,* x_{11} *varies between* 0 *and* -6 *and the corresponding* $y_{11} = 1$*. The dashed line in the top panel corresponds to* 1.96*, and in the middle panel to* $\log_{10}(0.05)$*.*

Chapter 3

Robustness Properties of Two-stage M-Estimators

Many estimators in the statistics and econometrics literature are obtained following a two-stage procedure. Typically, the first stage is preliminary and provides the necessary input for the second stage, which is of main interest. Sometimes, the first stage is also of interest, as in the case, for instance, of time series where the trend and seasonality are removed in a first stage, and similarly in spatial statistics; see Genton (2001). Several papers in the literature discuss various statistical properties of two-stage estimators; see for instance Murphy and Topel (1985), Pagan (1986), and references therein. They mostly focus on two-stage Maximum Likelihood Estimators (MLE) or Least Squares Estimators (LSE) in linear models. It is well known that classical MLE and LSE are very sensitive to deviations from the underlying stochastic assumptions of the model or to outliers in the data. These deviations may lead to biased estimators and incorrect inference.

In the existing literature some authors have proposed robust versions of specific two-stage estimators. Kim and Muller (2007) proposed a two-stage Huber version of two-stage least squares whereas Cohen Freue et al. (2011) derived robust estimators with instrumental variables. Moreover, Hardin (2002) derived a robust variance estimator for two-stage models and Yeap and Davidian (2001) proposed a robust two-stage procedure for hierarchical nonlinear models. Finally, Dollinger and Staudte (1991) computed the influence function for the case of iteratively reweighted least squares estimators and Jorgensen (1993) investigated the influence functions of iteratively defined statistics. In spite of these developments,

a general framework to analyze the robustness properties of two-stage procedures is still missing.

In this chapter we present such a general framework based on M-estimators. It has the advantage to include most of the two-stage estimators available in the literature, to indicate a general way to robustify two-stage estimators, and to clarify the structure of their asymptotic variance. Although we focus on two-stage estimators, our results can be easily extended to multi-stage procedures.

This chapter is structured as follows. In Section 3.1 we present the two-stage M-estimation framework. In Section 3.2 we derive the influence function. In Section 3.3 we show the connection between the influence function and the asymptotic variance. The derivation of the change-of-variance function for the two-stage M-estimator is given in Section 3.4. Section 3.5 provides some specific examples of applications. Section 3.6 offers the extension for multi-stage estimation procedures.

3.1 Framework

To analyze the robustness properties of two-stage estimators, we consider the class of two-stage M-estimators. This class is general enough to cover the vast majority of classical estimators used in statistics and econometrics and it provides a convenient framework to develop robust versions of two-stage estimators.

Let F_N be the empirical distribution function putting mass $1/N$ at each observation $z_i = (z_i^{(1)}, z_i^{(2)})$, where $z_i^{(j)} = (x_{ji}, y_{ji})$, $j = 1, 2$, $i = 1, \ldots, N$, and let F be the distribution function of z_i. Also, let $\beta = (\beta_1, \beta_2)$ be a vector defining the parameters of the first and second stage, respectively.

Consider the following system of equations:

$$E_F\Big[\Psi_1(z^{(1)}; S(F))\Big] = 0, \tag{3.1}$$

$$E_F\Big[\Psi_2(z^{(2)}; h(z^{(1)}; S(F)), T(F))\Big] = 0, \tag{3.2}$$

where $\Psi_1(\cdot;\cdot)$ and $\Psi_2(\cdot;\cdot,\cdot)$ denote the score functions of the first and second stage estimators respectively, $h(\cdot;\cdot)$ is a given continuously piecewise differentiable function in the second variable. Here S is the functional for the parameters of the first stage, such that $S(F_N) = \hat{\beta}_1$ and at the model $S(F) = \beta_1$, while T is the functional for the second stage, such that

$T(F_N) = \hat{\beta}_2$ and at the model $T(F) = \beta_2$. Here $T(F)$ depends directly on F and indirectly on F through $S(F)$. Notice that we do not put any restrictions on the presence or absence of one or several components of the unit z.

3.2 Influence Function

For a given functional $T(F)$, the influence function (IF) is defined by Hampel (1974) as $IF(z; T, F) = \lim_{\epsilon \to 0} [T(F_\epsilon) - T(F)]/\epsilon$, where $F_\epsilon = (1-\epsilon)F + \epsilon\Delta_z$ and Δ_z is the probability measure which puts mass 1 at the point z. It describes the standardized asymptotic bias on the estimator due to a small amount of contamination ϵ at the point z. An estimator is considered to be robust if small departures from the assumed distribution have only small effects on the estimator. Therefore, a condition for (infinitesimal) robustness is a bounded IF with respect to z. In our case F_ϵ 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.

From (3.2), the functional $T(F_\epsilon)$ is defined by:

$$\int \Psi_2(z^{(2)}; h(z^{(1)}; S(F_\epsilon)), T(F_\epsilon)) dF_\epsilon = 0 \tag{3.3}$$

and the derivative of (3.3) with respect to ϵ evaluated at $\epsilon = 0$ is

$$\begin{aligned} \frac{\partial}{\partial \epsilon}(1-\epsilon) \int \Psi_2(\tilde{z}^{(2)}; h(\tilde{z}^{(1)}; S(F_\epsilon)), T(F_\epsilon)) dF(\tilde{z}) \Big|_{\epsilon=0} \\ + \frac{\partial}{\partial \epsilon}\epsilon \int \Psi_2(\tilde{z}^{(2)}; h(\tilde{z}^{(1)}; S(F_\epsilon)), T(F_\epsilon)) d\Delta_z \Big|_{\epsilon=0} \quad = \quad 0. \end{aligned} \tag{3.4}$$

The second term of (3.4) is given by

$$\frac{\partial}{\partial \epsilon}\epsilon \int \Psi_2(\tilde{z}^{(2)}; h(\tilde{z}^{(1)}; S(F_\epsilon)), T(F_\epsilon)) d\Delta_z \Big|_{\epsilon=0} = \Psi_2(z^{(2)}; h(z^{(1)}; S(F)), T(F)),$$

and the first term by

$$\begin{aligned}
& \frac{\partial}{\partial\epsilon}(1-\epsilon)\int \Psi_2(\tilde{z}^{(2)}; h(\tilde{z}^{(1)}; S(F_\epsilon)), T(F_\epsilon))dF(\tilde{z})\Big|_{\epsilon=0} \\
= & \frac{\partial}{\partial\epsilon}\int \Psi_2(\tilde{z}^{(2)}; h(\tilde{z}^{(1)}; S(F_\epsilon)), T(F_\epsilon))dF(\tilde{z})\Big|_{\epsilon=0} \\
= & \int \frac{\partial}{\partial\theta}\Psi_2(\tilde{z}^{(2)}; \theta, T(F))\frac{\partial}{\partial\eta}h(\tilde{z}^{(1)}; \eta)dF(\tilde{z})\frac{\partial}{\partial\epsilon}S(F_\epsilon)\Big|_{\epsilon=0} \\
& + \int \frac{\partial}{\partial\xi}\Psi_2(\tilde{z}^{(2)}; h(\tilde{z}^{(1)}; S(F)), \xi)dF(\tilde{z}) \cdot IF(z; T, F),
\end{aligned}$$

where the derivative with respect to θ is evaluated at $\theta = h(\tilde{z}^{(1)}; S(F))$, the derivative with respect to η is evaluated at $\eta = S(F)$, the derivative with respect to ξ is evaluated at $\xi = T(F)$, and the derivative of S with respect to ϵ is the influence function of the estimator of the first stage, i.e. $\frac{\partial}{\partial\epsilon}S(F_\epsilon)|_{\epsilon=0} = IF(z; S, F)$.

Combining the derivatives of the two terms of (3.4), we obtain the IF of the two-stage M-estimator:

$$\begin{aligned}
IF(z; T, F) = M^{-1}\Bigg(& \Psi_2(z^{(2)}; h(z^{(1)}; S(F)), T(F)) \\
& + \int \frac{\partial}{\partial\theta}\Psi_2(\tilde{z}^{(2)}; \theta, T(F))\frac{\partial}{\partial\eta}h(\tilde{z}^{(1)}; \eta)dF(\tilde{z}) \cdot IF(z; S, F)\Bigg),
\end{aligned} \tag{3.5}$$

where $M = -\int \frac{\partial}{\partial\xi}\Psi_2(\tilde{z}^{(2)}; h(\tilde{z}^{(1)}; S(F)), \xi)dF(\tilde{z})$.

Here are some remarks on the IF obtained in (A.6) and its sources of unboundedness.

[i] If x_1 and y_1 are not contaminated, i.e. the distribution of $z^{(1)}$ is the marginal $F^{(1)}$ of F, then $IF(z; S, F)$ drops out and the IF of the estimator of the second stage collapses to $IF((x_2, y_2); T, F)$, which implies that the robustness properties of the estimator are determined just by the boundedness of the score function of the second stage;

[ii] If $h(\cdot;\cdot)$ does not appear in (3.2), then the IF of the two-stage estimator is equal to the IF of the one-stage estimator, because $\frac{\partial}{\partial\theta}\Psi_2(z^{(2)}; \theta, T(F)) = 0$;

[iii] Robust estimators are obtained by bounding the IFs at both stages. If the score function of the first stage is unbounded, the final estimator is non-robust. Of course, if the score function of the second stage is unbounded, the final estimator is also non-robust.

Depending on the location of the contamination (1st, 2nd or both stages), a robust estimation procedure can be proposed. We suggest two different approaches. The first is to ensure robustness by bounding the IFs of both stages. All the terms in (A.6) except the score function of the second stage and IF of the first stage are constants. Hence, we need to have bounded score functions on both stages to produce a bounded-influence two-stage estimator. The contamination can also emerge in only one of the stages and in this case there is no need to use robust estimators in both stages.

When y_1 and/or x_1 are contaminated, the second approach uses the robust estimator in the first stage and computes robustly $h(\cdot;\cdot)$. In the second stage using the property [i], we are in the situation of classical one-stage M-estimation.

3.3 Asymptotic Variance

Using the result in (Hampel et al., 1986, p. 85), we can derive the expression of the asymptotic variance. For the one-stage estimator we have

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

Denote the components of the IF as follows:

$$\begin{aligned} a(z) &= \Psi_2(z^{(2)}; h(z^{(1)}; S(F)), T(F)), \\ b(z) &= \int \frac{\partial}{\partial\theta}\Psi_2(\tilde{z}^{(2)}; \theta, T(F)) \frac{\partial}{\partial\eta} h(\tilde{z}^{(1)}; \eta) dF(\tilde{z}) \cdot IF(z; S, F). \end{aligned}$$

Using the expression of IF in (A.6) and integrating, we obtain the asymptotic variance of $\hat{\beta}_2$:

$$\begin{aligned} V(T,F) &= M^{-1} \int \big(a(z)a(z)^\top + a(z)b(z)^\top \\ &\quad + b(z)a(z)^\top + b(z)b(z)^\top\big)\, dF(z)\, M^{-1}. \end{aligned} \tag{3.6}$$

The form (3.6) is general for any two-stage M-estimator. In particular this expression of the asymptotic variance is the generalization of the

result in Murphy and Topel (1985). Then, specifying the vectors $a(z)$ and $b(z)$, we can derive the asymptotic variances for the particular cases. Given particular score functions and $h(\cdot;\cdot)$ functions, we can obtain the asymptotic variance for any M-estimator. If we assume the function $h(\cdot;\cdot)$ to be linear, then our result matches the result of Newey (1984) for the fully identified case. If $h(\cdot;\cdot)$ does not depend on the first stage equation (for instance it is fixed) then all the vectors $b(z)$ become equal to zero, and (3.6) collapses to the asymptotic variance of the one-stage M-estimator. In the cases when the error terms are independent the $\int a(z)b(z)^\top dF(z)$ and $\int b(z)a(z)^\top dF(z)$ are equal to zero.

3.4 Change-of-Variance Function

The change-of-variance function (CVF) of an M-estimator T at the model distribution F is defined by the matrix $CVF(z;T,F) = \Big[(\partial/\partial\epsilon)V(T,(1-\epsilon)F+\epsilon\Delta_z)\Big]_{\epsilon=0}$, for all z where this expression exists; see Hampel et al. (1981). It reflects the influence of a small amount of contamination on the variance of the estimator, and hence on the length of the confidence intervals.

For the case of a two-stage M-estimator the CVF has the following form:

$$\begin{aligned}
CVF(z;S,T,F) =& V(T,F) - M^{-1}\left[\int D^{(2S)}dF(z)\right]V(T,F) \\
&- M^{-1}\left[\frac{\partial}{\partial\theta}\Psi_2(z^{(2)};h(z^{(1)};S(F)),\theta)\right]V(T,F) \\
&+ M^{-1}\int\left(Aa(z)^\top + Ba(z)^\top + Ab(z)^\top + Bb(z)^\top\right)dF(z)M^{-1} \\
&+ M^{-1}\int\left(a(z)A^\top + b(z)A^\top + a(z)B^\top + b(z)B^\top\right)dF(z)M^{-1} \\
&+ M^{-1}\left(a(z)a(z)^\top + a(z)b(z)^\top + b(z)a(z)^\top + b(z)b(z)^\top\right)M^{-1} \\
&- V(T,F)\left[\int D^{(2S)}dF(z) + \frac{\partial}{\partial\theta}\Psi_2(z^{(2)};h(z^{(1)};S(F)),\theta)\right]M^{-1},
\end{aligned} \tag{3.7}$$

where $D^{(2S)}$ is a matrix with elements

$$\begin{aligned} D_{ij}^{(2S)} &= \left(\frac{\partial}{\partial h}\frac{\partial \Psi_{2i}(z^{(2)};h,\theta)}{\partial \theta_j}\right)^{\top}\frac{\partial h(z^{(1)};s)}{\partial s}IF(z;S,F) \\ &\quad + \left(\frac{\partial}{\partial \theta}\frac{\partial \Psi_{2i}(z^{(2)};h,\theta)}{\partial \theta_j}\right)^{\top} IF(z,T,F). \end{aligned}$$

Matrix A is given by

$$\begin{aligned} A &= \frac{\partial}{\partial h}\Psi_2(z^{(2)};h,T(F))\frac{\partial}{\partial s}h(z^{(1)};s)IF(z,S,F) \\ &\quad + \frac{\partial}{\partial \theta}\Psi_2(z^{(2)};h(z^{(1)};S(F)),\theta)\cdot IF(z;T,F). \end{aligned}$$

The matrix B has the following form

$$\begin{aligned} B = &\int R_1\frac{\partial}{\partial s}h(z^{(1)};s)dFIF(z,S,F) + \int \frac{\partial}{\partial h}\Psi_2(z^{(2)};h,T(F))R_2 dFIF(z,S,F) \\ &- \int \frac{\partial}{\partial h}\Psi_2(z^{(2)};h,T(F))\frac{\partial}{\partial s}h(z^{(1)};s)dFM_1^{-1}\int D^{(1)}dFIF(z,S,F) \\ &- \int \frac{\partial}{\partial h}\Psi_2(z^{(2)};h,T(F))\frac{\partial}{\partial s}h(z^{(1)};s)dFM_1^{-1}\frac{\partial}{\partial \theta}\Psi_1(z^{(1)};\theta)IF(z,S,F) \\ &+ \int \frac{\partial}{\partial h}\Psi_2(z^{(2)};h,T(F))\frac{\partial}{\partial s}h(z^{(1)};s)dFM_1^{-1}\frac{\partial}{\partial \theta}\Psi_1(z^{(1)};\theta)IF(z,S,F) \\ &+ \frac{\partial}{\partial h}\Psi_2(z^{(2)};h,T(F))\frac{\partial}{\partial s}h(z^{(1)};s)\cdot IF(z;S,F), \end{aligned}$$

where $D^{(1)}$ denotes the matrix with elements

$$D_{ij}^{(1)} = \left(\frac{\partial}{\partial \theta}\frac{\partial \Psi_{1i}(z^{(1)};\theta)}{\partial \theta_j}\right)^{T}\cdot IF(z;S,F), \tag{3.8}$$

$R^{(1)}$ is the matrix with elements

$$\begin{aligned} R_{ij}^{(1)} = &\left(\frac{\partial}{\partial h}\frac{\partial \Psi_{2i}(z^{(2)};h,T(F))}{\partial h_j}\right)^{\top}\frac{\partial}{\partial s}h(z^{(1)};s)IF(z;S,F) \\ + &\left(\frac{\partial}{\partial \theta}\frac{\partial \Psi_{2i}(z^{(2)};h,\theta)}{\partial h_j}\right)^{\top} IF(z;T,F), \end{aligned}$$

$R^{(2)}$ is the matrix with elements $R_{ij}^{(2)} = \left(\frac{\partial}{\partial s}\frac{\partial h_i(z^{(1)};s)}{\partial s_j}\right)^{\top} IF(z;S,F)$, and M_1 denotes the M matrix of the first stage. The derivation of the CVF

function is similar to the derivation of the IF. The detailed computations are given in the appendix.

Analogously to the properties of the IF of a two-stage M-estimator, in case that the second stage estimator does not depend on $h(\cdot;\cdot)$, the CVF of the two-stage estimator collapses to the CVF of one-stage M-estimator. The same happens if there is no contamination on the first stage, i.e. if $z^{(1)} \sim F$. The CVF of the one-stage M-estimator has been recently studied by Ferrari and La Vecchia (2012). The boundedness of the CVF function is determined by the boundedness of the IF's and the derivatives of Ψ-functions.

3.5 Examples

3.5.1 Two-Stage Maximum Likelihood

Equation (3.6) gives the general form of the asymptotic variance. We can use it to obtain the expression of the variance for the two-stage MLE derived in the paper Murphy and Topel (1985) and generalized by Hardin (2002). Recall that

$$\begin{aligned} \Psi_1(z^{(1)}; S(F)) &= \frac{\partial \log f_1}{\partial \beta_1}, \\ \Psi_2(z^{(2)}; h(z^{(1)}; S(F)), T(F)) &= \frac{\partial \log f_2}{\partial \beta_2}, \end{aligned}$$

where f_1, f_2 are the probability densities and β_1, β_2 are the parameter vectors of the first and second stages, respectively. If we use these expressions in (3.6) then we immediately obtain the result in Murphy and Topel (1985).

3.5.2 Two-Stage Least Squares

The Two-Stage Least Squares (2SLS) is an important method of estimation in the case when the exogenous variables are correlated with the error term. Consider the simplest case

$$y = x^\top \beta + u,$$

where x is a $p \times 1$ vector consisting of p_1 exogenous variables $x^{(1)}$ and for simplicity of notation one endogenous $x^{(2)}$ such that $x^\top = (x^{(1)\top}, x^{(2)})$.

The general case for multivariate $x^{(2)}$ is treated in appendix. We assume $\mathrm{cov}(x^{(2)}, u) \neq 0$ and $\mathrm{cov}(x_j^{(1)}, u) = 0$ for all j. In this case the ordinary least squares (OLS) estimator is biased due to the endogeneity of $x^{(2)}$. To find an unbiased estimator we need first to regress $x^{(2)}$ on w, which is the vector of instrumental exogenous variables such that it is correlated with $x^{(2)}$ but uncorrelated with u, i.e. we have the first stage regression $x^{(2)} = w^\top \alpha + u_2$, where u_2 is the error term of the auxiliary regression. In this case y, $x^{(1)}$, $x^{(2)}$, w correspond to y_2, x_2, y_1, x_1 from (3.1)-(3.2), respectively, and $z^{(2)} = (x^{(1)}, y)$ and $z^{(1)} = (w, x^{(2)})$. Here $h(\cdot;\cdot)$ is linear. The functional form of $\hat{\alpha}$ is $(\int ww^\top dF)^{-1} \int wx^{(2)} dF$, where F is the distribution function of the statistical unit $z = (x^{(1)}, y, w, x^{(2)})$. Then we replace $x^{(2)}$ by its estimate $\hat{x}^{(2)} = w^\top \hat{\alpha}$ and regress y on $x^{(1)}$ and $\hat{x}^{(2)}$.

The score functions are equal to:

$$\begin{aligned} \Psi_1((w, x^{(2)}); S(F)) &= \left(x^{(2)} - w^\top \alpha\right) w \\ \Psi_2((y, x^{(1)}); w^\top \alpha, T(F)) &= \left(y - (x^{(1)})^\top \beta_1 - w^\top \alpha \beta_2\right) \begin{pmatrix} x^{(1)} \\ w^\top \alpha \end{pmatrix}. \end{aligned}$$

Using the general formula (A.6) we compute the IF for 2SLS as a special case:

$$\begin{aligned} M &= -\int \frac{\partial}{\partial \xi} \Psi_2((y, x^{(1)}); w^\top \alpha, \xi) dF(z) \\ &= \int \begin{pmatrix} x^{(1)} \\ w^\top \alpha \end{pmatrix} \begin{pmatrix} (x^{(1)})^\top & w^\top \alpha \end{pmatrix} dF(z). \end{aligned}$$

The derivative of $\Psi_2(\cdot;\cdot,\cdot)$ with respect to $h(\cdot;\cdot)$, which is the linear predictor from the first equation, is:

$$\begin{aligned} \frac{\partial}{\partial \theta} \Psi_2((y, x^{(1)}); \theta, T(F)) &= \frac{\partial}{\partial w^\top \alpha} \Psi_2((y, x^{(1)}); w^\top \alpha, T(F)) \\ &= \begin{pmatrix} -x^{(1)} \beta_2 \\ y - (x^{(1)})^\top \beta_1 - 2w^\top \alpha \beta_2 \end{pmatrix}. \end{aligned}$$

Combining the formulas above we find

$$\begin{aligned} IF(z; T, F) = M^{-1} \Bigg\{ & \left(y - (x^{(1)})^\top \beta_1 - w^\top \alpha \beta_2\right) \begin{pmatrix} x^{(1)} \\ w^\top \alpha \end{pmatrix} \\ & - \left(\int \begin{pmatrix} x^{(1)} \beta_2 \\ w^\top \alpha \beta_2 \end{pmatrix} w^\top dF(z) \right) \cdot IF(z; S, F) \Bigg\}, \end{aligned} \tag{3.9}$$

where

$$IF(z; S, F) = \left(\int ww^\top dF(z) \right)^{-1} (x^{(2)} - w^\top \alpha)w.$$

The IF function of the classical 2SLS estimator is unbounded in any component of z, which means that a deviation from the assumed model can bias the estimator. We illustrate this fact by a simulation study provided in the next section. Also note that from (3.9) we can obtain the asymptotic variance of the 2SLS estimator using formula (3.6).

Simulations

We illustrate the robustness issues in this model via Monte Carlo simulations. In our experiment, for simplicity of exposition, we omit $x^{(1)}$ and have $u \sim N(0,1)$, $x^{(2)} \sim N(0,1)$, $\text{corr}(x^{(2)}, u) = -0.6$, $\beta_2 = 1$, and an intercept $\beta_0 = 0$. There exists one instrumental variable w, such that $\text{corr}(x^{(2)}, w) = 0.6$ and $\text{corr}(w, u) = 0$. We find the 2SLS estimate of α without contamination and with two types of contamination. In the first scenario we contaminate $x^{(2)}$. We generate observations from the model described above and replace them with probability $\epsilon = 0.01$ from the degenerate distribution putting mass 1 at the point $(-1, -1, 9)$, corresponding to $(y, w, x^{(2)})$. In the second scenario we contaminate w, using the same idea as with $x^{(2)}$, but the degenerate distribution is now equal to the constant vector $(0, 5, -2)$. Both types of contaminations generate outliers only in one of four dimensions, either in $x^{(2)}$ or in w. Two other coordinates belong to the bulk of the data while $x^{(1)}$ is omitted. The sample size is $N = 200$, and we repeated the experiment 200 times. The values of average bias, variance, and Mean Square Error (MSE) presented in Table 3.1 confirm the theoretical results derived above. Even under a relatively weak contamination the estimates are seriously biased. Also note that the variances of the parameters under contamination increase. It can be explained by the fact that the CVF in (A.32) depends on the IF of the 2SLS estimator and is unbounded. The unshaded boxplots in Figure 3.1 correspond to the classical 2SLS estimator. Three types of contamination are denoted by (a), (b), and (c), which correspond to the non-contaminated case, the contamination of w, and the contamination of $x^{(2)}$, respectively.

We did not consider the case when y is contaminated because it appears only in the second stage and the treatment is obvious. When there are outliers in $x^{(2)}$ or in w the solution is less evident. Leaving the problem

of optimality beyond the scope of this work, the problem of outliers in $x^{(2)}$ can be treated by using robust first stage estimator. In the case when the instrumental variable is contaminated, a robust first stage is not enough, because the contamination emerges on the second stage anyhow. If we use the non-robust estimator on the first stage, then the structure of the data changes arbitrarily. If we use the robust estimator, then we correct the bias of $\hat{\alpha}$, but $\hat{x}^{(2)} = w^\top \hat{\alpha}$ still depends on w, which means that we have a retained outlier in the main equation. A straightforward solution is to use robust estimators for both stages, which preserve the structure of the data after the first stage and downweight the outliers, moving them to the bulk of the data in the second stage. We implemented the robust estimation procedures for both types of contamination. The results are shown in Table 3.1 and Figure 3.1. The grey shaded boxplots are the robust versions of 2SLS based on MM-estimators introduced by Yohai (1987). We can see that the robust version works well, there is no considerable bias, and most importantly, the loss of efficiency is not dramatic. In Table 3.1 we see that the variances of the parameters under the model for the robust estimator are only slightly larger than for the classical estimator.

Table 3.1: Bias, Variance and MSE of the classical and robust 2SLS at the model and under two types of contamination

$N = 200$	Not contaminated			w is contaminated			$x^{(2)}$ is contaminated		
Classical	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
α_0	0.0055	0.0035	0.0035	−0.0343	0.0039	0.0051	0.0962	0.0088	0.0180
α_1	0.0007	0.0036	0.0036	−0.2002	0.0151	0.0551	−0.0907	0.0088	0.0170
β_0	−0.0004	0.0057	0.0057	0.0265	0.0081	0.0088	−0.1359	0.0383	0.0568
β_2	0.0044	0.0137	0.0137	0.2657	0.0839	0.1545	0.2354	0.1260	0.1814
Robust	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
α_0	0.0066	0.0036	0.0037	0.0065	0.0035	0.0036	0.0060	0.0037	0.0037
α_1	−0.0004	0.0041	0.0041	−0.0004	0.0042	0.0042	0.0002	0.0041	0.0041
β_0	−0.0019	0.0059	0.0059	−0.0129	0.0059	0.0061	−0.0051	0.0058	0.0058
β_2	0.0082	0.0158	0.0159	−0.0283	0.0230	0.0239	0.0131	0.0151	0.0153

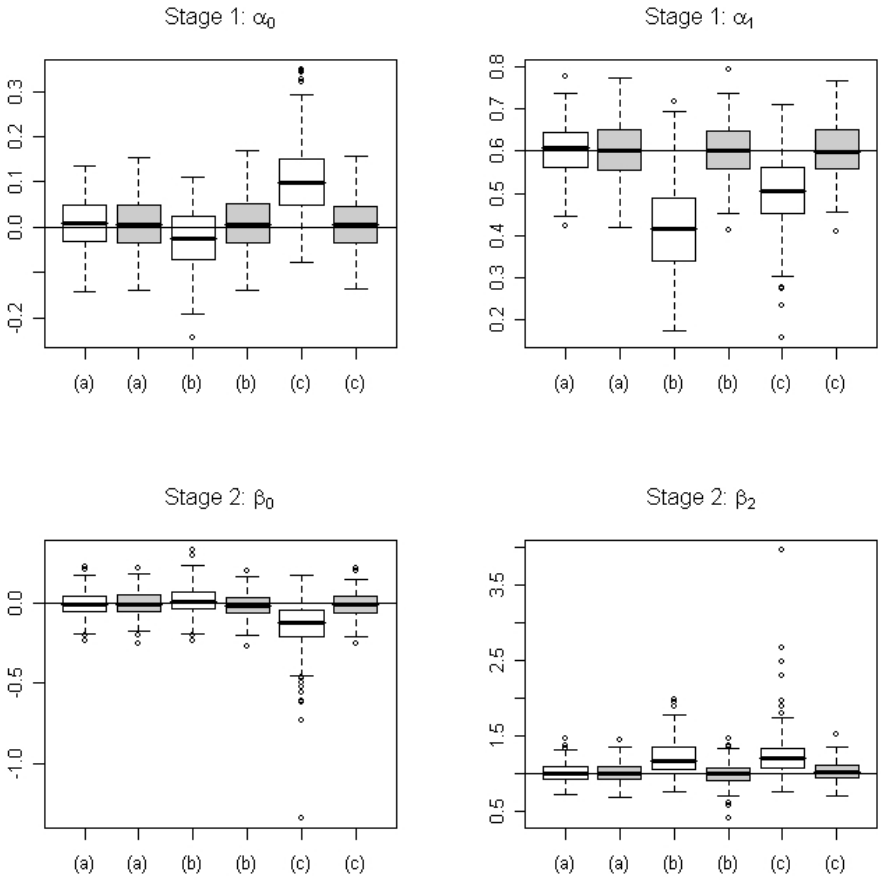

Figure 3.1: *2SLS, parameter estimates. Unshaded boxplots correspond to the classical 2SLS and shaded boxplots correspond to the robust 2SLS. Case (a) is without contamination, (b) is with contamination of w, and (c) is with contamination of $x^{(2)}$. The top panels correspond to the auxiliary regression, the bottom panels to the regression of interest. Horizontal lines mark the true values of the parameters.*

3.5.3 Time Series

In time series analysis, it is usually necessary to decompose the deterministic and the stochastic components and to make corresponding inferences.

For this purpose, the two-stage estimation is usually preferred. In the first stage, the trend is removed and in the second stage, the stochastic component is modeled. To illustrate the results developed in the previous sections, we consider the standard process with linear trend and autoregression (AR) of order 1, i.e.

$$y_t = \alpha_1 t + a_t, \tag{3.10}$$

where a_t follows an autoregressive process of order 1,

$$a_t = \beta_1 a_{t-1} + u_t, \tag{3.11}$$

and u_t are independent and identically distributed according to a $N(0, \sigma^2)$.

The typical way to proceed is to estimate the trend by OLS or MLE and to estimate the 2nd stage by MLE. Both estimators are non-robust. The score function of the first stage is:

$$\Psi_1(y_t; S(F)) = t(y_t - \alpha t), \tag{3.12}$$

which is clearly unbounded in y_t. The score function of the second stage is

$$\Psi_2((y_t, y_{t-1}); h(y_t; S(F)), T(F)) = (y_{t-1} - \alpha(t-1))(y_t - \alpha t - \beta(y_{t-1} - \alpha(t-1))), \tag{3.13}$$

which is also unbounded because of y_t and y_{t-1}. Hence, one possible outlier in the time series can bias both the estimators of trend and of AR. The solution in this case is less evident. Clearly, the use of a robust estimator for the trend is necessary, but the question of the type of estimator for the AR is more complicated, see e.g. Künsch (1984) and de Luna and Genton (2001), and a review in (Maronna et al., 2006, Ch. 8).

Simulations

We simulate the process defined by (3.10), (3.11). We have $\alpha_1 = 0.02$, $\beta_1 = 0.5$, $u_t \sim N(0, 1)$. The intercepts α_0 and β_0 are equal to zero. The length of the trajectory is $N = 500$. We estimate this model with and without contamination. The contamination is introduced by replacing the observations of the time series generated from the model by random numbers from $N(0, 1)$ without trend with probability $\epsilon = 0.05$. The experiment is repeated 200 times. The values of average bias and variance are presented in the table 3.2. The boxplots of the estimates are in the Figure 3.2.

For this simple example we propose a natural robustification procedure similar to one explained in the last paragraph of Section 3.2. We use the MM-estimator for the trend and compute the residuals using the robust weights, i.e.

$$\hat{r}_i = (y_t - \hat{\alpha}t)\omega(y_t - \hat{\alpha}t), \tag{3.14}$$

where r_i denotes the residual, and $\omega(\cdot)$ denotes the robustness weight. The weighting function allows to approach the outliers to the bulk of the data and to reduce their influence in the second estimation stage. Hence, in the second step we can use the classical estimator of the AR process.

From the boxplots in Figure 3.2 it is clear that for the classical procedure the estimates are biased both for trend and autoregression. Even if the test for the presence of trend is not going to change, the prediction will be seriously biased. The robust estimator performs well. There is small bias, but it is almost negligible. The loss of efficiency at the model is not dramatic neither (see Table 3.2).

Table 3.2: Bias, Variance and MSE of the classical and robust estimators at the model and with contamination

$N = 200$	Not contaminated		y_t is contaminated	
Classical	Bias	Var	Bias	Var
α_0	0.0011	0.0296	0.0016	0.0345
α_1	0.000001	0.00001	−0.0005	0.00001
β_0	−0.00001	0.0001	−0.0001	0.00001
β_2	−0.0100	0.0015	−0.2918	0.0092
Robust	Bias	Var	Bias	Var
α_0	0.0008	0.0301	−0.0082	0.0301
α_1	0.00001	0.00001	0.00003	0.00001
β_0	0.0003	0.0001	−0.0001	0.00001
β_2	−0.0129	0.0015	−0.0343	0.0029

3.6 Three- and n-Stage Estimation

Suppose we have the model with more than two stages. There are many examples, for instance, estimation of a MA with trend component. The popular method is to estimate the trend component in the first stage and to estimate the MA component by Durbin method which requires two

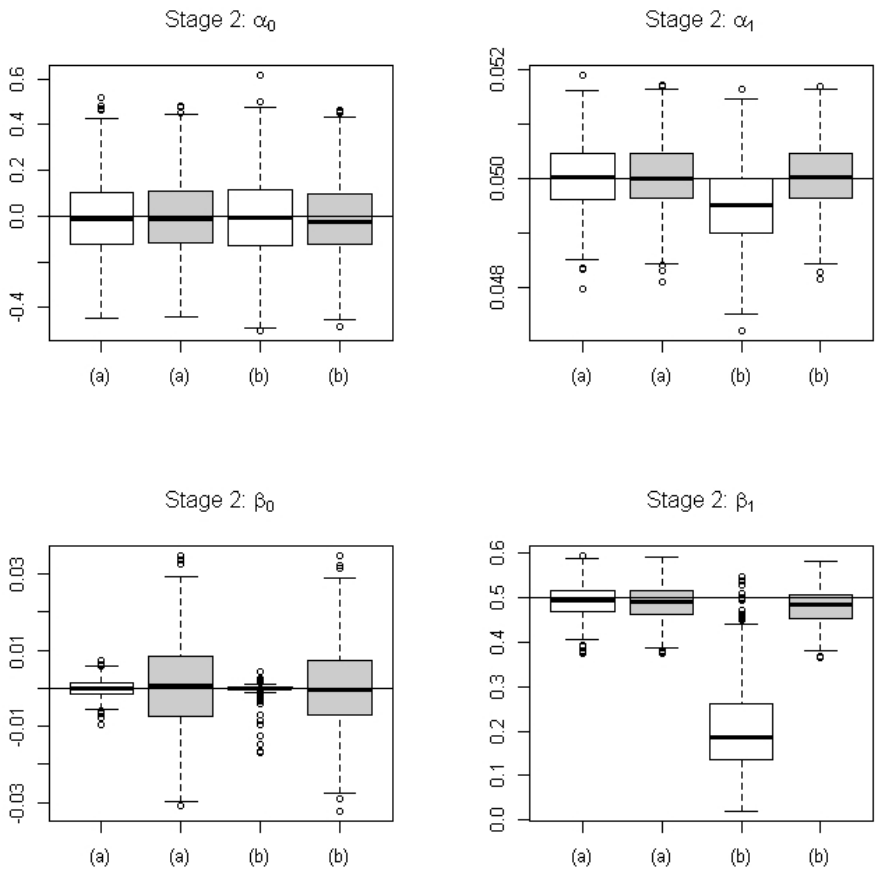

Figure 3.2: *Parameter estimates of autoregression with trend. Unshaded boxplots correspond to the classical estimation procedure and shaded boxplots correspond to the robust estimator. Case (a) is without contamination, (b) is with contamination. The top panels correspond to the estimation of trend, the bottom panels to the estimation of autoregression. Horizontal lines mark the true values of the parameters.*

estimation stages. Another example is the simultaneous equations models with selectivity, it requires the estimation of the selection equation in the first stage, and 2SLS in the second stage, which in general is estimated in two steps. We discuss this model in detail in Chapter 5. In this section we obtain the IF for the n-stage M-estimator. This result is formalized

in the following proposition.

Proposition 1. *Assume that z is a statistical datum containing the exogenous and endogenous variables. $h_k\{z_{k-1}; T_{k-1}(F)\}$ is a piecewise differentiable function transmitting information from stage $k-1$ to k. For the model given by the following sequence of M-estimators*

$$\begin{aligned}
&\textit{Stage 1}: && E_F\left[\Psi_1\{z_1; T_1(F)\}\right] = 0,\\
&\textit{Stage 2}: && E_F\left(\Psi_2[z_2; h_2\{z_1; T_1(F)\}, T_2(F)]\right) = 0\\
&\dots &&\\
&\textit{Stage n}: && E_F\left(\Psi_n[z_n; h_n\{z_{n-1}; T_{n-1}(F)\}, T_n(F)]\right) = 0, \quad (3.15)
\end{aligned}$$

the influence function of the n'th stage estimator is given by

$$\begin{aligned}
IF(z; T_n, F) \;=\; & M(\Psi_n)^{-1}\Bigg(\Psi_n[z_n; h_n\{z_{n-1}; T_{n-1}(F)\}, T_n(F)]\\
& + \int \frac{\partial}{\partial\theta}\Psi_n \frac{\partial}{\partial\eta} h_n dF \cdot IF(T_{n-1})\Bigg), \qquad (3.16)
\end{aligned}$$

where $\frac{\partial}{\partial\eta}h_n = \frac{\partial}{\partial\eta}h_n(z_{n-1};\eta)$, $\frac{\partial}{\partial\theta}\Psi_n = \frac{\partial}{\partial\theta}\Psi_n\{z_n;\theta, T_n(F)\}$, *and* $IF(T_{n-1}) = IF(z; T_{n-1}, F)$.

Proof. Straightforward generalization of the IF for two-stage M-estimator derived in Section 3.2. □

Clear that expanding the $IF(z; T_{n-1}, F)$ term in (3.16) gives us the dependence of the n'th stage on all the previous stages up to the first one. The estimator $T_n(F)$ depends on F directly and indirectly through the sequence of estimators $T_{n-1}(F), \dots, T_1(F)$. This dependence occurs because of the presence of $h(\cdot;\cdot)$ function. If this function is not present then we obtain a sequence of independent estimators and the second line of (3.16) vanishes.

The robustness properties of the n-stage estimator depend on the boundedness of all the IF's. It means that if the IF in one of the stages is unbounded then the complete estimator is non-robust.

Chapter 4

Robust Inference in Sample Selection Models

In this chapter we present the main results concerning the robustness issues in standard selection model. The chapter is structured as follows. In Section 4.1 we investigate the robustness properties of the Heckman's two-stage estimator, we compute its influence function, the change-of-variance function, give the connection between the influence function and the asymptotic variance of the estimator, and show the nonrobustness of the standard sample selection bias test. Section 4.2 is devoted to the robust estimation and inference. We explore the possibilities to obtain a robust estimator and propose a simple robust alternative to the sample selection bias test. The Monte Carlo simulation study is given in Section 4.3. The real data application is presented in Section 4.4. The simple extension to the switching regressions model is offered in Section 4.5.

4.1 Robustness Issues with Heckman's Two-stage Estimator

In this section we present the main results concerning the two-stage estimator. We derive its influence function (IF) and its change-of-variance function (CVF) and discuss the robustness properties of the classical estimator. Moreover, we explain the connection between its IF and the asymptotic variance. Finally, we explore the robustness properties of the SSB test.

We consider a parametric sample selection model $\{F_\theta\}$, where $\theta =$

(β_1, β_2) lies in Θ, a compact subset of $\mathbb{R}^{p_1+p_2}$. 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. The Heckman's estimator is a particular case of general two-stage M-estimators, 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 $\mathbb{R}^{p_1+p_2}$ and its range is a vector in $\mathbb{R}^{p_2}$.

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

$$\int \Psi_1\{(x_1, y_1); S(F)\} dF = 0, \tag{4.1}$$

$$\int \Psi_2[(x_2, y_2); \lambda\{(x_1, y_1); S(F)\}, T(F)] dF = 0, \tag{4.2}$$

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 $\Psi_1(\cdot;\cdot)$ is given by (A.12), and $\Psi_2(\cdot;\cdot,\cdot)$ is given by (A.8). 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 $S(F)$.

4.1.1 Influence Function

For a given functional $T(F)$, the influence function (IF) is defined by Hampel (1974) as $IF(z; T, F) = \lim_{\epsilon \to 0} [T\{(1-\epsilon)F + \epsilon\Delta_z\} - T(F)]/\epsilon$, 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. An estimator is considered to be locally robust if small departures from the assumed distribution have only small effects on the estimator.

Assume that we have a contaminated distribution $F_\epsilon = (1-\epsilon)F + \epsilon G$, where G is some arbitrary distribution function. Using a von Mises (1947) expansion, we can approximate the statistical functional $T(F_\epsilon)$ at the assumed distribution F as

$$T(F_\epsilon) = T(F) + \epsilon \int IF(z; T, F) dG + o(\epsilon) \tag{4.3}$$

and the maximum bias over the neighborhood described by F_ϵ 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. Notice also that (4.3) can be used to approximate numerically the bias of the estimator T at a given underlying distribution F_ϵ for a given G. The next proposition gives the influence function of Heckman's two-stage estimator.

Proposition 2. *For the model (2.1)-(2.5), the IF of the Heckman's two-stage estimator is*

$$IF(z; T, F) = \left\{ \int \begin{pmatrix} x_2 x_2^T & \lambda x_2 \\ \lambda x_2^T & \lambda^2 \end{pmatrix} y_1 dF \right\}^{-1} \left\{ (y_2 - x_2^T\beta_2 - \lambda\beta_\lambda) \begin{pmatrix} x_2 \\ \lambda \end{pmatrix} y_1 + \int \begin{pmatrix} x_2\beta_\lambda \\ \lambda\beta_\lambda \end{pmatrix} y_1 \lambda' dF \cdot IF(z; S, F) \right\}, \quad (4.4)$$

where

$$IF(z; S, F) = \left(\int \left[\frac{\phi(x_1^T\beta_1)^2 x_1 x_1^T}{\Phi(x_1^T\beta_1)\{1 - \Phi(x_1^T\beta_1)\}} \right] dF \right)^{-1} \times \{y_1 - \Phi(x_1^T\beta_1)\} \frac{\phi(x_1^T\beta_1) x_1}{\Phi(x_1^T\beta_1)\{1 - \Phi(x_1^T\beta_1)\}}. \quad (4.5)$$

Proofs and derivations of all results are given in Appendix A.

The first term of (4.4) 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 it tends to zero from the right (the solid line in Figure 4.3). From (4.5) 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.

If there is no selection mechanism involved, then the IF has only the first term. The same occurs if there is no contamination in the selection stage. In Section 4.2 we will present two ways to construct two-stage estimator with a bounded influence function.

Graphical representation

We generate $x_1 \sim N(0,1)$, $x_2 \sim N(0,1)$, and the errors e_1 and e_2 from a bivariate normal distribution with expectation zero, $\sigma_2 = 1$, and $\rho = 0.5$. The graphs of the IF for four types of contamination can be seen in Figure 4.1 and Figure 4.2. The solid lines represent the IF. The boxplots represent the standardized biases [1] of the estimators of the simulated samples under contamination of one observation number k. The top panel of Figure 4.1 represents the IF depending on x_1 when the corresponding value of $y_1 = 1$, i.e. the outlier in the selection equation which is transmitted to the main equation via λ. The bottom panel of Figure 4.1 represents the IF depending on x_1 when the corresponding value of $y_1 = 0$, i.e. the outlier in the selection equation which is not transmitted to the main equation, but as we can see it still influences the final estimator. The top panel of Figure 4.2 represents the IF depending on x_2, i.e. the sensitivity to the leverage outlier. The bottom panel of Figure 4.2 represents the IF depending on y_2, i.e. the outlier in the dependent variable. Note that the graphs in Figure 4.2 essentially the same as the graphs of the IF for normal linear regression.

We can see that the IF's are unbounded, which corresponds to the theoretical results. The bias of the estimator under contamination is not bounded. As it was mentioned the MLE is also non-robust. It's IF is unbounded, the figures can be seen in Salazar (2008). The IF of the probit estimator is plotted in Figure A.1 in appendix.

4.1.2 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

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

[1] standardized bias= $n\{\hat{\beta}(F_\epsilon) - \hat{\beta}(F)\}$

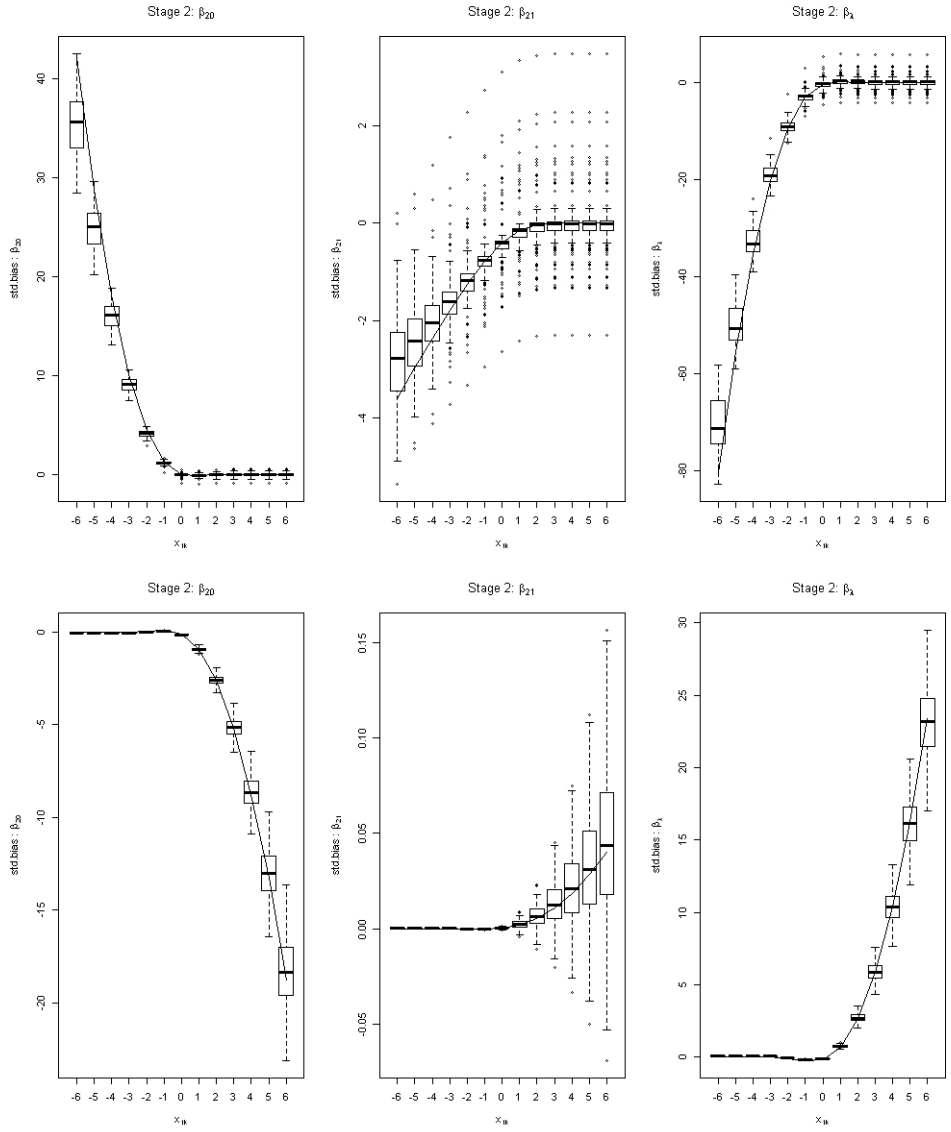

Figure 4.1: *IF of the Heckman two-stage estimator. The solid lines represent the IF. The boxplots represent the standardized biases of the estimators of the simulated samples under contamination of one observation number k. Top panel corresponds to the contamination of Case A, i.e. x_{1k} varies from -6 to 6 and $y_{1k} = 1$. The bottom panel corresponds to the contamination of Case B, i.e. x_{1k} varies from -6 to 6 and $y_{1k} = 0$.*

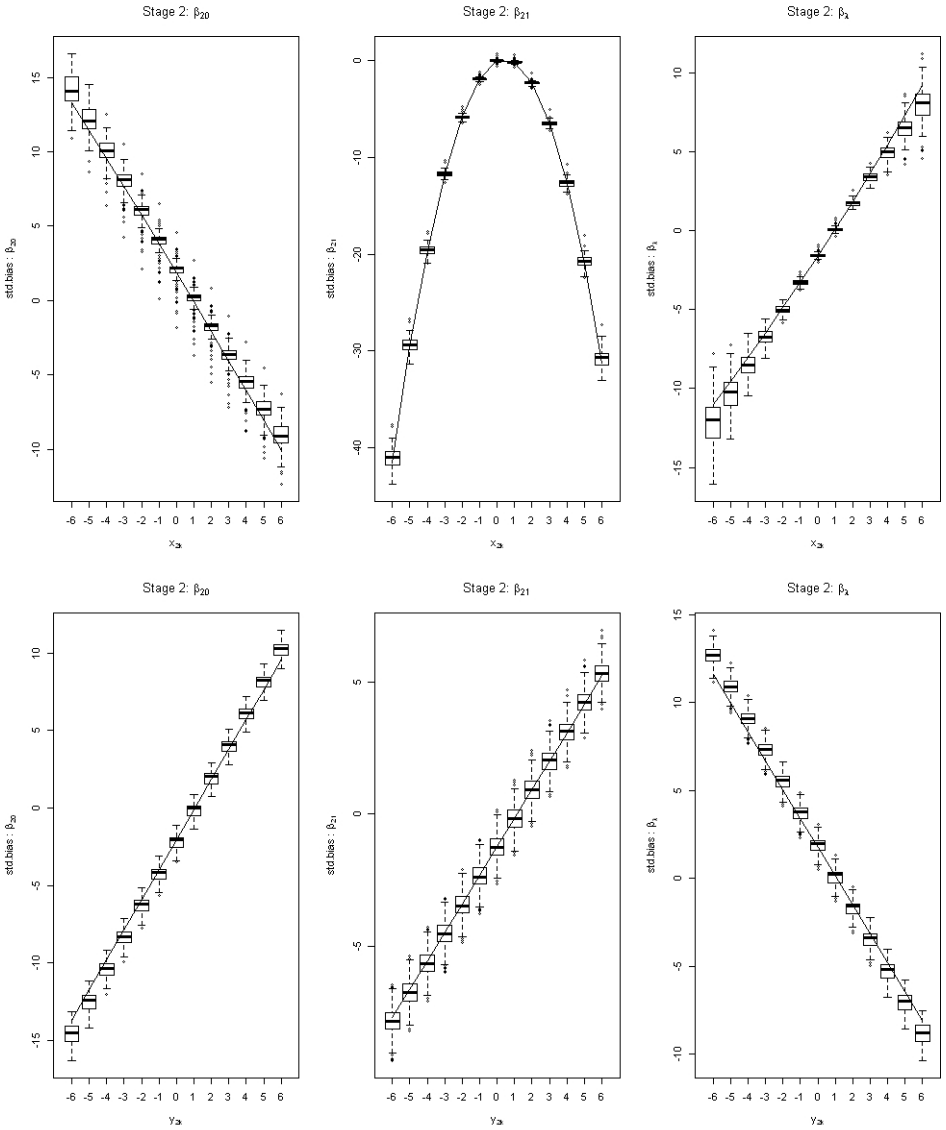

Figure 4.2: *IF of the Heckman two-stage estimator. The solid lines represent the IF. The boxplots represent the standardized biases of the estimators of the simulated samples under contamination of one observation number* k*. Top panel corresponds to the contamination of Case C, i.e.* x_{2k} *varies from* -6 *to* 6*. The bottom panel corresponds to the contamination of Case D, i.e.* y_{2k} *varies from* -6 *to* 6*.*

Specifically, denote the components of the IF as follows:

$$\begin{aligned} a(z) &= (y_2 - x_2^T\beta_2 - \lambda\beta_\lambda)\begin{pmatrix} x_2 \\ \lambda \end{pmatrix}, \\ b(z) &= \int \begin{pmatrix} x_2\beta_\lambda \\ \lambda\beta_\lambda \end{pmatrix} \lambda' dF \cdot IF(z;S,F), \\ M(\Psi_2) &= \int \begin{pmatrix} x_2x_2^T & \lambda x_2 \\ \lambda x_2^T & \lambda^2 \end{pmatrix} dF. \end{aligned} \tag{4.6}$$

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

$$\begin{aligned} V(T,F) &= M(\Psi_2)^{-1} \int \left\{ a(z)a(z)^T + a(z)b(z)^T \right. \\ &\quad \left. + b(z)a(z)^T + b(z)b(z)^T \right\} dF(z)\, M(\Psi_2)^{-1}. \end{aligned}$$

After integration and some simplifications (see Appendix B) we obtain the asymptotic variance matrix of the classical estimator

$$\begin{aligned} V\left\{ \begin{pmatrix} \beta_2 \\ \beta_\lambda \end{pmatrix}, F \right\} &= (X^TX)^{-1}\left[\sigma_2^2 \left\{ X^T \left(I - \frac{\beta_\lambda^2}{\sigma_2^2}\Delta \right) X \right\} \right. \\ &\quad \left. + \beta_\lambda^2 X^T \Delta X_1 \mathrm{Var}(S,F) X_1^T \Delta X \right] (X^TX)^{-1}, \end{aligned} \tag{4.7}$$

where Δ is a diagonal matrix with elements $\delta_{ii} = \frac{\partial\lambda(x_{1i}\beta_1)}{\partial(x_{1i}\beta_1)}$, matrix X consists of vectors $x_i = \begin{pmatrix} x_{2i} \\ \lambda_i \end{pmatrix}$, 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 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 will be used in Section 4.1.3 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). As in (4.3) a von Mises (1947) expansion of $\log V(T, F_\epsilon)$ at F gives

$$V(T, F_\epsilon) \cong V(T, F) \exp\left\{\epsilon \int \frac{CVF(z; T, F)}{V(T, F)} dG\right\}. \tag{4.8}$$

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). Similarly to the approximation of the bias, using (4.8) 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 3. *The CVF of the Heckman's two-stage estimator is given by*

$$\begin{aligned} CVF(z; S, T, F) =& V - M(\Psi_2)^{-1} \left\{ \int D_H dF + \begin{pmatrix} x_2 x_2^T & \lambda x_2 \\ \lambda x_2^T & \lambda^2 \end{pmatrix} \right\} V \\ &+ M(\Psi_2)^{-1} \int \left\{ A_H a(z)^T + A_H b(z)^T \right\} dF M(\Psi_2)^{-1} \\ &+ M(\Psi_2)^{-1} \int \left\{ B_H b(z)^T + a(z) A_H^T \right\} dF M(\Psi_2)^{-1} \\ &+ M(\Psi_2)^{-1} \int \left\{ b(z) A_H^T + b(z) B_H^T \right\} dF M(\Psi_2)^{-1} \\ &+ M(\Psi_2)^{-1} \left\{ a(z) + b(z) \right\} \left\{ a(z) + b(z) \right\}^T M(\Psi_2)^{-1} \\ &- V \left\{ \int D_H dF + \begin{pmatrix} x_2 x_2^T & \lambda x_2 \\ \lambda x_2^T & \lambda^2 \end{pmatrix} \right\} M(\Psi_2)^{-1}, \end{aligned} \tag{4.9}$$

where V denotes the variance of the Heckman (1979) estimator, $M(\Psi_2)$ is defined by (4.6), $A_H = \frac{\partial}{\partial \epsilon} a(z)$, $B_H = \frac{\partial}{\partial \epsilon} b(z)$, and $D_H = \frac{\partial}{\partial \epsilon} \frac{\partial}{\partial \theta} \Psi_2(z; \lambda, \theta)$. All these terms are given explicitly in Appendix A.

The CVF has several sources of unboundedness. The first line of (4.9) 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 fifth line. Finally, in the fourth line there are two terms 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).

4.1.3 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 by applied researchers. Here we investigate the effect of contamination on this test statistic

$$\tau_n = \sqrt{n}\frac{\hat{\beta}_\lambda}{\sqrt{V(\beta_\lambda, F)}}.$$

Using the expressions for the IF and CVF of the estimator and its asymptotic variance, we obtain the von Mises (1947) 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.

4.2 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.

4.2.1 Robust Two-Stage Estimator

From the expression of the IF in (4.4), 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 (4.1) and (4.2). We can obtain a robust estimator by bounding both score functions. In the first stage, we construct a robust probit estimator following the idea of Cantoni and Ronchetti (2001). We use a general class of M-estimators of Mallows (1975) type, where the influence of deviations on y_1 and x_1 are bounded separately. The estimator is defined by the following score function:

$$\Psi_1^R\{z_1; S(F)\} = \nu(z_1; \mu)\omega_1(x_1)\mu^{'} - \alpha(\beta_1), \tag{4.10}$$

where $\alpha(\beta_1) = \frac{1}{n}\sum_{i=1}^{n} E\{\nu(z_{1i}; \mu_i)\}\omega_1(x_{1i})\mu_i^T$ 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, and $\mu_i = \mu_i(z_{1i}, \beta_1) = \Phi(x_{1i}^T\beta_1)$.

The weight functions are defined by

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

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

$$\psi_{c_1}(r) = \begin{cases} r, & |r| \leq c_1, \\ c_1 \text{sign}(r), & |r| > c_1. \end{cases} \tag{4.11}$$

The tuning constant c_1 is chosen to ensure a given level of asymptotic efficiency at the model. 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 $H = X(X^TX)^{-1}X^T$. 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} (see Rousseeuw 1985 and Rousseeuw and Van Driessen 1999). 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

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

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:

$$\Psi_2^R(z_2;\lambda,T) = \Psi_{c_2}(y_2 - x_2^T\beta_2 - \lambda\beta_\lambda)\omega(x_2,\lambda)y_1, \tag{4.12}$$

where $\Psi_{c_2}(\cdot)$ is the classical Huber function from (4.11), but with a different tuning constant c_2, $\omega(\cdot)$ is the weight function, which can also be based on the robust Mahalanobis distance $d(x_2,\lambda)$, e.g.

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

We summarize the result in the following proposition.

Proposition 4. *Under the assumptions stated in Appendix A, Heckman's two-stage estimator defined by (4.10) and (4.12) is robust, consistent, and asymptotically normal, with asymptotic variance given by:*

$$V(T,F) = M\left(\Psi_2^R\right)^{-1}\int\{a_R(z)a_R(z)^T + b_R(z)b_R(z)^T\}dFM\left(\Psi_2^R\right)^{-1}, \tag{4.13}$$

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

$$b_R(z) = \int\frac{\partial}{\partial\lambda}\Psi_2^R(z;\lambda,T)\lambda' dF\left\{\int\frac{\partial}{\partial\beta_1}\Psi_1^R(z;S)dF\right\}^{-1}\Psi_1^R(z;S).$$

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. Huber function, it is relatively simple.

4.2.2 Robust Inverse Mills Ratio

Often the outliers appear only in one or a few components of the observation $z = (x_1, y_1, x_2, y_2)$. In these cases the use of robust estimators in both stages is not necessary. If outliers are in y_2 and/or x_2, then a robust estimator for the equation of interest is needed. If outliers are in x_1 and $y_1 = 0$, then a robust estimator of the probit model is needed. The most complicated case is when a leverage point x_1 with $y_1 = 1$ in the selection equation is transferred to the equation of interest through the exogenous variable λ, a nonlinear transformation of $x_1^T\beta_1$.

In this case a natural solution is to bound the influence of the outliers that come into the main equation when computing λ. Since λ is unbounded and approximately linear in the predictor $x_1^T\beta_1$ from the left (see Figure 4.3), we can transform the linear predictor in such a way that it becomes bounded, ensuring the boundedness of λ, and hence the boundedness of the IF of the final estimator.

To achieve this we rewrite the linear predictor $x_1^T\beta_1 = \Phi^{-1}\{y_1 - rV^{1/2}(\mu)\}$, where $r = (y_1 - \mu)/V^{1/2}(\mu)$ and $\mu = \Phi(x_1^T\beta_1)$. Then, using the bounded function ψ_{c_1} from (4.11), we obtain the bounded linear predictor

$$\eta(x_1^T\beta_1) = \Phi^{-1}\left\{y_1 - \psi_{c_1}(r)V^{1/2}(\mu)\right\}, \tag{4.14}$$

and (4.14) ensures the boundedness of the inverse Mills ratio. It also has the advantage to avoid introducing an additional weight to bound directly λ, which would increase the complexity of the estimator. Note that, if $c_1 \to \infty$, then $\eta = x_1^T\beta_1$. The classical and robust versions of the inverse Mills ratio are plotted in Figure 4.3. Depending on the tuning parameter c_1 of the robust probit, the influence of large linear predictors can be reduced, which ensures the robustness of the estimator.

4.2.3 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 (4.13).

The first term of (4.13), $M(\Psi_2^R)^{-1}\int a_R(z)a_R(z)^T dF M(\Psi_2^R)^{-1}$, 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 es-

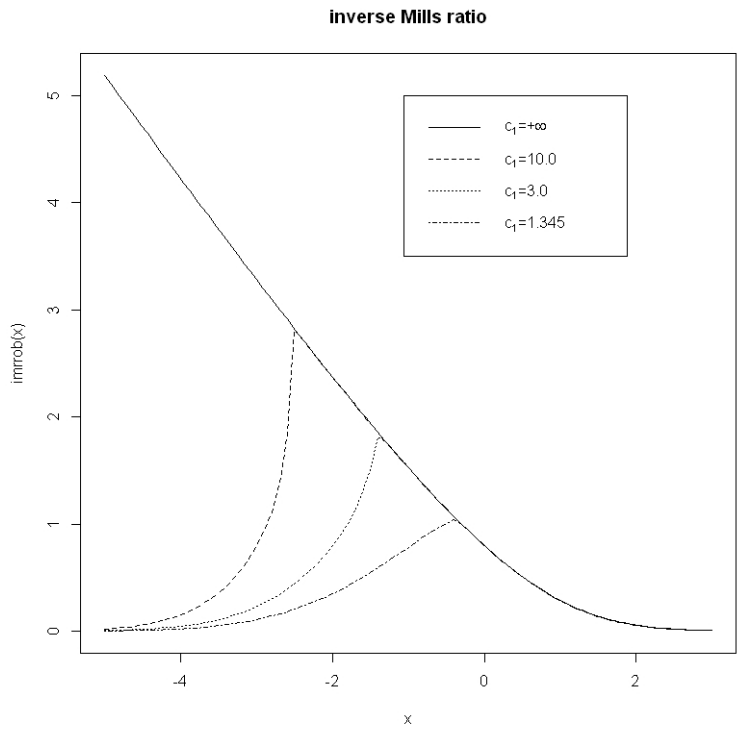

Figure 4.3: *The solid line corresponds to the classical inverse Mills ratio. The dotted lines correspond to the robust inverse Mills ratio from Section 4.2.2.*

timate 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 in (4.13), $M(\Psi_2^R)^{-1}\int b_R(z)b_R(z)^T dF M(\Psi_2^R)^{-1}$, 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

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

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

4.3 Simulation Study

Consider the model described in Section 2.1. We carry out a Monte Carlo simulation study to illustrate the robustness issues in this model and compare different estimators. In our experiment we generate $x_1 \sim N(0,1)$, $x_2 \sim N(0,1)$, and the errors e_1 and e_2 from a bivariate normal distribution with expectation zero, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.5$, which gives $\beta_\lambda = 0.5$. The degree of censoring is controlled by the intercept in the selection equation, denoted by β_{10} and set to 0, which corresponds to 50% of censoring. Results (presented in Appendix C) are similar for other censoring proportions such as 75% and 25%. The intercept in the equation of interest is $\beta_{20} = 0$. The slope coefficients are $\beta_{11} = \beta_{21} = 1$. 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 $(4, 0, 1, 1)$, corresponding to (x_1, y_1, x_2, y_2). 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 $(-4, 1, 1, 1)$. This is the most complicated case, because the outliers influence not only the first estimation stage, but also the second stage through λ. The sample size is $N = 200$ and we repeat the experiment 500 times. We do not show in this section (see Appendix C) the cases when only the second stage is contaminated because this is essentially the situation of standard linear regression which has been studied extensively; see e.g. Hampel et al. (1986) and Maronna et al. (2006).

In Table 4.1 we present the results of the estimation of the first stage using a classical probit MLE and a robust probit M-estimator. Under the model we see that the robust estimator is less efficient, but in the presence of a small amount of contamination it remains stable, both for bias and for variance. The classical estimator is clearly biased and becomes much less efficient. In Table C.1 we present the results of the two-stage procedure for four different estimators. We compare the classical estimator, robust two-stage (robust 2S) from Section 3.1, robust inverse Mills ratio (IMR) from Section 3.2, and the estimator using only the robust probit in the first stage and OLS in the second. First of all, notice that all the estimators perform well without contamination. The loss of efficiency for the robust versions is reasonable and the bias is close to zero. Obviously, under

contamination the classical estimator breaks down. This effect can be seen in Figure 4.4. The boxplots correspond to the four estimators described above. The classical estimator is denoted by (a), the robust probit with OLS by (b), the robust IMR by (c), and the robust 2S by (d). In the case when the outlier is not transferred to the equation of interest (Figure 4.4 top panel) it is enough to use a robust probit, but when the outlier emerges in the equation of interest (Figure 4.4 bottom panel), a robust estimation of the second stage is necessary. The robust IMR and the robust two-stage estimators are both stable regardless of the presence or absence of outliers in λ or x_1. The robust IMR estimator has smaller variance than the robust two-stage, but in the case of outliers in y_2 or x_2 it will require a robust estimator in the second stage anyhow. So this estimator cannot be considered as a self-contained estimator, but it is a useful tool in some situations.

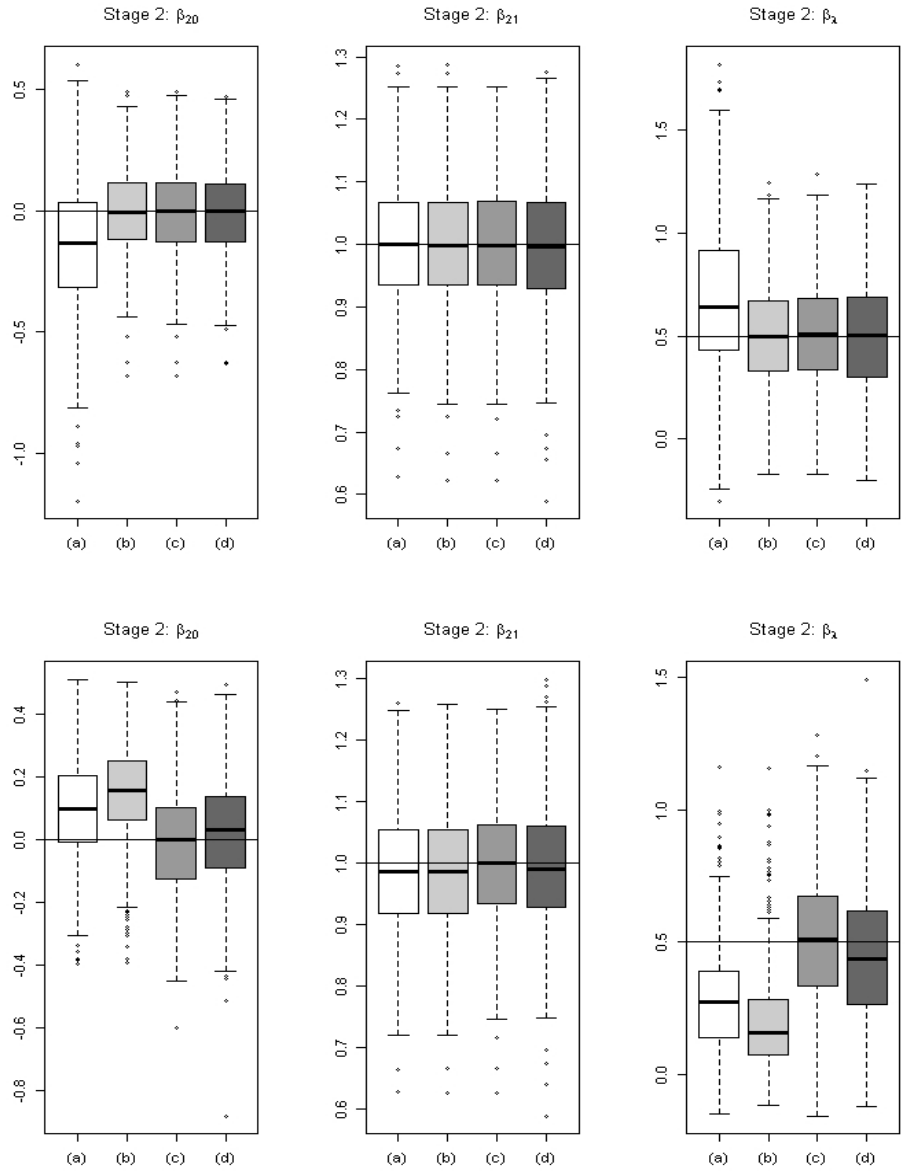

Figure 4.4: *Parameter estimates by classical and robust two-stage estimators with contamination of x_1. Top panel corresponds to $y_1 = 0$, and bottom panel corresponds to $y_1 = 1$. Case (a) corresponds to the classical estimator, (b) corresponds to robust probit with OLS on the second stage, (c) corresponds to robust IMR, and (d) to robust two-stage. Horizontal lines mark the true values of the parameters.*

Table 4.1: *Bias, Variance and MSE of the classical and robust probit MLE at the model and under two types of contamination.*

$N = 200$	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}	0.0007	0.0115	0.0115	0.0474	0.0104	0.0127	−0.0502	0.0093	0.0119
β_{11}	0.0274	0.0213	0.0221	−0.2943	0.0378	0.1244	−0.3160	0.0341	0.1339
Robust									
β_{10}	0.0003	0.0118	0.0118	0.0025	0.0119	0.0120	−0.0022	0.0118	0.0119
β_{11}	0.0281	0.0229	0.0237	0.0105	0.0287	0.0288	0.0107	0.0294	0.0295

Table 4.2: Bias, Variance and MSE of the classical and robust two-stage estimators at the model and under two types of contamination

$N = 200$	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}	−0.0025	0.0299	0.0299	0.0998	0.0269	0.0368	−0.1584	0.0725	0.0976
β_{21}	0.0005	0.0100	0.0100	−0.0141	0.0101	0.0103	0.0002	0.0100	0.0100
$\beta_{2\lambda}$	0.0030	0.0603	0.0603	−0.2161	0.0410	0.0876	0.1828	0.1331	0.1665
Robust probit + OLS									
β_{20}	−0.0026	0.0301	0.0301	0.1504	0.0231	0.0457	−0.0089	0.0320	0.0320
β_{21}	0.0005	0.0100	0.0100	−0.0134	0.0101	0.0103	0.0001	0.0100	0.0100
$\beta_{2\lambda}$	0.0032	0.0607	0.0607	−0.2968	0.0370	0.1251	0.0112	0.0639	0.0640
Robust IMR									
β_{20}	−0.0033	0.0306	0.0306	−0.0084	0.0299	0.0299	−0.0096	0.0326	0.0327
β_{21}	0.0004	0.0099	0.0099	−0.0003	0.0098	0.0098	0.0004	0.0099	0.0099
$\beta_{2\lambda}$	0.0100	0.0646	0.0647	0.0170	0.0639	0.0640	0.0178	0.0679	0.0680
Robust 2S									
β_{20}	−0.0041	0.0325	0.0326	0.0225	0.0317	0.0320	−0.0106	0.0346	0.0347
β_{21}	−0.0027	0.0107	0.0107	−0.0058	0.0110	0.0109	0.0027	0.0108	0.0108
$\beta_{2\lambda}$	0.0005	0.0731	0.0731	−0.0433	0.0643	0.0662	0.0085	0.0773	0.0774

4.4 Example: Ambulatory Expenditures

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, with 526 (15.8%) corresponding 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 could be introduced, by means of the income variable, but the use of this variable for this purpose is arguable. All these variables are significant for the decision to spend, and all except education status and insurance status are significant for the spending amount.

The p-value of the SSB t-test is 0.0986, which is close to the border of the 10% level. Although the 10% significance level can be chosen, at 5% level (which is usually chosen by some conventional wisdom) the variable is not significant. The estimation of this model using joint MLE returns the p-value of the Wald test equal to 0.38 (see Cameron and Trivedi 2009). Such behavior of classical MLE is not surprising due to the fact that it uses stronger distributional assumption than Heckman's estimator, and is even more sensitive to the presence of contamination. The possible conclusion of no selection bias seems to be doubtful.

In Table 4.3 we present the estimation results of the data by the classical estimator obtained using the `R` package `sampleSelection` (Toomet and Henningsen 2008) and by the robust two-stage estimator proposed in Section 4.2.1. For all the variables the differences between the estimates are not dramatic, except for the inverse Mills ratio (IMR) parameter. The robust estimator returns $\hat{\beta}_{IMR} = -0.6768$, compared to the classical $\hat{\beta}_{invMillsRatio} = -0.4802$. We can remark that the classical estimator is downward biased. Moreover if we consider the standard errors, we see that the standard error of the robust estimator (0.2593) is smaller than that of the classical one (0.2907). If the distributional assumptions were satisfied then such situation would be impossible, but in this example it is not surprising, because the classical estimator of the variance is not robust (see Section 4.1). The p-value of the robust SSB test is $p = 0.009$, which leads to the conclusion of the presence of SSB.

Table 4.3: Estimation results of the Medical Expenditures data by the classical estimator and by the robust two-stage estimator from Section 4.2.1. The standard errors are given in parentheses. Significance codes: "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1.

	Classical	Robust
Selection		
intercept	−0.71771 (0.19247)***	−0.74914 (0.19507)***
age	0.09732 (0.02702)***	0.10541 (0.19507)***
female	0.64421 (0.06015)***	0.68741 (0.06226)***
educ	0.07017 (0.01134)***	0.07012 (0.01147)***
blhisp	−0.37449 (0.06175)***	−0.39775 (0.06507)***
totchr	0.79352 (0.07112)***	0.83284 (0.08028)***
ins	0.18124 (0.06259)**	0.18256 (0.06371)**
Outcome		
intercept	5.30257 (0.29414)***	5.40154 (0.27673)***
age	0.20212 (0.02430)***	0.20062 (0.02451)***
female	0.28916 (0.07369)***	0.25501 (0.06992)***
educ	0.01199 (0.01168)	0.01325 (0.01162)
blhisp	−0.18106 (0.06585)**	−0.15508 (0.06507)*
totchr	0.49833 (0.04947)***	0.48116 (0.03822)***
ins	−0.04740 (0.05315)	−0.06707 (0.05159)
inverse Mills ratio	−0.4802 (0.2907)·	−0.67676 (0.25928)**

4.5 Switching Regressions Model

A natural extension of Heckman's model is the switching regression model or "Tobit type-5 model". In this case we have two regimes, and the regime depends on the selection process. This model consists of three equations:

$$\begin{aligned} y_{1i}^* &= x_{1i}^T\beta_1 + e_{1i}, \\ y_{21i}^* &= x_{21i}^T\beta_{21} + e_{2i}, \\ y_{22i}^* &= x_{22i}^T\beta_{22} + e_{3i}, \end{aligned}$$

where the error terms follow a multivariate normal distribution. The observed variables are $y_{1i} = I(y_{1i}^* > 0)$ and

$$y_{2i} = \begin{cases} y_{21i}^*, & \text{if } y_{1i} = 1, \\ y_{22i}^*, & \text{if } y_{1i} = 0. \end{cases}$$

The model can be estimated by a two-stage procedure with the following switching regressions:

$$y_{2i} = \begin{cases} x_{21i}^T\beta_{21} + \beta_{\lambda 1}\lambda_i + v_{1i}, & \text{if } y_{1i} = 1, \\ x_{22i}^T\beta_{22} - \beta_{\lambda 2}\tilde{\lambda}_i + v_{2i}, & \text{if } y_{1i} = 0, \end{cases} \tag{4.15}$$

where $\lambda_i = \lambda_i(x_{1i}^T\beta_1)$ and $\tilde{\lambda}_i = \lambda_i(-x_{1i}^T\beta_1)$. This system can be estimated as two independent linear models.

The IF for the first regression in (4.15) is exactly the same as that in the basic Heckman's model and is given by (4.4). For the second regression in (4.15), a slight modification is required. The IF for the second regression is given by (4.4), where λ is replaced by $\tilde{\lambda}$, $\lambda^{'}$ by

$$\tilde{\lambda}^{'} = \frac{-\{1 - \Phi(x_1^T\beta_1)\}\phi(x_1^T\beta_1)x_1^T\beta_1 + \phi(x_1^T\beta_1)^2}{\{1 - \Phi(x_1^T\beta_1)\}^2}x_1^T,$$

and y_1 by $1 - y_1$.

Obviously, the $\tilde{\lambda}$ is unbounded from the right and therefore the conclusion about the robustness properties remains the same. A robust estimator can be easily obtained using the procedure described in Section 4.2.

Chapter 5

Robust Estimation of Simultaneous Equations Models with Selectivity

In Chapter 4 we discussed the robustness issues for standard Heckman's selection model. In this chapter we extend our results to the case of Simultaneous Equations Models (SEM). This class of models is an extension of multivariate linear models and is one of the central topics in modern econometric theory. SEM with truncated or censored dependent variables show a considerable interest in various applications, e.g. see recent works by Michel-Kerjan et al. (2011) and Di Falco et al. (2011). Many more examples can be found in a book by Maddala (1983).

The classical estimators for SEM are sensitive to deviations from the assumed distribution. There are several proposals of estimators of SEM without selectivity. Krasker (1986) proposed a bounded influence two-stage estimator. Krishnakumar and Ronchetti (1997) suggested the use of robust estimators based on maximum likelihood (ML) and investigated the optimality problem. Recently, Cohen Freue et al. (2011) proposed a robust instrumental variables estimator. For a review of the robustness issue in SEM see Maronna and Yohai (1997) and references therein. But the problem of robust estimation in presence of censoring or selectivity has not been treated.

In this chapter we fill this gap by providing a robust alternative to the classical estimator. The chapter is organized as follows. In Section 5.1 we present the model and discuss the estimation methods. Section 5.2 presents a discussion of the robustness issue. A robust estimator is pro-

posed in Section 5.3. A Monte Carlo simulation study is presented in Section 5.4. In Section 5.5 we illustrate our methodology on a real-data application.

5.1 Simultaneous Equations Models with Selectivity

5.1.1 Definition

The simultaneous equations model with selectivity is given by the following system of equations

$$\begin{aligned} I_i &= w_i^T \alpha + e_{1i}, && (5.1)\\ \Gamma Y_i &= BX_i + e_{2i}, \text{ if } I_i > 0, && (5.2)\\ Y_i &= 0, \text{ if } I_i \leq 0, && (5.3) \end{aligned}$$

where e_1 and e_2 are the error terms following a multivariate normal distribution with zero mean and covariance matrix Σ:

$$\Sigma = \begin{pmatrix} 1 & \sigma_{12}^T \\ \sigma_{12} & \Sigma_{22} \end{pmatrix}, \tag{5.4}$$

Y_i is a $q \times 1$ vector of endogenous variables, X_i is a $p \times 1$ vector of exogenous variables, w is a $l \times 1$ vector consisting of some or all variables from X and also additional exogenous variables. In many applications X includes both exogenous and lagged endogenous variables. The variable I is unobserved, we only know whether $I > 0$ or $I \leq 0$. B and Γ are $p \times q$ and $q \times q$ matrices of parameters, respectively, α is a $l \times 1$ vector of parameters.

Equation (5.1) defines the selectivity rule. The system given by (5.1)-(5.3) is the truncated SEM. The switching SEM, a straightforward generalization, is discussed briefly below (see Remark 1).

5.1.2 Estimation

There are two popular methods to estimate the SEM with selectivity. The first is to use a Heckman (1976, 1979) type procedure. In the first stage we obtain an estimate $\hat{\alpha}$ of α by probit MLE. It is known that

$$E(e_{2i}|I_i > 0) = -\sigma_{12}\frac{\phi(w_i^T\alpha)}{\Phi(w_i^T\alpha)} = -\sigma_{12}\lambda_i, \tag{5.5}$$

where $\phi()$ and $\Phi()$ denote the standard normal probability density and cumulative distribution functions respectively. Without loss of generality, consider the first structural equation in (5.2). We can rewrite it as

$$y_{1i} = \gamma_{12} y_{2i} + \cdots + \gamma_{1q} y_{qi} + \beta_1 X_i - \sigma_{12} \lambda_i + v_i, \tag{5.6}$$

where $E(v_i | I_i > 0) = 0$. Replacing λ_i by its estimate $\hat{\lambda}_i = \phi(w_i^T \hat{\alpha}) / \Phi(w_i^T \hat{\alpha})$ we obtain the parameters γ, β, and σ_{12} by 2SLS (Wooldridge 2002, p. 568). The computation of the asymptotic variance is given in Lee et al. (1980).

Remark 1. If instead of truncated SEM we have a system of switching SEM then the inverse Mills ratio term for the second regime will be

$$\tilde{\lambda} = \frac{\phi(w_i^T \alpha)}{1 - \Phi(w_i^T \alpha)}. \tag{5.7}$$

Apart from this, the estimation procedure remains the same.

The second estimation procedure is to use the full information maximum likelihood (FIML). The major virtue of FIML is its asymptotic efficiency, but if the distributional assumptions are not exactly satisfied the virtue becomes a vice because of very high sensitivity of this estimator to different deviations from the assumptions. For the simple Heckman (1979) selection model different Monte Carlo studies investigating performance of the estimators under non-standard assumptions have been carried out, e.g. see Paarsch (1984) or Puhani (2000) and references therein. Also FIML has higher computational complexity, and the likelihood function of this estimator is not globally concave (Olsen 1982, Toomet and Henningsen 2008), which requires more complicated numerical algorithms. The robustness properties of MLE in general are well known. The influence function is not bounded and the estimator is not robust. The robustification of such estimators is even more complicated, because of the introduction of additional non-linearities via bounding of the score functions. It leads to obstructed tractability of the results, and difficulty to use it in practice.

5.2 Robustness Properties

In this section we derive the influence function of the estimator and discuss the robustness properties and possible problems.

The Heckman-type estimation procedure explained in Section 5.1 can be rewritten in a more general form. Assume that the set of exogenous

variables X consists of the variables $X^{(1)}$, which are used as explanatory variables for the equation of interest, and $X^{(2)}$, which are used as instruments for endogeneity correction equation. Note that the sets $X^{(1)}$ and $X^{(2)}$ can be overlapping. The estimation procedure consists of two steps, where the second step is the 2SLS, which itself can be splitted into two steps. Finally we have a three-stage M-estimator

$$\begin{aligned} E_F[\Psi_1\{(I,w);\alpha\}] &= 0, && (5.8)\\ E_F[\Psi_2\{(y,x^{(2)});\lambda,b\}] &= 0, && (5.9)\\ E_F[\Psi_3\{(y,x^{(1)});\lambda,h(x^{(2)};b),(\beta,\gamma)\}] &= 0, && (5.10) \end{aligned}$$

where $\Psi_1(\cdot;\cdot)$, $\Psi_2(\cdot;\cdot,\cdot)$, and $\Psi_3(\cdot;\cdot,\cdot,\cdot)$ denote the score functions of the first, second and third stage estimators respectively, b is a vector of auxiliary parameters, and $h(\cdot;\cdot) = {x^{(2)}}^T b$ is a function computing the estimated values of $\hat{y}$. The classical estimator can be recovered by using the MLE and OLS Ψ-functions.

The score function of probit MLE is given by

$$\Psi_1\{(I,w);\alpha\} = \left(\frac{I-\Phi(w^T\alpha)}{\Phi(w^T\alpha)(1-\Phi(w^T\alpha))}\right)\phi(w^T\alpha)w. \qquad (5.11)$$

For simplicity of exposition assume that we have only one endogenous y_{2i} in (5.6). Then the score function of the auxiliary stage is

$$\Psi_2\{(y_2,x^{(2)});\lambda,b\} = (y_2 - (x^{(2)T},\lambda)b)\begin{pmatrix} x^{(2)} \\ \lambda \end{pmatrix}, \qquad (5.12)$$

and taking into account that $\hat{y}_2 = (x^{(2)T},\lambda)\hat{b}$, we obtain the score function of the last stage

$$\begin{aligned} \Psi_3\{(y_1,x^{(1)});\lambda,h(x^{(2)};b),(\beta\gamma)\} &= \left[y_1 - \{(x^{(2)T},\lambda)b, x^{(1)T},\lambda\}\begin{pmatrix}\gamma\\ \beta\\ \sigma_{12}\end{pmatrix}\right] \\ &\quad \times \begin{pmatrix}(x^{(2)T},\lambda)b\\ x^{(1)}\\ \lambda\end{pmatrix}. \qquad (5.13) \end{aligned}$$

Note that $\lambda = \lambda(w;\alpha)$, which means that it depends on the first estimation stage. The last estimation stage depends on both previous stages directly through λ and $h(x^{(2)};b)$ and indirectly via the second stage, which itself depends on the first stage through λ.

The following proposition characterizes the IF of the classical Heckman-type estimator.

Proposition 5. *For the model (5.1)-(5.3) the influence function of the Heckman type estimator given in Section 5.1 is unbounded in all the components of statistical datum* $z = (Y, X, w)$.

Proof. The proposition is the corollary of the Proposition 1 from Section 3.6. The score functions are given by (5.11)-(5.13). Consider the first line in (3.16), the score function Ψ_n corresponds to the score function in (5.13). In the second line of (3.16) the $IF(z; T_{n-1}, F)$ depends on the score function (5.12) and on the IF of the previous stage, which itself depends on (5.11). All these score functions are unbounded, which gives unboundedness of the final IF. □

The unboundedness of the IF means that a small amount of contamination can make the estimator arbitrarily biased. Notice that the IF is unbounded in all the components of z, which means that if the contamination appears in the first estimation stage, i.e. in the selection equation, the final estimator will be biased. Taking into account that the inverse Mills ratio term can be also used as an instrument, the direction of the bias and the behavior of the estimator becomes almost unpredictable. In the next section we construct an estimator with a bounded IF.

5.3 Robust Estimation

Robust estimators are usually complicated from the technical point of view. They often require complex numerical methods and/or sophisticated computational algorithms. This technical issue sometimes becomes the reason for practitioners to avoid robust estimators at all. We construct a robust estimator of SEM with selectivity based on a three-stage M-estimator of Mallows type. It is structurally similar to the classical Heckman-type procedure, which makes it a simple-in-practice and useful complement to the classical estimator.

First we estimate the selection equation by the robust version of probit ML. The robustness is achieved by using the bounded score function (5.11). The robustness weights are introduced to control the effect of the leverage outliers in w and residuals $r = \{I - \Phi(w^T\alpha)\}/\{\Phi(w^T\alpha)(1 - \Phi(w^T\alpha))\}$. For a detailed treatment of robust probit see Zhelonkin et al. (2013), and for robust generalized linear models in general see Cantoni and Ronchetti (2001).

Having obtained the robust estimator of α, we can use it to estimate the inverse Mills ratio. Then we use the 2SLS to get rid of the endogeneity

and to consistently estimate the parameters of the equation of interest. We propose to use a two stage robust Mallows type M-estimator, because of its structural closeness to the classical estimator and good robustness properties. The robust Ψ_2-function is the following:

$$\Psi_2^R\{(y, x^{(2)}); \lambda, T_2\} = \Psi_{c_2}\{y_2 - (x^{(2)T}, \lambda)b\}\omega_2\left\{\begin{pmatrix} x^{(2)} \\ \lambda \end{pmatrix}\right\}, \qquad (5.14)$$

where $\Psi_{c_2}(\cdot)$ is the classical Huber (1973) function:

$$\psi_{c_2}(r) = \begin{cases} r, & |r| \leq c_2, \\ c_2 \text{sign}(r), & |r| > c_2, \end{cases} \qquad (5.15)$$

$\omega_2(\cdot)$ is the weight function, which can be based on the robust Mahalanobis distance $d(x^{(2)}, \lambda)$, e.g.

$$\omega_2(x^{(2)}, \lambda) = \begin{cases} x^{(2)}, & \text{if } d(x^{(2)}, \lambda) < c_m, \\ \frac{x^{(2)} c_m}{d(x^{(2)}, \lambda)}, & \text{if } d(x^{(2)}, \lambda) \geq c_m, \end{cases} \qquad (5.16)$$

where c_2 and c_m are tuning constants controlling the degree of robustness. The robust Ψ_3-function is given by:

$$\Psi_3^R\{(y_1, x^{(1)}); \lambda, h(x^{(2)}; T_2), T_3\} = \Psi_{c_3}(r_3)\omega_3\begin{Bmatrix} (x^{(2)T}, \lambda)b \\ x^{(1)} \\ \lambda \end{Bmatrix}, \qquad (5.17)$$

where Ψ_{c_3} is also a Huber-function, ω_3 is a leverage weight function, and

$$r_3 = \left[y_1 - \{(x^{(2)T}, \lambda)b, x^{(1)T}, \lambda\}\begin{pmatrix} \gamma \\ \beta \\ \sigma_{12} \end{pmatrix}\right].$$

The weight functions ω_3 can have the same form as ω_2 in (5.16), possibly even with the same tuning constant. The choice of the tuning constants depends on the desired levels of efficiency and robustness. The optimality problem in the case of linear regression has been studied extensively in literature (Hampel et al., 1986, Ch. 6).

Similarly to the standard selection model discussed in Chapter 4 and formulated in Proposition 4 the robust estimator proposed in this section is also consistent and asymptotically normal at the model F. The test statistics have the same form as in the classical case (see Section 4.2),

which allows to compare the estimators and to see the deviations of the estimators from each other.

The structural similarity of the classical and robust estimators allows to use it also as a diagnostic tool. At the model both estimators are consistent and must show close results. If it is not the case then it is a signal about the distributional deviation of the data from the assumed model or/and about the presence of errors in the data.

Remark 2. If we estimate not under the model, but we allow some deviation from it, then the classical estimator can become arbitrarily biased. The robust estimator allows some bias but it is finite. In case when there are several estimation stages and the atypical observations pass through all or several stages then the accumulated bias of the final estimator can become noticeable even if the robust estimators were used in every stage. In this case it can be controlled by reducing the tuning constants.

5.4 Simulation Study

Consider a system of two equations with two endogenous variables. For each equation we have one instrument x_1 and x_2 generated from a standard normal distribution independently of each other. In the selection equation we have one explanatory variable w, also following a standard normal distribution independent from x_1 and x_2. The error terms follow a multivariate normal distribution

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_{21} \\ \epsilon_{22} \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & 0.5 & 0.75 \\ & 1 & 0.5 \\ & & 1 \end{pmatrix} \right\}.$$

The sample size is $N = 200$ and we repeat the experiment 500 times.

We study the effect of contamination in the selection stage. With probability 0.01 we replace the values in the datum (I_i, w_i) by $(1, -6)$. We expect the outliers to affect the estimator of the selection equation, and to emerge in the estimation of the SEM via the inverse Mills ratio. The results can be seen in Figure 5.1. The white boxplots correspond to the non-contaminated sample, and the shaded boxplots correspond to the contaminated sample. Letter (c) denotes classical estimator, and (r) denotes robust. It is clear that without contamination both classical and robust estimators perform well. The variability of the robust estimator is a bit higher than that of the classical, which is natural. Under contamination we see that the β_λ coefficients of both equations are seriously

biased towards zero, which will also affect the sample selection bias test. The robust estimator is more stable.

In the second scenario of contamination we study the effect of contamination in the exogenous variable x_1, which also appears as a regressor in the selection equation. The exogenous variable is used twice as a regressor in selection and in 2SLS, and moreover it appears in 2SLS stage via the inverse Mills ratio. The results can be seen in Figure 5.2. Under contamination, the classical estimator clearly breaks down. Obviously the β_λ's are seriously biased, but moreover all the other coefficients except γ_{11} are affected. The robust estimator remains stable except for a little bias in β_λ's.

5.5 Wage Data Application

To illustrate our methodology in practice we consider the example 17.7 in Wooldridge (2002). The data about women's labor force participation consist of 753 observations, with 428 (56.8%) having non-zero wage. In the selection equation we have such explanatory variables as non-wife income (*nwifeinc*), experience (*exper*), squared experience (*expersq*), age, number of children less than 6 years of age (*kidslt6*), and number of children between 6 and 18 years (*kidsge6*). We assume that the variable education (*educ*) is endogenous. The set of instrumental variables consists of mother's education (*motheduc*), father's education (*fatheduc*), and husband's education (*huseduc*). The variable of interest is the wage in a log scale (*lwage*), which is regressed on *educ*, *exper*, *expersq*, and *age*. We use the Heckman-type estimation procedure, which means that the inverse Mills ratio is used as an instrument and as a regressor in the equation of interest. Note, that *educ*, *exper*, *expersq* are moderately correlated with each other (variance inflation factors are between 15 and 25).

Estimation of the data set provides no evidence for the presence of selection bias. In order to check the robustness of this conclusion we perform a sensitivity analysis. For the observations 126 and 348 we change the values of *kidslt6* from 0 to 3. These observations are given by

```
126
 lwage      nwifeinc  exper  expersq  age  kidslt6  kidsge6
 -1.822631  17.90008  17     289      35   3        2

348
```

```
 lwage      nwifeinc  exper  expersq  age  kidslt6  kidsge6
 -1.766677  9.000047  10     100      44   3        2
```

Recall that we have modified only the number of children less than 6 years. Of course having three small kids is not an extreme case, and it can be hardly assigned to be a mistake in the data or a clear outlier.

The results of the classical and robust estimation are the following

```
Call:
lm(formula = lwage ~ st2SEM.MROZ$fitted + exper +
                     expersq + age + INVMILLSRAT)

Residuals:
     Min       1Q   Median       3Q      Max
-2.53481 -0.33563  0.03285  0.38715  2.53239

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.1600032  0.3809733  -0.420  0.67471
educ$fitted         0.0884884  0.0233268   3.793  0.00017 ***
exper               0.0017975  0.0202225   0.089  0.92921
expersq            -0.0001823  0.0004829  -0.377  0.70603
age                 0.0136759  0.0067873   2.015  0.04455 *
INVMILLSRAT        -0.5849879  0.1978288  -2.957  0.00328 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6947 on 422 degrees of freedom
------------------------------------------------------------

Call: rlm(formula = lwage ~ st2SEMrob.MROZC$fitted + exper +
               expersq + age + INVMILLSRAT, weights = xweight)
Residuals:
     Min       1Q   Median       3Q      Max
-2.03521 -0.33401  0.02231  0.34019  2.41157

Coefficients:
                      Value   Std. Error t value
(Intercept)           -0.4743  0.3364    -1.4099
educ$fitted            0.1103  0.0211     5.2285
```

```
exper                0.0425  0.0188     2.2598
expersq             -0.0010  0.0005    -1.9907
age                  0.0015  0.0060     0.2463
INVMILLSRAT         -0.1723  0.1946    -0.8854

Residual standard error: 0.501 on 422 degrees of freedom
-------------------------------------------------------------
```

We can see that the sample selection bias test of the classical estimator becomes significant. The robust estimator is a bit affected, but much less than the classical one. The robust estimator returns the same set of significant variables as without contamination. While by the classical estimator, *exper* and *expersq* become non-significant, and *age* instead becomes significant.

From this example, one can clearly see that the classical estimator is highly sensitive even to a negligible amount of contamination. The robust estimator is more stable, and its use cannot be ignored. We suggest to use it in parallel to the classical estimator as a complementary routine, to indicate possible problems. In the presence of deviations from the model and/or errors in the data, the robust estimator becomes an alternative to the classical ones.

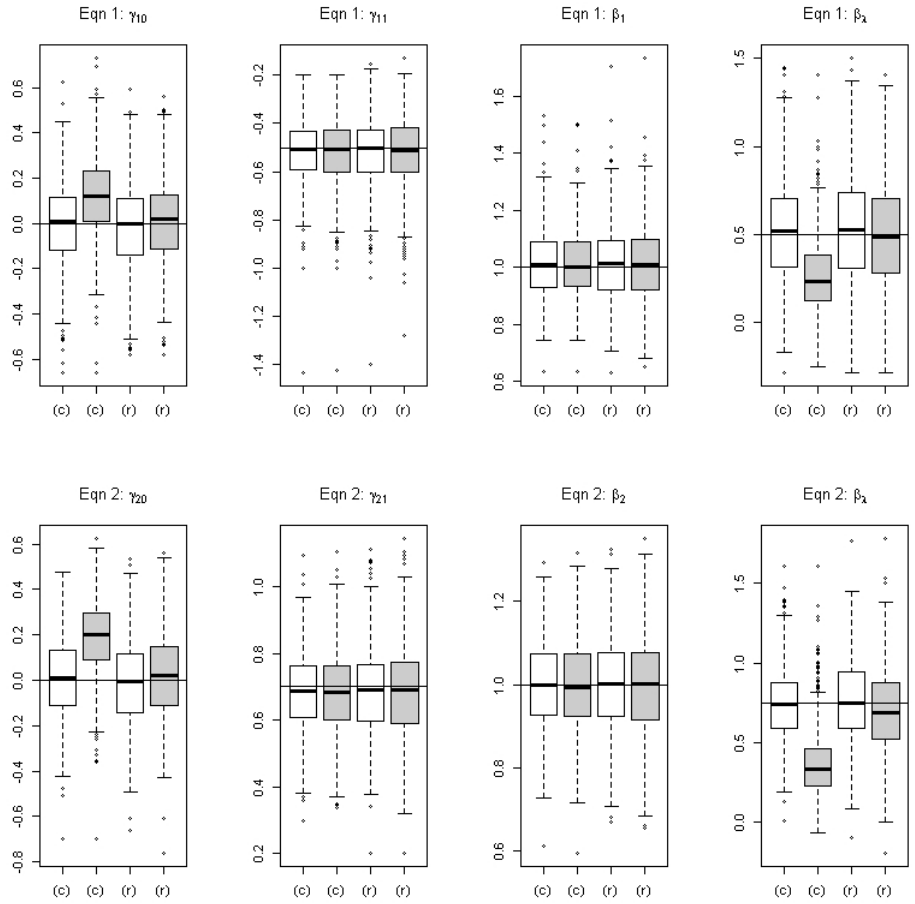

Figure 5.1: *Parameter estimates by classical and robust estimators with contamination of w. The unshaded boxplots correspond to the non-contaminated case, shaded boxplots correspond to the contaminated case. Letter (c) denotes classical estimator, (r) denotes robust estimator. Horizontal lines mark the true values of the parameters.*

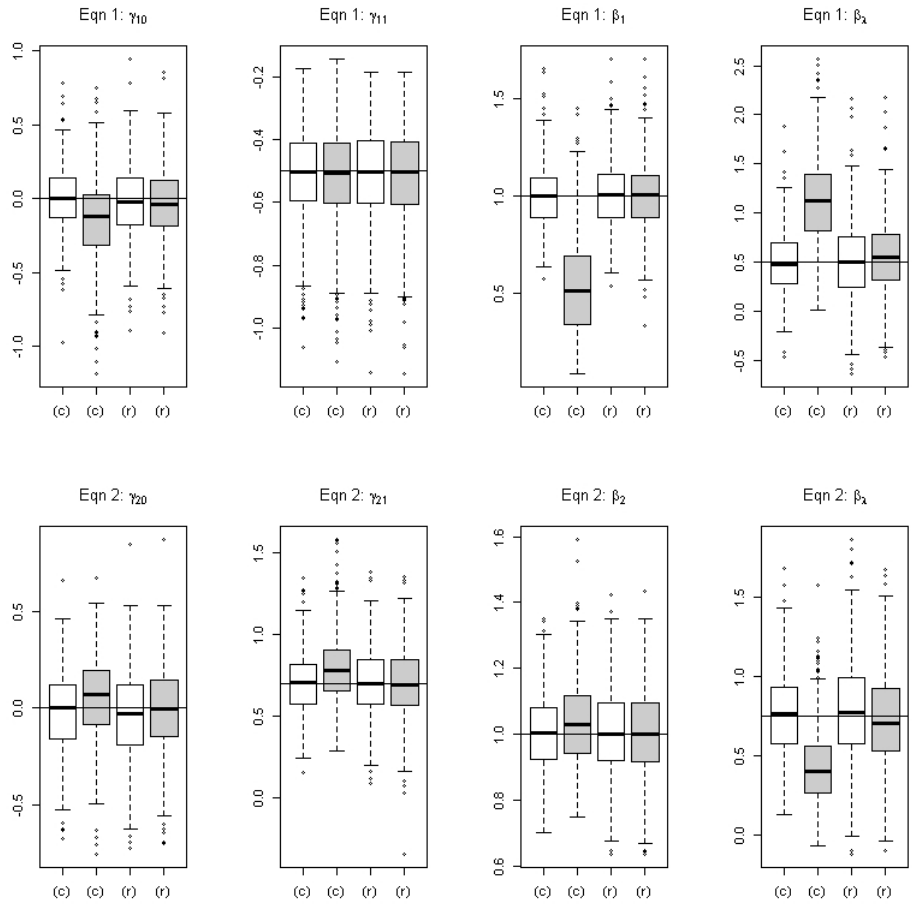

Figure 5.2: *Parameter estimates by classical and robust estimators with contamination of x_1, when it is also used as a regressor in selection equation. The unshaded boxplots correspond to the non-contaminated case, shaded boxplots correspond to the contaminated case. Letter (c) denotes classical estimator, (r) denotes robust estimator. Horizontal lines mark the true values of the parameters.*

Chapter 6

The R Package ssmrob

To facilitate the applications of ideas of this thesis, we provide an R (R Core Team 2012) package for robust estimation and inference in sample selection models. The implementation of the classical estimators of sample selection models in R was made by Toomet and Henningsen (2008), and can be found in package sampleSelection.

The chapter is organized as follows. In Section 6.1 we discuss the implementation of the package. Section 6.2 contains the description of the functions. In Section 6.3 we explain how to use the package. Section 6.4 offers two real data examples.

6.1 Implementation

The main function in the package is ssmrob. It works as a router determining the type of the model and choosing the necessary estimator. In the current version (version 0.2) there are two options: censored Heckman's selection model (Tobit-2) and switching regressions model with probit selection mechanism (Tobit-5). If the Tobit-2 model is chosen then the heckitrob function is called, if the Tobit-5 model is chosen then the heckit5rob is called. More detailed description of the functions with their parameters and options is given in the next section.

The command ssmrob returns the object of class heckitrob or heckit5rob for Tobit-2 or Tobit-5 respectively. The package provides the generic functions for these classes: print function prints the estimation results, summary calculates and prints the summary of estimation with standard errors and t-values of the estimates, coef function extracts the estimated coefficients, vcov function returns the variance-covariance ma-

trices for the two estimation stages, `fitted` function calculates the fitted values, and `residuals` function returns the residuals of the model.

The `summary` function returns the object of class `summary.heckitrob` or `summary.heckit5rob` for Tobit-2 or Tobit-5 respectively. Generic function `vcov` returns two or three variance-covariance matrices, one for selection equation and one or two (depending on the model) for the outcome equation. Functions `fitted` and `residuals` return one or two vectors of fitted values or residuals, also depending on the model (one vector for Tobit-2, two vectors for Tobit-5).

The package also contains several auxiliary functions such as `dLambdadSM`, `dLambdadSM5`, `MmatrM`, `PsiMest`, `x2weight.covMcd`, and `x2weight.robCov`. They are needed for computation of the asymptotic variances and of the robustness weights.

The package is written completely in `R`. It depends on packages `sampleSelection` (Toomet and Henningsen 2008), `robustbase` (Rousseeuw et al. 2012), and `mvtnorm` (Genz et al. 2012). The package `mvtnorm` is required only for the examples based on simulated data, which requires simulation from the multivariate normal distribution. All these packages are available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/`.

6.2 Description of the Functions

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

This function is a router, depending on the type of parameters in `selection` and `outcome`, chooses the model and calls the corresponding estimator. Parameter `outcome` is a simple formula for the case of truncated selection model, or a list of two formulas for the case of switching regressions. Parameter `selection` is a formula for the selection equation. Parameter `control` defines the accuracy and the robustness tuning parameters, see the description of `heckitrob.control` function below.

The default method is the two-stage M-estimator of Huber's type, i.e. without leverage weights.

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

Tuning parameters for the robust two-stage Mallows-type M-estimator. Parameters `acc` and `test.acc` control for the accuracy of estimation. Maximum number of iterations is defined by `maxit`.

The leverage weights for the first and the second stage estimators are defined by `weights.x1` and `weights.x2` respectively. If `none` is chosen then the weights are equal to 1. If `hat` is chosen, then weights on the design of the form $\sqrt{1-h_{ii}}$ are used, where h_{ii} are the diagonal elements of the hat matrix. If `robCov` is chosen, then weights based on the robust Mahalanobis distance of the design matrix are used, where the covariance matrix is estimated by the `rob.cov` method from package `MASS` (Venables and Ripley 2002) using the minimum volume ellipsoid estimator (Rousseeuw 1985). Similarly, if `covMcd` is chosen, but the covariance is estimated by minimum covariance determinant estimator (Rousseeuw and Van Driessen 1999). However, it should be noted, that the use of robust Mahalanobis distance can be limited in some specific cases, e.g. binary data (Hubert and Rousseeuw 1997, Maronna and Yohai 2000). In such cases, the hat matrix based weights can be used to avoid numerical problems.

Parameters `tcc` and `t.c` are the tuning constants for the Huber-functions of the first and second stage estimators respectively.

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

Function presents the robust two-stage estimator of the simple selection model (Tobit-2). Parameters `outcome` and `selection` must be formulas. Note that, if the Huber tuning parameters are large and the leverage weights are ones, then the estimator converges to the classical Heckman's two-stage estimator.

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

Similarly to the previous function, but `heckit5rob` estimates the switching regressions model.

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

6.3 Using the ssmrob Package

In this section we provide the illustrative simulation experiments. We demonstrate the usage of the `ssmrob` function.

6.3.1 Tobit-2 Model

First, we simulate the data:

```
R> library(mvtnorm)
R> N=1000
R> covmat = matrix(c(1,0.75,0.75,1),2,2)
R> eps = rmvnorm(N, mean =rep(0,2), sigma=covmat)
R> x1 <- rnorm(N)
R> y1 <- x1 + eps[,1] > 0
R> x2 <- rnorm(N)
R> y2=ifelse(y1 > 0.5, x2 + eps[,2], 0)
```

We set the sample size equal to 1000. The errors are generated from a bivariate normal distribution with correlation equal to 0.75. Then we generate the explanatory variables (`x1` and `x2`) following a standard normal distribution independently from each other. One could use the same variable for the selection equation and for the outcome equation, but we would like to study the robustness problem, and leave the exclusion restriction issue beyond the scope of this work. Reader, interested in this issue, can consult Toomet and Henningsen (2008). And finally, using the explanatory variables and the errors we compute the response variables (`y1` and `y2`). The data generated from the explained procedure is not contaminated, and from the output below one can see that the estimator is close to the true parameters.

```
R> summary(ssmrob(y2 ~ x2, y1 ~ x1), control =
R>         heckitrob.control(weights.x1 = "robCov",
                             weights.x2 = "covMcd"))

------------------------------------------------------------
Robust 2-step Heckman / heckit M-estimation
Probit selection equation:
               Estimate  Std.Error t-value  p-value
(Intercept)   0.05280394 0.04704872   1.122 2.62e-01
x1            1.03283217 0.06905866  14.960 1.43e-50 ***
```

```
Outcome equation:
                  Estimate  Std.Error  t-value   p-value
(Intercept)   -0.004639395 0.07063611 -0.06568  9.48e-01
x2             0.907592366 0.03674991 24.70000 1.17e-134 ***
IMR            0.810391013 0.09893104  8.19100  2.58e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In order to test for the robustness of the estimator we introduce the contamination. With probability 0.01 we generate outliers from the degenerate distribution putting mass one at the point $(x_1, y_1, x_2, y_2) = (-6, 1, 1, 1)$. It generates leverage outliers in the selection stage, which is the same as the contamination of Case A in Section 4.3.

```
R> uni=runif(N,0,1)
R> for(i in 1:N)
R>  if(uni[i]<0.01) {x1[i]=-6; y1[i]=1; x2[i]=1; y2[i]=1}
```

The results of the estimation are the following:

```
R> summary(ssmrob(y2 ~ x2, y1 ~ x1, control =
R>  heckitrob.control(weights.x1 = "robCov",
                     weights.x2 = "covMcd")))

------------------------------------------------------------
Robust 2-step Heckman / heckit M-estimation
Probit selection equation:
                Estimate  Std.Error t-value  p-value
(Intercept)   0.06232741 0.04725277   1.319 1.87e-01
x1            1.02487851 0.06916874  14.820 1.14e-49 ***
Outcome equation:
                Estimate  Std.Error t-value   p-value
(Intercept)   0.05061578 0.07107007  0.7122  4.76e-01
x2            0.90783347 0.03726999 24.3600 4.74e-131 ***
IMR           0.70443818 0.10097710  6.9760  3.03e-12 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is clear, that the estimator allows some bias, but it is controlled. It can be made smaller by decreasing the tuning constants. For comparison we present the output of the classical estimator from the `sampleSelection` package.

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

--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
1000 observations (463 censored and 537 observed)
7 free parameters (df = 994)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12687    0.04172   3.041  0.00242 **
x1           0.42344    0.03467  12.212  < 2e-16 ***
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.23671    0.08207   2.884  0.00401 **
x2           0.88779    0.03809  23.306  < 2e-16 ***
Multiple R-Squared:0.5185,Adjusted R-Squared:0.5167
Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio   0.3258     0.1097   2.971  0.00304 **
sigma           0.9422         NA      NA       NA
rho             0.3458         NA      NA       NA
---
Signif. codes:  0 ë***í 0.001 ë**í 0.01 ë*í 0.05 ë.í 0.1 ë í 1
```

The estimator of the inverse Mills ratio drops to 0.326, while with the robust estimator it is 0.704, which is much closer to the true value of 0.75. Note, that we can obtain the classical estimates using the `ssmrob` function by using large values, e.g. 1000, of the tuning parameters `tcc` and `t.c`, and setting the leverage weights `weights.x1` and `weights.x2` equal to `'none'`.

6.3.2 Tobit-5 Model

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

```
R> library(mvtnorm)
R> covm <- diag(3)
R> covm[lower.tri(covm)] <- c(0.75, 0.5, 0.25)
R> covm[upper.tri(covm)] <- covm[lower.tri(covm)]
```

```
R> eps <- rmvnorm(1000, rep(0, 3), covm)
R> x1 <- rnorm(1000)
R> y1 <- x1 + eps[,1] > 0
R> x21 <- rnorm(1000)
R> x22 <- rnorm(1000)
R> y2=ifelse(y1 > 0.5, x21 + eps[,2], x22 + eps[,3])
```

The DGP is similar to the tobit-2 case, but with minor modifications. We generate two explanatory variables for the outcome equation, namely `x21` and `x22` for the first and second regimes respectively. The error terms follow a trivariate normal distribution. The response variable (`y2`) in the outcome equation has two regimes, depending on the selection variable `y1`. Without contamination we have the following output:

```
R> summary(ssmrob(list(y2 ~ x21, y2 ~ x22), y1 ~ x1,
R>  control = heckitrob.control(weights.x1 = "robCov",
R>                              weights.x2 = "covMcd")))

-----------------------------------------------------------------
Robust 2-step Heckman / heckit M-estimation
Probit selection equation:
                Estimate  Std.Error t-value  p-value
(Intercept)   -0.04487983 0.04717327 -0.9514 3.41e-01
x1             0.99888413 0.06514231 15.3300 4.54e-53 ***
Outcome equation, regime 1:
                Estimate  Std.Error t-value   p-value
(Intercept)    0.1207170 0.07596463   1.589  1.12e-01
x21            1.0036156 0.04215398  23.810 2.74e-125 ***
IMR1           0.6566172 0.09779090   6.715  1.89e-11 ***
Outcome equation, regime 2:
                  Estimate  Std.Error t-value   p-value
(Intercept)    -0.05188555 0.07260708 -0.7146  4.75e-01
x22             1.05217058 0.04251854 24.7500 3.41e-135 ***
IMR2            0.54288563 0.10716225  5.0660  4.06e-07 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

equation of interest in the second regime. This type of contamination corresponds to the Case B for the first regime and to the Case A for the second regime.

```
R> uni=runif(N,0,1)
R> for(i in 1:N)
R>   if(uni[i]<0.01) {x1[i]=6; y1[i]=0; x22[i]=1; y2[i]=1}
```

We estimate the contaminated sample and obtain the following output:

```
R> summary(ssmrob(list(y2 ~ x21, y2 ~ x22), y1 ~ x1,
R>  control = heckitrob.control(weights.x1 = "robCov",
R>                             weights.x2 = "covMcd")))

------------------------------------------------------------
Robust 2-step Heckman / heckit M-estimation
Probit selection equation:
                 Estimate  Std.Error t-value  p-value
(Intercept)   -0.05226743 0.04734428  -1.104 2.70e-01
x1             0.99741226 0.06534541  15.260 1.33e-52 ***
Outcome equation, regime 1:
                Estimate  Std.Error t-value   p-value
(Intercept)    0.1128210 0.07740272   1.458  1.45e-01
x21            1.0006546 0.04235204  23.630 2.03e-123 ***
IMR1           0.6613017 0.09943004   6.651  2.91e-11 ***
Outcome equation, regime 2:
                 Estimate  Std.Error t-value   p-value
(Intercept)    -0.0809203 0.07117386  -1.137  2.56e-01
x22             1.0516439 0.04197831  25.050 1.66e-138 ***
IMR2            0.4840658 0.10523967   4.600  4.23e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

```
------------------------------------------------------------
Robust 2-step Heckman / heckit M-estimation
Probit selection equation:
               Estimate  Std.Error t-value  p-value
(Intercept)   -0.1076174 0.04206950  -2.558 1.05e-02   *
x1             0.5705124 0.04682784  12.180 3.82e-34 ***
Outcome equation, regime 1:
                Estimate  Std.Error t-value   p-value
(Intercept)    -0.1970618 0.13091579  -1.505  1.32e-01
x21             1.0042791 0.04183671  24.000 2.48e-127 ***
IMR1            1.0339217 0.16239044   6.367  1.93e-10 ***
Outcome equation, regime 2:
                Estimate  Std.Error t-value   p-value
(Intercept)    -0.1761809 0.07785650  -2.263  2.36e-02   *
x22             1.0721582 0.04289107  25.000 6.55e-138 ***
IMR2            0.2621362 0.10677171   2.455  1.41e-02   *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimators of both regimes are seriously affected by the contamination. For the robust estimator the inverse Mills ratio coefficients are 0.661 and 0.484, and for the classical estimator they are 1.034 and 0.262. Recall that the true values are 0.75 and 0.5.

6.4 Examples

In this section we demonstrate how the package can be used with the real data. We examine two datasets already analyzed in literature. In the first example there is no robustness problem, and the results of estimation by classical and robust procedures are close. In the second example, on the contrary, the results of estimation by the classical and robust estimators are different, which indicates the robustness problem.

6.4.1 Wage Offer Data

The first dataset is an example from Wooldridge (2002). We consider the Example 17.6 (p. 565) about the wage offer for married women, with potential selectivity bias into the labor force. The dataset consists of

753 observations, with 325 (43.2%) truncated observations. The selection equation defining the labor force participation includes the following variables as age, education status (*educ*), non-wife income (*nwifeinc*), experience (*exper*), squared experience (*expersq*), number of children less than 6 years of age (*kidslt6*), and number of children greater than 6 years of age (*kidsge6*). In the equation of interest the log-wage offer depends on education, experience, and squared experience. Using the package `sampleSelection` we obtain the following output.

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

```
sigma          0.66363         NA      NA       NA
rho            0.04861         NA      NA       NA
--------------------------------------------
```

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

```
R> summary(ssmrob(outcomeEq, selectEq))

> summary(heckrob.MROZ)
------------------------------------------------------------
Robust 2-step Heckman / heckit M-estimation
Probit selection equation:
                   Estimate    Std.Error t-value  p-value
(Intercept)    0.185085844 0.5215843100  0.3549 7.23e-01
nwifeinc      -0.013812287 0.0051413318 -2.6870 7.22e-03  **
educ           0.131746879 0.0263491696  5.0000 5.73e-07 ***
exper          0.123029123 0.0192493260  6.3910 1.64e-10 ***
expersq       -0.001905781 0.0006133654 -3.1070 1.89e-03  **
age           -0.050790058 0.0087215189 -5.8240 5.76e-09 ***
kidslt6       -0.840732688 0.1223745428 -6.8700 6.41e-12 ***
kidsge6        0.039740117 0.0453318208  0.8766 3.81e-01
Outcome equation:
                    Estimate    Std.Error t-value  p-value
(Intercept) -0.4720491368 0.2595473904 -1.8190 6.90e-02   .
educ         0.1114265409 0.0132375403  8.4170 3.85e-17 ***
exper        0.0366974402 0.0134697652  2.7240 6.44e-03  **
expersq     -0.0007015977 0.0003696587 -1.8980 5.77e-02   .
IMR         -0.0495792781 0.1338459784 -0.3704 7.11e-01
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
------------------------------------------------------------
```

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

6.4.2 Ambulatory Expenditures Data

The second example is an example considered in Section 4.4. The results of the estimation obtained using the `R` package `sampleSelection` (Toomet and Henningsen 2008) are:

```
R> data(MEPS2001)
R> attach(MEPS2001)
R> selectEq <- dambexp ~ age + female + educ + blhisp +
                         totchr + ins
R> outcomeEq <- lnambx ~ age + female + educ + blhisp +
                         totchr + ins
R> summary(selection(selectEq, outcomeEq, method="2step"))
--------------------------------------------------------------
2-step Heckman / heckit estimation
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.71771    0.19247  -3.729 0.000195 ***
age          0.09732    0.02702   3.602 0.000320 ***
female       0.64421    0.06015  10.710  < 2e-16 ***
educ         0.07017    0.01134   6.186 6.94e-10 ***
blhisp      -0.37449    0.06175  -6.064 1.48e-09 ***
totchr       0.79352    0.07112  11.158  < 2e-16 ***
ins          0.18124    0.06259   2.896 0.003809 **
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.30257    0.29414  18.028  < 2e-16 ***
age          0.20212    0.02430   8.319  < 2e-16 ***
female       0.28916    0.07369   3.924 8.89e-05 ***
educ         0.01199    0.01168   1.026    0.305
blhisp      -0.18106    0.06585  -2.749    0.006 **
totchr       0.49833    0.04947  10.073  < 2e-16 ***
ins         -0.04740    0.05315  -0.892    0.373
Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio  -0.4802     0.2907  -1.652   0.0986 .
sigma           1.2932         NA      NA       NA
rho            -0.3713         NA      NA       NA
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
------------------------------------------------------------
```

Using the robust two-stage estimator we obtained:

```
R> summary(ssmrob(outcomeEq, selectEq,
    control=heckitrob.control(tcc=3.2, weights.x1="robCov")))
------------------------------------------------------------
Robust 2-step Heckman / heckit M-estimation
Probit selection equation:
                  Estimate  Std.Error t-value  p-value
(Intercept)    -0.74914476 0.19506999  -3.840 1.23e-04 ***
age             0.10541500 0.02769588   3.806 1.41e-04 ***
female          0.68740832 0.06225762  11.040 2.41e-28 ***
educ            0.07011568 0.01146521   6.116 9.62e-10 ***
blhisp         -0.39774532 0.06264878  -6.349 2.17e-10 ***
totchr          0.83283613 0.08027772  10.370 3.24e-25 ***
ins             0.18256005 0.06371471   2.865 4.17e-03  **
Outcome equation:
                 Estimate  Std.Error t-value  p-value
(Intercept)    5.40154264 0.27672891  19.520 7.53e-85 ***
age            0.20061658 0.02450765   8.186 2.70e-16 ***
female         0.25501033 0.06992954   3.647 2.66e-04 ***
educ           0.01324867 0.01161609   1.141 2.54e-01
blhisp        -0.15508435 0.06506654  -2.383 1.72e-02   *
totchr         0.48115830 0.03822948  12.590 2.52e-36 ***
ins           -0.06706633 0.05159205  -1.300 1.94e-01
IMR           -0.67676033 0.25927579  -2.610 9.05e-03  **
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
------------------------------------------------------------
```

In this case there is clear evidence, that the distributional assumptions are violated. The classical estimator produces underestimated inverse Mills ratio and leads to conclusion of no selection bias. The robust estimator is more reliable and should be preferred in such situations.

Chapter 7

Discussion and Conclusion

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 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 we focused on the basic sample selection model, our methodology can be easily extended to more complex models. This was done in the end of Chapter 4 for the switching regressions model, and in Chapter 5 for the case of simultaneous equations models with selectivity. Other important extensions include problems with different censoring rules, such as ordered or multinomial models, and models with multiple selection rules. Another interesting class of problems are the models with non-additive errors. Leaving these issues beyond the scope of this work we believe that these problems will be addressed in future research.

The results in Chapter 3 provide a more general framework than that in the selection models. We explore the robust estimation and inference issues in two-stage models. There, we presented three simple examples of how our approach can be used. Certainly, there are many other possible situations where the robust two-stage procedures are useful. In particular the results for time series can be extended to spatial statistics. Also, we should note that others than sample selection models empirical examples in modern economics are based on latent two-stage procedures, e.g. series of regressions, use of composite indexes as variables, and so on. The fields of economics and social sciences are growing and becoming more and more complex involving more sophisticated models and estimation techniques.

Given the increasing complexity of the data, the use of robust analysis cannot be neglected.

The proposed robust estimators and tests can be used not only as alternatives to the classical procedures, but also as complements to them. One can use classical and robust analysis in parallel. If there is no robustness problem, then both procedures return similar results (see wage offer example in Chapter 6), and the classical methods can be used. But if the results are different, then it is a flag that a more careful analysis is required, or that the robust methods should be preferred (see ambulatory expenditures example). The importance of robust procedures in the analysis of real data has been proven many times in the literature. We believe that the methodology presented in this thesis will be useful for practitioners in various fields of science.

Appendix A

Technical Derivations

A.1 IF of the probit MLE

In general, probit model belongs to the class of generalized linear models (GLM), see McCullagh and Nelder (1989). And the inverse of the normal cdf is used as the link function in GLM when the distribution of the response variable is binomial, i.e. when the conditional expectation of the response variable y given set of explanatory variables x is modeled, we have

$$E(y|x) = g^{-1}(x^T\beta) = \Phi(x^T\beta),$$

where g denotes the link function, and β is a vector of parameters. This link is not canonical for the binomial distribution, the canonical link is logit.

The use of probit model instead of logit is motivated by the formulation via the latent underlying process. Assume that there is an unobservable response variable y^* defined by the following regression

$$y^* = x^T\beta + e,$$

the observed variable is

$$y = \begin{cases} y = 1 & \text{if } y^* > 0, \\ y = 0 & \text{if } y^* \leq 0, \end{cases}$$

where e is normally distributed random noize, which provides the connection with the use of the normal cdf as the link. For more details about probit model see Maddala (1983).

The log-likelihood is given by:

$$\sum_{i=1}^{N} \big[y \log\{\Phi(x^T\beta)\} + (1-y) \log\{1-\Phi(x^T\beta)\} \big]. \tag{A.1}$$

It is well known (see Hampel et al. 1986) that, MLE belongs to the class of M-estimators and has the following form of the IF:

$$IF(z; T, F) = M(\Psi, F)^{-1} \Psi(z; T(F)), \tag{A.2}$$

where Ψ is the derivative of the log-likelihood with respect to β, matrix M is given by

$$M(\Psi, F) = -\int \left\{ \frac{\partial}{\partial \beta} \Psi(z; \beta) \right\} dF(z), \tag{A.3}$$

and z denotes the datum (y, x).

We only need to specify Ψ and M for this particular case. Taking the derivative of (A.1) with respect to β we obtain

$$\Psi(z; \beta) = \frac{y - \Phi(x^T\beta)}{\Phi(x^T\beta)\{1-\Phi(x^T\beta)\}} \phi(x^T\beta) x, \tag{A.4}$$

and M is given by

$$M(\Psi, F) = \int \left[\frac{\phi(x^T\beta)^2}{\Phi(x^T\beta)\{1-\Phi(x^T\beta)\}} \right] x x^T dF(z). \tag{A.5}$$

The influence function of the probit MLE is unbounded, which means that the estimator is not robust. The plot of the IF can be seen in Figure A.1. The solid lines represent the IF, and the sequences of boxplots correspond to the standardized biases of the estimators. The top panel is the case when x varies from -6 to 6 and the corresponding $y = 1$. In the bottom panel the corresponding $y = 0$.

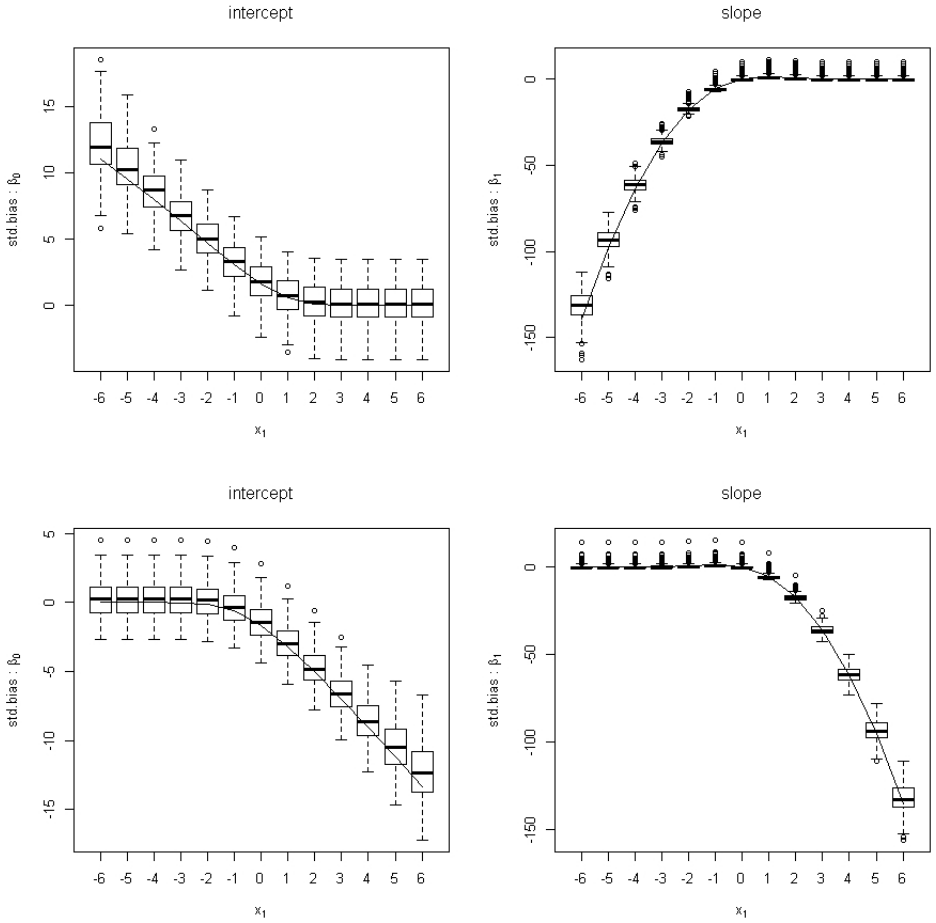

Figure A.1: *IF of the probit MLE. The solid lines represent the IF's. The boxplots represent the standardized biases of the estimators of the simulated samples. We vary x from -6 to 6, and in the top panel the corresponding $y = 1$ and in the bottom panel the corresponding $y = 0$.*

A.2 IF of the Heckman's two-stage estimator

Proof of Proposition 1. The Heckman's estimator belongs to the class of two-stage M-estimators and satisfies the conditions required in Zhelonkin et al. (2012). The IF of the general M-estimator has the following form:

$$IF(z;T,F) = M(\Psi_2)^{-1}\Bigg(\Psi_2[(x_2,y_2);h\{(x_1,y_1);S(F)\},T(F)] + \int \frac{\partial}{\partial\theta}\Psi_2\{(x_2,y_2);\theta,T(F)\}\frac{\partial}{\partial\eta}h\{(x_1,y_1);\eta\}dFIF(z;S,F)\Bigg), \tag{A.6}$$

where $M(\Psi_2) = -\int \frac{\partial}{\partial\xi}\Psi_2[(x_2,y_2);h\{(x_1,y_1);S(F)\},\xi]dF$ and $IF(z;S,F)$ denotes the IF of the first stage. Note that in (A.6) T and S denote general M-estimators.

To find the IF of the Heckman's estimator we need to identify the derivatives in (A.6). The $h(\cdot;\cdot)$ function is the inverse Mills ratio and its derivative with respect to β_1 is:

$$\lambda' = \frac{\partial}{\partial\eta}\lambda\{(x_1,y_1);\eta\} = \frac{-\Phi(x_1^T\beta_1)\phi(x_1^T\beta_1)x_1^T\beta_1 - \phi(x_1^T\beta_1)^2}{\Phi(x_1^T\beta_1)^2}x_1^T. \tag{A.7}$$

The score function of the second stage is

$$\Psi_2[(x_2,y_2);\lambda\{(x_1,y_1);S(F)\},T(F)] = (y_2 - x_2^T\beta_2 - \lambda\beta_\lambda)\begin{pmatrix} x_2 \\ \lambda \end{pmatrix} y_1. \tag{A.8}$$

The $M(\Psi_2)$ matrix is given by

$$-\int\left(-\frac{\partial}{\partial\xi}\Psi_2[(x_2,y_2);\lambda\{(x_1,y_1);S(F)\},\xi]\right)dF = \int\begin{pmatrix} x_2x_2^T & \lambda x_2 \\ \lambda x_2^T & \lambda^2 \end{pmatrix} y_1 dF. \tag{A.9}$$

The derivative with respect to the second parameter of Ψ_2 is

$$\frac{\partial}{\partial\theta}\Psi_2\{(x_2,y_2);\theta,T(F)\} = \begin{pmatrix} 0 \\ y_2 - x_2^T\beta_2 - \lambda\beta_\lambda \end{pmatrix} y_1 - \begin{pmatrix} x_2 \\ \lambda \end{pmatrix}\beta_\lambda y_1. \tag{A.10}$$

After taking the expectation, the first term of the righthand side of (A.10) becomes zero. The $IF(z; S, F) = M(\Psi_1)^{-1}\Psi_1\{(x_1, y_1); S(F)\}$ is the IF of the MLE estimator for probit regression, where:

$$M(\Psi_1) = \int \left[\frac{\phi(x_1^T\beta_1)^2}{\Phi(x_1^T\beta_1)\{1-\Phi(x_1^T\beta_1)\}}\right] x_1 x_1^T dF, \tag{A.11}$$

$$\Psi_1\{(x_1, y_1); S(F)\} = \{y_1 - \Phi(x_1^T\beta_1)\}\frac{\phi(x_1^T\beta_1)}{\Phi(x_1^T\beta_1)\{1-\Phi(x_1^T\beta_1)\}} x_1. \tag{A.12}$$

Inserting the expressions in formulas (A.7)-(A.12) in (A.6) we obtain the IF of Heckman's estimator. □

A.3 CVF of the one-stage M-estimator

The particular cases of the CVF for the location model and for the linear regression framework were explored in Hampel et al. (1986). Here we need to obtain the expression of CVF for general two-stage M-estimator. In order to do this we first derive the CVF for general one-stage M-estimator. Consider the equation (3.1), for simplicity of notation we omit the subscript denoting the estimation stage. Suppose that the contaminating distribution is $F_\epsilon = (1-\epsilon)F + \epsilon\Delta_z$, then under F_ϵ the asymptotic variance is

$$V_\epsilon = M_\epsilon^{-1} \int \Psi\{z; S(F_\epsilon)\}\Psi\{z; S(F_\epsilon)\}^T dF_\epsilon M_\epsilon^{-1} \tag{A.13}$$

We need to compute the derivative of V_ϵ with respect to ϵ at $\epsilon = 0$.

$$\left.\frac{\partial V_\epsilon}{\partial \epsilon}\right|_{\epsilon=0} = \left.\frac{\partial}{\partial \epsilon} M_\epsilon^{-1}\right|_{\epsilon=0} Q M^{-1} + M^{-1}\left.\frac{\partial}{\partial \epsilon} Q_\epsilon\right|_{\epsilon=0} M^{-1} + M^{-1} Q \left.\frac{\partial}{\partial \epsilon} M_\epsilon^{-1}\right|_{\epsilon=0}, \tag{A.14}$$

where

$$M_\epsilon = -\int \frac{\partial}{\partial \theta}\Psi(z;\theta) dF_\epsilon, \tag{A.15}$$

note that $\Psi\{z; S(F_\epsilon)\}$ depends on ϵ through the estimator $S(F_\epsilon)$, and

$$Q_\epsilon = \int \Psi\{z; S(F_\epsilon)\}\Psi\{z; S(F_\epsilon)\}^T dF_\epsilon. \tag{A.16}$$

For simplicity of notation denote $\Psi_\epsilon = \Psi\{z; S(F_\epsilon)\} = \begin{pmatrix} \Psi_{1\epsilon} & \Psi_{2\epsilon} & \dots & \Psi_{p\epsilon} \end{pmatrix}^T$ and $\Psi = \Psi\{z; S(F)\} = \begin{pmatrix} \Psi_1 & \Psi_2 & \dots & \Psi_p \end{pmatrix}^T$.

$$\begin{aligned}\left.\frac{\partial V_\epsilon}{\partial \epsilon}\right|_{\epsilon=0} =& - M^{-1}\left(\int \frac{\partial}{\partial\epsilon}\frac{\partial}{\partial\theta}\Psi_\epsilon dF\right|_{\epsilon=0} - M + \frac{\partial}{\partial\theta}\Psi\Big) M^{-1}QM^{-1} \\ &+ M^{-1}\left(\int \frac{\partial}{\partial\epsilon}\Psi_\epsilon\Psi_\epsilon^T dF\right|_{\epsilon=0} - \int \Psi\Psi^T dF + \Psi\Psi^T\Big) M^{-1} \\ &- M^{-1}QM^{-1}\left(\int \frac{\partial}{\partial\epsilon}\frac{\partial}{\partial\theta}\Psi_\epsilon dF\right|_{\epsilon=0} - M + \frac{\partial}{\partial\theta}\Psi\Big) M^{-1}.\end{aligned} \tag{A.17}$$

Note that the matrix in brackets in the third line is symmetric and is the same as the one in the first line.

Now we need to find the derivatives in (A.17). Clear that

$$\frac{\partial\Psi}{\partial\theta} = \begin{pmatrix} \frac{\partial\Psi_1}{\partial\theta_1} & \dots & \frac{\partial\Psi_1}{\partial\theta_p} \\ \frac{\partial\Psi_2}{\partial\theta_1} & \dots & \frac{\partial\Psi_2}{\partial\theta_p} \\ & \dots & \\ \frac{\partial\Psi_p}{\partial\theta_1} & \dots & \frac{\partial\Psi_p}{\partial\theta_p} \end{pmatrix}.$$

Every element of this matrix is a scalar. Then the derivative of this matrix with respect to ϵ must be the matrix of the same dimension. And the derivative in the first and third lines of (A.17) is

$$\begin{aligned}\left.\frac{\partial}{\partial\epsilon}\frac{\partial}{\partial\theta}\Psi_\epsilon\right|_{\epsilon=0} &= \left.\begin{pmatrix} \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{1\epsilon}}{\partial\theta_1} & \dots & \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{1\epsilon}}{\partial\theta_p} \\ \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{2\epsilon}}{\partial\theta_1} & \dots & \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{2\epsilon}}{\partial\theta_p} \\ & \dots & \\ \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{p\epsilon}}{\partial\theta_1} & \dots & \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{p\epsilon}}{\partial\theta_p} \end{pmatrix}\right|_{\epsilon=0} \\ &= \begin{pmatrix} \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_1}{\partial\theta_1}\right)^T IF(z;S,F) & \dots & \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_1}{\partial\theta_p}\right)^T IF(z;S,F) \\ \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_2}{\partial\theta_1}\right)^T IF(z;S,F) & \dots & \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_2}{\partial\theta_p}\right)^T IF(z;S,F) \\ & \dots & \\ \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_p}{\partial\theta_1}\right)^T IF(z;S,F) & \dots & \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_p}{\partial\theta_p}\right)^T IF(z;S,F) \end{pmatrix} \\ &= D.\end{aligned} \tag{A.18}$$

The derivative in the second line of (A.17) is

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\Psi_\epsilon\Psi_\epsilon^T\Big|_{\epsilon=0} &= \left(\begin{array}{ccc} \frac{\partial}{\partial\epsilon}\Psi_{1\epsilon}^2 & \dots & \frac{\partial}{\partial\epsilon}\Psi_{1\epsilon}\Psi_{p\epsilon} \\ \frac{\partial}{\partial\epsilon}\Psi_{2\epsilon}\Psi_{1\epsilon} & \dots & \frac{\partial}{\partial\epsilon}\Psi_{2\epsilon}\Psi_{p\epsilon} \\ & \dots & \\ \frac{\partial}{\partial\epsilon}\Psi_{p\epsilon}\Psi_{1\epsilon} & \dots & \frac{\partial}{\partial\epsilon}\Psi_{p\epsilon}^2 \end{array}\right)\Bigg|_{\epsilon=0} \\
&= \left(\begin{array}{ccc} \left(\frac{\partial}{\partial\theta}\Psi_1\right)^T IF(z;S,F)\Psi_1 & \dots & \left(\frac{\partial}{\partial\theta}\Psi_1\right)^T IF(z;S,F)\Psi_p \\ \left(\frac{\partial}{\partial\theta}\Psi_2\right)^T IF(z;S,F)\Psi_1 & \dots & \left(\frac{\partial}{\partial\theta}\Psi_2\right)^T IF(z;S,F)\Psi_p \\ & \dots & \\ \left(\frac{\partial}{\partial\theta}\Psi_p\right)^T IF(z;S,F)\Psi_1 & \dots & \left(\frac{\partial}{\partial\theta}\Psi_p\right)^T IF(z;S,F)\Psi_p \end{array}\right) \\
&\quad + \left(\begin{array}{ccc} \Psi_1\left(\frac{\partial}{\partial\theta}\Psi_1\right)^T IF(z;S,F) & \dots & \Psi_1\left(\frac{\partial}{\partial\theta}\Psi_p\right)^T IF(z;S,F) \\ \Psi_2\left(\frac{\partial}{\partial\theta}\Psi_1\right)^T IF(z;S,F) & \dots & \Psi_2\left(\frac{\partial}{\partial\theta}\Psi_p\right)^T IF(z;S,F) \\ & \dots & \\ \Psi_p\left(\frac{\partial}{\partial\theta}\Psi_1\right)^T IF(z;S,F) & \dots & \Psi_p\left(\frac{\partial}{\partial\theta}\Psi_p\right)^T IF(z;S,F) \end{array}\right) \\
&= \left(\begin{array}{c} \left(\frac{\partial}{\partial\theta}\Psi_1\right)^T IF(z;S,F) \\ \left(\frac{\partial}{\partial\theta}\Psi_2\right)^T IF(z;S,F) \\ \vdots \\ \left(\frac{\partial}{\partial\theta}\Psi_p\right)^T IF(z;S,F) \end{array}\right)\left(\begin{array}{cccc} \Psi_1 & \Psi_2 & \dots & \Psi_p \end{array}\right) \\
&\quad + \left\{\left(\begin{array}{c} \left(\frac{\partial}{\partial\theta}\Psi_1\right)^T IF(z;S,F) \\ \left(\frac{\partial}{\partial\theta}\Psi_2\right)^T IF(z;S,F) \\ \vdots \\ \left(\frac{\partial}{\partial\theta}\Psi_p\right)^T IF(z;S,F) \end{array}\right)\left(\begin{array}{cccc} \Psi_1 & \Psi_2 & \dots & \Psi_p \end{array}\right)\right\}^T \\
&= \left(\frac{\partial}{\partial\theta}\Psi\right) IF(z;S,F)\Psi^T + \left\{\left(\frac{\partial}{\partial\theta}\Psi\right) IF(z;S,F)\Psi^T\right\}^T \\
&= \left(\frac{\partial}{\partial\theta}\Psi\right)\Psi IF(z;S,F)^T + IF(z;S,F)\Psi^T\left(\frac{\partial}{\partial\theta}\Psi\right) \quad \text{(A.19)}
\end{aligned}$$

Finally, we get

$$\begin{aligned}
CVF(z;S,F) =& V - M^{-1}\left(\int DdF + \frac{\partial}{\partial\theta}\Psi\right)V - V\left(\int DdF + \frac{\partial}{\partial\theta}\Psi\right)M^{-1} \\
&+ M^{-1}\left(\int RdF + \int R^T dF + \Psi\Psi^T\right)M^{-1}, \qquad \text{(A.20)}
\end{aligned}$$

where

$$D_{ij} = \left(\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta_j}\Psi_i\right)^T IF(z;S,F),$$

$$R = \left(\frac{\partial}{\partial\theta}\Psi\right) IF(z; S, F)\Psi^T,$$

and V is the variance at the model without contamination.

A.4 CVF of the two-stage M-estimator

Assume that we need to estimate q parameters in the first stage, p parameters in the second stage and we construct k variables $h(z^{(1)}; S(F))$, such that $k < p$. Then variance-covariance matrix of the two-stage estimator is $p \times p$ dimensional.

In order to obtain the CVF of the two-stage M-estimator we need to find the following derivative

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\mathrm{V}_\epsilon\Big|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\left(M_\epsilon^{-1}Q_\epsilon^{(2S)}M_\epsilon^{-1}\right)\Big|_{\epsilon=0}, \\
&= \frac{\partial}{\partial\epsilon}M_\epsilon^{-1}\Big|_{\epsilon=0} Q^{(2S)}M^{-1} + M^{-1}\frac{\partial}{\partial\epsilon}Q_\epsilon^{(2S)}\Big|_{\epsilon=0} M^{-1} \\
&\quad + M^{-1}Q^{(2S)}\frac{\partial}{\partial\epsilon}M_\epsilon^{-1}\Big|_{\epsilon=0}, \qquad \text{(A.21)}
\end{aligned}$$

where $Q^{(2S)}$ denotes the Q matrix of the two-stage estimator, $M = -\int \frac{\partial}{\partial\theta}\Psi_2[\tilde{z}^{(2)}; h\{\tilde{z}^{(1)}; S(F)\}, \theta]dF(\tilde{z})$, ϵ denotes the dependence of the estimating functional on the contaminated distribution. To get rid of possible ambiguity denote M_1 the M-matrix of the first stage.

Let us compute the necessary derivatives separately.

$$\begin{aligned}
\frac{\partial M_\epsilon}{\partial\epsilon}\Big|_{\epsilon=0} &= \int \frac{\partial}{\partial\epsilon}\frac{\partial}{\partial\theta}\Psi_2[\tilde{z}^{(2)}; h\{\tilde{z}^{(1)}; S(F_\epsilon)\}, \theta]dF\Big|_{\epsilon=0} \\
&\quad + \frac{\partial}{\partial\theta}\Psi_2[z^{(2)}; h\{z^{(1)}; S(F)\}, \theta] - M. \qquad \text{(A.22)}
\end{aligned}$$

Denote $\Psi_2 = \Psi_2[z^{(2)}; h\{z^{(1)}; S(F)\}, T(F)] = \begin{pmatrix} \Psi_{21} & \Psi_{22} & \dots & \Psi_{2p} \end{pmatrix}^T$ and $\Psi_{2\epsilon} = \Psi_2[z^{(2)}; h\{z^{(1)}; S(F_\epsilon)\}, T(F_\epsilon)] = \begin{pmatrix} \Psi_{21\epsilon} & \Psi_{22\epsilon} & \dots & \Psi_{2p\epsilon} \end{pmatrix}^T$.

The derivative under the integral of the first term of (A.22) is

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\frac{\partial}{\partial\theta}\Psi_{2\epsilon}\Big|_{\epsilon=0} &= \left.\begin{pmatrix} \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{21\epsilon}}{\partial\theta_1} & \cdots & \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{21\epsilon}}{\partial\theta_p} \\ \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{22\epsilon}}{\partial\theta_1} & \cdots & \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{22\epsilon}}{\partial\theta_p} \\ & \cdots & \\ \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{2p\epsilon}}{\partial\theta_1} & \cdots & \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{2p\epsilon}}{\partial\theta_p} \end{pmatrix}\right|_{\epsilon=0} \\
&= \begin{pmatrix} \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{21}}{\partial\theta_1}\right)^T \frac{\partial h}{\partial s}IF(z;S,F) & \cdots & \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{21}}{\partial\theta_p}\right)\frac{\partial h}{\partial s}IF(z;S,F) \\ \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{22}}{\partial\theta_1}\right)^T \frac{\partial h}{\partial s}IF(z;S,F) & \cdots & \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{22}}{\partial\theta_p}\right)\frac{\partial h}{\partial s}IF(z;S,F) \\ & \cdots & \\ \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{2p}}{\partial\theta_1}\right)^T \frac{\partial h}{\partial s}IF(z;S,F) & \cdots & \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{2p}}{\partial\theta_p}\right)\frac{\partial h}{\partial s}IF(z;S,F) \end{pmatrix} \\
&\quad + \begin{pmatrix} \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{21}}{\partial\theta_1}\right)^T IF(z;T,F) & \cdots & \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{21}}{\partial\theta_p}\right)^T IF(z;T,F) \\ \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{22}}{\partial\theta_1}\right)^T IF(z;T,F) & \cdots & \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{22}}{\partial\theta_p}\right)^T IF(z;T,F) \\ & \cdots & \\ \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{2p}}{\partial\theta_1}\right)^T IF(z;T,F) & \cdots & \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{2p}}{\partial\theta_p}\right)^T IF(z;T,F) \end{pmatrix} \\
&= D^{(2S)}, \qquad \text{(A.23)}
\end{aligned}$$

where $h_\epsilon = h\{z^{(1)}; S(F_\epsilon)\}$, $h = h\{z^{(1)}, S(F)\}$, $\frac{\partial}{\partial h}\frac{\partial\Psi_{2i}}{\partial\theta_j} = \frac{\partial}{\partial\zeta}\frac{\partial}{\partial\theta_j}\Psi_{2i}(z^{(2)};\zeta,\theta_j)$, $\frac{\partial h}{\partial s} = \frac{\partial}{\partial\zeta}h(z^{(1)};\zeta)$. The derivative $\frac{\partial}{\partial\theta}\frac{\partial\Psi_{2i}}{\partial\theta_j} = \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta_j}\Psi_{2i}[z^{(2)}; h\{z^{(1)}; S(F)\}, \theta_j]$ is the derivative with respect to the parameter vector θ.

$$\frac{\partial Q_\epsilon^{(2S)}}{\partial\epsilon}\Big|_{\epsilon=0} = \frac{\partial}{\partial\epsilon}\int \{a_\epsilon(z)+b_\epsilon(z)\}\{a_\epsilon(z)+b_\epsilon(z)\}^T d\{(1-\epsilon)F+\epsilon\Delta_z\}\Big|_{\epsilon=0}. \qquad \text{(A.24)}$$

Splitting the integral, we have

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\int a_\epsilon(z)a_\epsilon(z)^T dF_\epsilon\Big|_{\epsilon=0} &= \int \frac{\partial}{\partial\epsilon}a_\epsilon(z)a_\epsilon(z)^T dF\Big|_{\epsilon=0} \\
&\quad - \int a(z)a(z)^T dF + a(z)a(z)^T, \\
\frac{\partial}{\partial\epsilon}\int a_\epsilon(z)b_\epsilon(z)^T dF_\epsilon\Big|_{\epsilon=0} &= \int \frac{\partial}{\partial\epsilon}a_\epsilon(z)b_\epsilon(z)^T dF\Big|_{\epsilon=0} \\
&\quad - \int a(z)b(z)^T dF + a(z)b(z)^T, \\
\frac{\partial}{\partial\epsilon}\int b_\epsilon(z)a_\epsilon(z)^T dF_\epsilon\Big|_{\epsilon=0} &= \int \frac{\partial}{\partial\epsilon}b_\epsilon(z)a_\epsilon(z)^T dF\Big|_{\epsilon=0} \\
&\quad - \int b(z)a(z)^T dF + b(z)a(z)^T, \\
\frac{\partial}{\partial\epsilon}\int b_\epsilon(z)b_\epsilon(z)^T dF_\epsilon\Big|_{\epsilon=0} &= \int \frac{\partial}{\partial\epsilon}b_\epsilon(z)b_\epsilon(z)^T dF\Big|_{\epsilon=0} \\
&\quad - \int b(z)b(z)^T dF + b(z)b(z)^T.
\end{aligned}$$

Now we need to compute the derivatives of $a_\epsilon(z)$ and $b_\epsilon(z)$:

$$\begin{aligned}
A = \frac{\partial}{\partial\epsilon}a_\epsilon(z)\Big|_{\epsilon=0} =& \frac{\partial}{\partial h}\Psi_2\{z^{(2)}; h, T(F)\}\frac{\partial}{\partial s}h(z^{(1)}; s)IF(z; S, F) \\
&+ \frac{\partial}{\partial\theta}\Psi_2[z^{(2)}; h\{z^{(1)}; S(F)\}, \theta]\cdot IF(z; T, F), \quad \text{(A.25)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}b_\epsilon(z)\Big|_{\epsilon=0} =& \frac{\partial}{\partial\epsilon}\int \frac{\partial}{\partial h}\Psi_{2\epsilon}\frac{\partial}{\partial s}h_\epsilon dF M_{1\epsilon}^{-1}\Psi_{1\epsilon}\Big|_{\epsilon=0} \\
&- \int \frac{\partial}{\partial h}\Psi_2\frac{\partial}{\partial s}h dF\cdot IF(z; S, F) + \frac{\partial}{\partial h}\Psi_2\frac{\partial}{\partial s}h\cdot IF(z; S, F)
\end{aligned} \quad \text{(A.26)}$$

The first term of (A.26) is

$$\begin{aligned}\frac{\partial}{\partial\epsilon}\int\frac{\partial\Psi_{2\epsilon}}{\partial h}\frac{\partial h_\epsilon}{\partial s}dFM_{1\epsilon}^{-1}\Psi_{1\epsilon}\bigg|_{\epsilon=0} &= \int\left(\frac{\partial}{\partial\epsilon}\frac{\partial}{\partial h}\Psi_{2\epsilon}\right)\bigg|_{\epsilon=0}\frac{\partial}{\partial s}hdFM_1^{-1}\Psi_1 \\ &+ \int\frac{\partial}{\partial h}\Psi_2\left(\frac{\partial}{\partial\epsilon}\frac{\partial}{\partial s}h_\epsilon\right)\bigg|_{\epsilon=0}dFM_1^{-1}\Psi_1 \\ &+ \int\frac{\partial}{\partial h}\Psi_2\frac{\partial}{\partial s}hdF\left(\frac{\partial}{\partial\epsilon}M_{1\epsilon}^{-1}\right)\bigg|_{\epsilon=0}\Psi_1 \\ &+ \int\frac{\partial}{\partial h}\Psi_2\frac{\partial}{\partial s}hdFM_1^{-1}\left(\frac{\partial}{\partial\epsilon}\Psi_{1\epsilon}\right)\bigg|_{\epsilon=0}\end{aligned} \tag{A.27}$$

We know that $M_1^{-1}\Psi_1 = IF(z;S,F)$, $\frac{\partial}{\partial\epsilon}M_{1\epsilon}^{-1}\Big|_{\epsilon=0}$ is the derivative of M^{-1} matrix of one-stage M-estimator (see first line of formula A.17), and $\frac{\partial}{\partial\epsilon}\Psi_{1\epsilon}\Big|_{\epsilon=0} = \frac{\partial}{\partial\theta}\Psi_1 IF(z;S,F)$. Hence,

$$\begin{aligned}\frac{\partial}{\partial\epsilon}\int\frac{\partial\Psi_{2\epsilon}}{\partial h}\frac{\partial h_\epsilon}{\partial s}dFM_{1\epsilon}^{-1}\Psi_{1\epsilon}\bigg|_{\epsilon=0} &= \int\left(\frac{\partial}{\partial\epsilon}\frac{\partial}{\partial h}\Psi_{2\epsilon}\right)\bigg|_{\epsilon=0}\frac{\partial}{\partial s}hdFIF(z;S,F) \\ &+ \int\frac{\partial}{\partial h}\Psi_2\left(\frac{\partial}{\partial\epsilon}\frac{\partial}{\partial s}h_\epsilon\right)\bigg|_{\epsilon=0}dFIF(z;S,F) \\ &- \int\frac{\partial}{\partial h}\Psi_2\frac{\partial}{\partial s}hdFM_1^{-1}\left(\int D^{(1)}dF\right)M_1^{-1}\Psi_1 \\ &- \int\frac{\partial}{\partial h}\Psi_2\frac{\partial}{\partial s}hdFM_1^{-1}\left(\frac{\partial}{\partial\theta}\Psi_1 - M_1\right)M_1^{-1}\Psi_1 \\ &+ \int\frac{\partial}{\partial h}\Psi_2\frac{\partial}{\partial s}hdFM_1^{-1}\frac{\partial}{\partial\theta}\Psi_1 IF(z;S,F),\end{aligned} \tag{A.28}$$

where $D^{(1)}$ denotes the D matrix of the first stage, i.e. $\frac{\partial}{\partial\epsilon}\frac{\partial}{\partial\theta}\Psi_{2\epsilon}\Big|_{\epsilon=0}$, which is defined in formula (A.18).

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\frac{\partial}{\partial h}\Psi_{2\epsilon}\Big|_{\epsilon=0} &= \begin{pmatrix} \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{21\epsilon}}{\partial h_1} & \cdots & \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{21\epsilon}}{\partial h_k} \\ \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{22\epsilon}}{\partial h_1} & \cdots & \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{22\epsilon}}{\partial h_k} \\ & \cdots & \\ \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{2p\epsilon}}{\partial h_1} & \cdots & \frac{\partial}{\partial\epsilon}\frac{\partial\Psi_{2p\epsilon}}{\partial h_k} \end{pmatrix}\Bigg|_{\epsilon=0} \\
&= \begin{pmatrix} \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{21}}{\partial h_1}\right)^T\frac{\partial}{\partial s}hIF(S) & \cdots & \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{21}}{\partial h_k}\right)^T\frac{\partial}{\partial s}hIF(S) \\ \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{22}}{\partial h_1}\right)^T\frac{\partial}{\partial s}hIF(S) & \cdots & \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{22}}{\partial h_k}\right)^T\frac{\partial}{\partial s}hIF(S) \\ & \cdots & \\ \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{2p}}{\partial h_1}\right)^T\frac{\partial}{\partial s}hIF(S) & \cdots & \left(\frac{\partial}{\partial h}\frac{\partial\Psi_{2p}}{\partial h_k}\right)^T\frac{\partial}{\partial s}hIF(S) \end{pmatrix} \\
&\quad + \begin{pmatrix} \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{21}}{\partial h_1}\right)^T IF(T) & \cdots & \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{21}}{\partial h_k}\right)^T IF(T) \\ \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{22}}{\partial h_1}\right)^T IF(T) & \cdots & \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{22}}{\partial h_k}\right)^T IF(T) \\ & \cdots & \\ \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{2p}}{\partial h_1}\right)^T IF(T) & \cdots & \left(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{2p}}{\partial h_k}\right)^T IF(T) \end{pmatrix} = R_1,
\end{aligned} \tag{A.29}$$

which is $p \times q$ dimensional.

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\frac{\partial}{\partial s}h_\epsilon\Big|_{\epsilon=0} &= \begin{pmatrix} \frac{\partial}{\partial\epsilon}\frac{\partial h_{1\epsilon}}{\partial s_1} & \cdots & \frac{\partial}{\partial\epsilon}\frac{\partial h_{1\epsilon}}{\partial s_q} \\ \frac{\partial}{\partial\epsilon}\frac{\partial h_{2\epsilon}}{\partial s_1} & \cdots & \frac{\partial}{\partial\epsilon}\frac{\partial h_{2\epsilon}}{\partial s_q} \\ & \cdots & \\ \frac{\partial}{\partial\epsilon}\frac{\partial h_{k\epsilon}}{\partial s_1} & \cdots & \frac{\partial}{\partial\epsilon}\frac{\partial h_{k\epsilon}}{\partial s_q} \end{pmatrix}\Bigg|_{\epsilon=0} \\
&= \begin{pmatrix} \left(\frac{\partial}{\partial\theta_2}\frac{\partial h_1}{\partial s_1}\right)^T IF(S) & \cdots & \left(\frac{\partial}{\partial\theta_2}\frac{\partial h_1}{\partial s_q}\right)^T IF(S) \\ \left(\frac{\partial}{\partial\theta_2}\frac{\partial h_2}{\partial s_1}\right)^T IF(S) & \cdots & \left(\frac{\partial}{\partial\theta_2}\frac{\partial h_2}{\partial s_q}\right)^T IF(S) \\ & \cdots & \\ \left(\frac{\partial}{\partial\theta_2}\frac{\partial h_k}{\partial s_1}\right)^T IF(S) & \cdots & \left(\frac{\partial}{\partial\theta_2}\frac{\partial h_k}{\partial s_q}\right)^T IF(S) \end{pmatrix} = R_2,
\end{aligned} \tag{A.30}$$

with the dimension of the matrix equal to $k \times q$.

Using R_1 and R_2 we obtain the matrix B

$$\begin{aligned}
B = \frac{\partial}{\partial \epsilon} b_\epsilon(z)\Big|_{\epsilon=0} &= \int R_1 \frac{\partial}{\partial s} h dF IF(z;S,F) + \int \frac{\partial}{\partial h}\Psi_2 R_2 dF IF(z;S,F) \\
&\quad - \int \frac{\partial}{\partial h}\Psi_2 \frac{\partial}{\partial s} h dF M_1^{-1} \left(\int D^{(1)} dF - M_1 + \frac{\partial}{\partial \theta}\Psi_1 \right) IF(z;S,F) \\
&\quad + \int \frac{\partial}{\partial h}\Psi_2 \frac{\partial}{\partial s} h dF M_1^{-1} \frac{\partial}{\partial \theta}\Psi_1 IF(z;S,F) \\
&\quad - \int \frac{\partial}{\partial h}\Psi_2 \frac{\partial}{\partial s} h dF \cdot IF(z;S,F) + \frac{\partial}{\partial h}\Psi_2 \frac{\partial}{\partial s} h \cdot IF(z;S,F) \\
&= \int R_1 \frac{\partial}{\partial s} h dF IF(z;S,F) + \int \frac{\partial}{\partial h}\Psi_2 R_2 dF IF(z;S,F) \\
&\quad - \int \frac{\partial}{\partial h}\Psi_2 \frac{\partial}{\partial s} h dF M_1^{-1} \left(\int D^{(1)} dF + \frac{\partial}{\partial \theta}\Psi_1 \right) IF(z;S,F) \\
&\quad + \int \frac{\partial}{\partial h}\Psi_2 \frac{\partial}{\partial s} h dF M_1^{-1} \frac{\partial}{\partial \theta}\Psi_1 IF(z;S,F) \\
&\quad + \frac{\partial}{\partial h}\Psi_2 \frac{\partial}{\partial s} h \cdot IF(z;S,F). \qquad \text{(A.31)}
\end{aligned}$$

Combining all the terms we obtain

$$\begin{aligned}
CVF(z;S,T,F) = &- M^{-1} \left(\int D^{(2S)} dF - M \right) M^{-1} Q^{(2S)} M^{-1} \\
&- M^{-1} \left(\frac{\partial}{\partial \theta}\Psi_2[z^{(2)}; h\{z^{(1)}; S(F)\}, \theta] \right) M^{-1} Q^{(2S)} M^{-1} \\
&+ M^{-1} \int \left\{ Aa(z)^T + a(z)A^T + Ab(z)^T + a(z)B^T \right\} dF M^{-1} \\
&+ M^{-1} \int \left\{ Ba(z)^T + b(z)A^T + Bb(z)^T + b(z)B^T \right\} dF M^{-1} \\
&- M^{-1} Q^{(2S)} M^{-1} \\
&+ M^{-1} \left\{ a(z) + b(z) \right\} \left\{ a(z) + b(z) \right\}^T M^{-1} \\
&- M^{-1} Q^{(2S)} M^{-1} \left(\int D^{(2S)} dF - M \right) M^{-1} \\
&- M^{-1} Q^{(2S)} M^{-1} \left(\frac{\partial}{\partial \theta}\Psi_2[z^{(2)}; h\{z^{(1)}; S(F)\}, \theta] \right) M^{-1}.
\end{aligned}$$

Recall that $V = M^{-1} Q^{(2S)} M^{-1}$ we can factorize V and simplify the for-

mula

$$\begin{aligned} CVF(z;S,T,F) =& V - M^{-1}\left(\int D^{(2S)}dF + \frac{\partial}{\partial\theta}\Psi_2[z^{(2)};h\{z^{(1)};S(F)\},\theta]\right)V \\ &+ M^{-1}\int\left\{Aa(z)^T + Ba(z)^T + Ab(z)^T + Bb(z)^T\right\}dFM^{-1} \\ &+ M^{-1}\int\left\{a(z)A^T + b(z)A^T + a(z)B^T + b(z)B^T\right\}dFM^{-1} \\ &+ M^{-1}\left\{a(z)+b(z)\right\}\left\{a(z)+b(z)\right\}^T M^{-1} \\ &- V\left(\int D^{(2S)}dF + \frac{\partial}{\partial\theta}\Psi_2[z^{(2)};h\{z^{(1)};S(F)\},\theta]\right)M^{-1}. \end{aligned} \tag{A.32}$$

Note that the expression of the CVF in (A.32) has the same structure as the CVF in (A.20). The first line in (A.20) corresponds to the first and the last lines of (A.32). In the last line of (A.20) the integrals of R and R^T have the analogs in the lines 2 and 3 in (A.32). The main source of unboundedness of the one-stage CVF, i.e. $\Psi\Psi^T$ in (A.20), corresponds to the line 4 in (A.32). Also, the CVF of the 2-stage estimator linearly depends on the CVF of the 1-stage estimator through $Bb(z)^T$, $b(z)B^T$ and $b(z)b(z)^T$ terms.

A.5 CVF of the Heckman's two-stage estimator

Proof of Proposition 2. Using the result in Section A.4 we can proceed with the computation of the CVF for the Heckman's estimator. Recall that the function $h\{z_1;S(F)\}$ is a scalar function $\lambda(x_1^T\beta_1)$ and its derivative with respect to β_1 is given by (A.7). The IF's of $T(F)$ and $S(F)$ are given by (4.4) and (4.5), respectively. The $M(\Psi_1)$ and $M(\Psi_2)$ matrices are given by (A.11) and (A.9), respectively and the scores $\Psi_1\{z;S(F)\}$ and $\Psi_2\{z;\lambda,T(F)\}$ are given by (A.12) and (A.8). The derivative of the score function with respect to λ is given in the formula (A.10).

The second term of the matrix D is equal to zero, hence the D matrix for Heckman's estimator takes the form

$$D_H = \begin{pmatrix} 0 & x_2 \\ x_2^T & 2\lambda \end{pmatrix}\lambda' y_1 IF(z;S,F),$$

where 0 is a $(p_2 - 1) \times (p_2 - 1)$ matrix of zeroes.

Denote the analog of the A matrix for the Heckman's estimator as A_H. we obtain

$$A_H = \begin{pmatrix} -x_2\beta_\lambda \\ y_2 - x_2^T\beta_2 - 2\lambda\beta_\lambda \end{pmatrix} y_1\lambda' IF(z;S,F) + \begin{pmatrix} x_2x_2^T & x_2\lambda \\ x_2^T\lambda & \lambda^2 \end{pmatrix} y_1 IF(z;T,F).$$

Next, we need to get the expression of B_H. Define $R_H^{(1)}$ and $R_H^{(2)}$ corresponding to the $R^{(1)}$ and $R^{(2)}$ matrices :

$$R_H^{(1)} = \begin{pmatrix} 0 \\ -2\beta_\lambda \end{pmatrix} y_1\lambda' IF(z;S,F) + \begin{pmatrix} 0 & -x_2 \\ -x_2 & -2\lambda \end{pmatrix} y_1 IF(z;T,F),$$

where 0 in the first term is a $(p_2 - 1)$ vector, and in the second term 0 is a $(p_2 - 1) \times (p_2 - 1)$ matrix. Let us compute $R_H^{(2)}$:

$$\begin{aligned} R_H^{(2)} =& \frac{\partial}{\partial\epsilon}\frac{\partial}{\partial s}\lambda_\epsilon\Big|_{\epsilon=0} = \frac{\partial}{\partial\epsilon}\lambda_\epsilon'\Big|_{\epsilon=0} \\ =& \left[\frac{\left\{\Phi(x_1^T\beta_1)(x_1^T\beta_1)^2 - \phi(x_1^T\beta_1)x_1^T\beta_1 - \Phi(x_1^T\beta_1) + 2\phi(x_1^T\beta_1)\right\}\phi(x_1^T\beta_1)}{\Phi(x_1^T\beta_1)^2}\right. \\ & \left. - \frac{2\phi(x_1^T\beta_1)^2\left\{-\Phi(x_1^T\beta_1)x_1^T\beta_1 - \phi(x_1^T\beta_1)\right\}}{\Phi(x_1^T\beta_1)^3}\right]\left\{x_1x_1^T IF(z;S,F)\right\}^T. \end{aligned}$$

The expression above is a p_1 vector. The $D^{(1)}$ matrix for the probit estimator is:

$$\begin{aligned} D_{probit} =& \frac{\partial}{\partial\epsilon}\frac{\partial}{\partial\theta}\Psi_1(z;\theta)\Big|_{\epsilon=0} \\ =& \left[\frac{-\phi(x_1^T\beta_1)^2x_1^T\beta_1}{\Phi(x_1^T\beta_1)\{1-\Phi(x_1^T\beta_1)\}}x_1^T IF(z;S,F)\right]x_1x_1^T \\ & - \left[\frac{\phi(x_1^T\beta_1)^3\{1-2\Phi(x_1^T\beta_1)\}}{\Phi(x_1^T\beta_1)^2\{1-\Phi(x_1^T\beta_1)\}^2}x_1^T IF(z;S,F)\right]x_1x_1^T, \end{aligned}$$

where we first take the derivative with respect to θ, and when we take the derivative with respect to ϵ, we evaluate the second derivative with respect to θ at $\theta = S(F)$. We defined all the ingredients of the B_H matrix, and we can use them in (A.31).

In order to obtain the expression of the CVF we need to insert the expressions obtained above into (A.32). By noting that the integrals of $B_Ha(z)^T$ and $a(z)B_H^T$ are equal to zero, we finally obtain (4.9). □

A.6 Assumptions and proof of Proposition 4

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

(i) $z_1, \ldots, z_N$ 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 $z \in \mathcal{Z}$ $|\Psi^R(z;\theta)\Psi^R(z;\theta)^T| \leq \xi_1$ 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. □

Appendix B

Complementary Materials

B.1 Selection bias under contaminated normal distribution

Consider a bivariate random vector with contaminated normal distribution:

$$\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \sim (1-\epsilon)N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix};\ \ \Sigma_1 \right\} + \epsilon N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix};\ \ \Sigma_2 \right\}, \quad \text{(B.1)}$$

where

$$\Sigma_1 = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} \nu_1^2 & \tau\nu_1\nu_2 \\ \tau\nu_1\nu_2 & \nu_2^2 \end{pmatrix}.$$

We need to derive the conditional expectation $E(e_1|e_2 > -d)$, where d corresponds to the linear predictor $x_{1i}^T\beta_1$ in the case of sample selection models. From (B.1) we can derive the following probabilities:

$$P(e_2 > -d) = (1-\epsilon)\Phi(d) + \epsilon\Phi\left(\frac{d}{\nu_2}\right),$$

$$P(e_2 > -d|e_1) = \frac{(1-\epsilon)\Phi_2(e_1, d, \sigma_1, 1, \rho) + \epsilon\Phi_2(e_1, d, \nu_1, \nu_2, \tau)}{(1-\epsilon)\Phi\left(\frac{e_1}{\sigma_1}\right) + \epsilon\Phi\left(\frac{e_1}{\nu_1}\right)},$$

where $\Phi_2()$ denotes the bivariate normal cdf. The necessary probability $P(e_1|e_2 > -d)$ is equal to

$$P(e_1|e_2 > -d) = \frac{P(e_1 < d_1)P(e_2 > -d|e_1)}{P(e_2 > -d)}.$$

Taking into account that $P(e_2 > -d|e_1)$ has $P(e_1 < d_1)$ in denominator, we obtain

$$P(e_1 < c|e_2 > -d) = \frac{(1-\epsilon)\Phi_2(c,d,\sigma_1,1,\rho) + \epsilon\Phi_2(c,d,\nu_1,\nu_2,\tau)}{(1-\epsilon)\Phi(d) + \epsilon\Phi\left(\frac{d}{\nu_2}\right)}. \quad \text{(B.2)}$$

Using the result in Kotz et al. (2000) (page 255), that the $\frac{\partial\Phi_2(e_1,e_2,\sigma_1,\sigma_2,\rho)}{\partial e_1} = \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{e_1^2}{2\sigma_1^2}\right)\Phi\left(\frac{e_2}{\sigma_2\sqrt{1-\rho^2}} - \frac{\rho e_1}{\sigma_1\sqrt{1-\rho^2}}\right)$, we obtain

$$\begin{aligned}
f(e_1|e_2 > -d) &= \frac{\frac{(1-\epsilon)}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{e_1^2}{2\sigma_1^2}\right)\Phi\left(\frac{-\rho e_1}{\sigma_1\sqrt{1-\rho^2}} + \frac{d}{\sqrt{1-\rho^2}}\right)}{(1-\epsilon)\Phi(d)+\epsilon\Phi\left(\frac{d}{\nu_2}\right)} \\
&\quad + \frac{\frac{\epsilon}{\nu_1\sqrt{2\pi}}\exp\left(-\frac{e_1^2}{2\nu_1^2}\right)\Phi\left(\frac{-\tau e_1}{\nu_1\sqrt{1-\tau^2}} + \frac{d}{\nu_2\sqrt{1-\tau^2}}\right)}{(1-\epsilon)\Phi(d)+\epsilon\Phi\left(\frac{d}{\nu_2}\right)} \\
&= \frac{(1-\epsilon)\phi\left(e_1,0,\sigma_1^2\right)\Phi\left(\frac{-\rho e_1}{\sigma_1\sqrt{1-\rho^2}} + \frac{d}{\sqrt{1-\rho^2}}\right)}{(1-\epsilon)\Phi(d)+\epsilon\Phi\left(\frac{d}{\nu_2}\right)} \\
&\quad + \frac{\epsilon\phi\left(e_1,0,\nu_1^2\right)\Phi\left(\frac{-\tau e_1}{\nu_1\sqrt{1-\tau^2}} + \frac{d}{\nu_2\sqrt{1-\tau^2}}\right)}{(1-\epsilon)\Phi(d)+\epsilon\Phi\left(\frac{d}{\nu_2}\right)} \\
&= \frac{(1-\epsilon)\phi\left(e_1,0,\sigma_1^2\right)\Phi\left(\frac{\frac{-\rho e_1}{\sigma_1}+d}{\sqrt{1-\rho^2}}\right) + \epsilon\phi\left(e_1,0,\nu_1^2\right)\Phi\left(\frac{\frac{-\tau e_1}{\nu_1}+\frac{d}{\nu_2}}{\sqrt{1-\tau^2}}\right)}{(1-\epsilon)\Phi(d)+\epsilon\Phi\left(\frac{d}{\nu_2}\right)}.
\end{aligned}$$

The densities can be represented in the form of the densities of extended skew-normal type (Arellano-Valle and Genton 2010). The expectations are known, therefore we obtain

$$\begin{aligned}
E(e_1|e_2 > -d) &= \left\{(1-\epsilon)\Phi(d) + \epsilon\Phi\left(\frac{d}{\nu_2}\right)\right\}^{-1} \\
&\quad \times \left\{(1-\epsilon)\Phi(d)\rho\sigma_1\frac{\phi(d)}{\Phi(d)} + \epsilon\Phi\left(\frac{d}{\nu_2}\right)\tau\nu_1\frac{\phi\left(\frac{d}{\nu_2}\right)}{\Phi\left(\frac{d}{\nu_2}\right)}\right\},
\end{aligned}$$

and finally

$$\begin{aligned}
E(e_1|e_2 > -d) &= \left\{(1-\epsilon)\Phi(d) + \epsilon\Phi\left(\frac{d}{\nu_2}\right)\right\}^{-1} \\
&\quad \times \left\{(1-\epsilon)\rho\sigma_1\phi(d) + \epsilon\tau\nu_1\phi\left(\frac{d}{\nu_2}\right)\right\}.
\end{aligned} \quad \text{(B.3)}$$

The correspondence between the analytical expression of the selection bias in (B.3) and Monte Carlo simulation is presented in Figure B.1. Different values of the parameters are considered, and the graphs match in all panels.

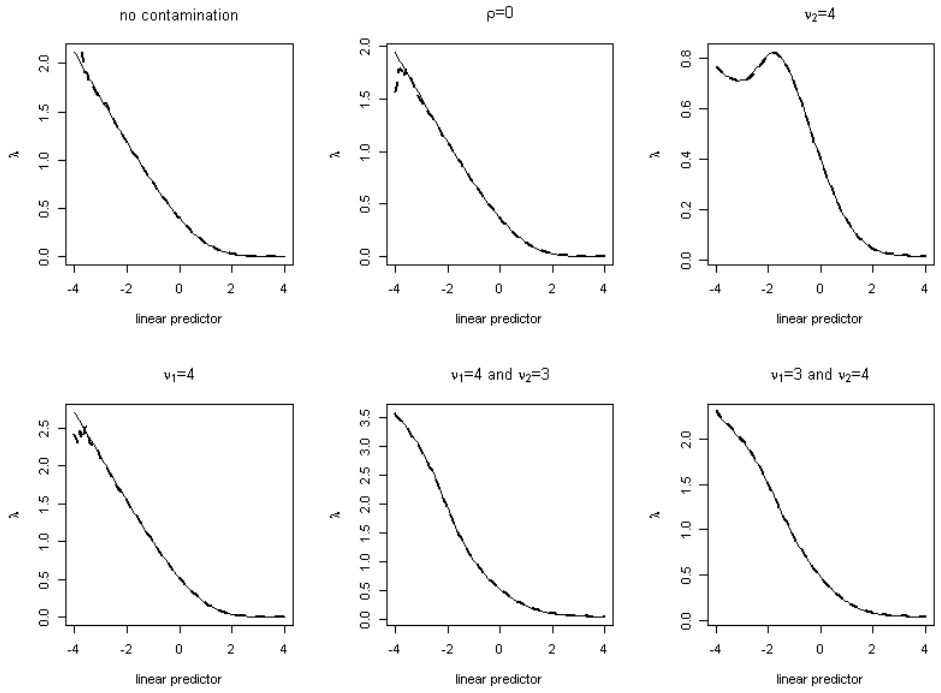

Figure B.1: *Comparison of analytically computed selection bias and Monte Carlo simulation.*

B.2 IF of 2SLS, general case

Assume that we have the following regression model:

$$y = x_1^T \beta_1 + x_2^T \beta_2 + e,$$

where $\text{cov}(x_1, e) = 0$ and $\text{cov}(x_2, e) \neq 0$, i.e. x_1 is exogenous and x_2 is endogenous. Denote $x_1 = \begin{pmatrix} x_{11} \\ \dots \\ x_{1p_1} \end{pmatrix}$ and $x_2 = \begin{pmatrix} x_{21} \\ \dots \\ x_{2p_2} \end{pmatrix}$. So the total

number of explanatory variables is $p_1 + p_2 = p$. Suppose that for each x_{2i} we have a vector of instruments w_i such that $cor(w_{ij}, e) = 0$ for $\forall j$. Each w_i is a $q_i \times 1$ dimensional vector and the total number of instruments is equal to $\sum_i^{p_2} q_i = q$.

The most popular solution of this problem is 2SLS. The idea is to regress the endogenous variables x_{2i} on corresponding set of instruments w_i, then estimate $\hat{x}_{2i}$ and on the second stage regress y on x_1 and $\hat{x}_2$. It means that on the first stage we have the following system of equations:

$$\begin{cases} x_{21} &= w_1^T \alpha_1 + e_{21}, \\ \dots & \\ x_{2p_2} &= w_{p_2}^T \alpha_{p_2} + e_{2p_2}, \end{cases}$$

where e_{2i} are the error terms with zero expectation.

In each datum $z = (y, x_1, x_2, w)$ we have the following matrix of instruments:

$$w = \begin{pmatrix} w_{11} & \dots & w_{1q_1} & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & w_{21} & \dots & w_{2q_1} & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & w_{p_2 1} & \dots & w_{p_2 q_{p_2}} \end{pmatrix}.$$

In the first stage we need to estimate the q dimensional vector of parameters:

$$\alpha = \begin{pmatrix} \alpha_{11} \\ \dots \\ \alpha_{1q_1} \\ \alpha_{21} \\ \dots \\ \alpha_{2q_2} \\ \dots\dots \\ \alpha_{p_2 1} \\ \dots \\ \alpha_{p_2 q_{p_2}} \end{pmatrix}.$$

The MLE of α is:

$$\alpha = \left(\int w^T w dF \right)^{-1} \int w^T x_2 dF,$$

where F is the distribution of the datum z. Next, we estimate $\hat{x}_2 = w\hat{\alpha}$ and regress $y = x_1^T \beta_1 + \hat{x}_2 \beta_2$.

Let us compute the IF of the second stage estimator. For this we can use the general formula (A.6) and need to specify its components. The IF of the first stage is proportional to the score function of the estimator.

$$\Psi_1\{(x_2, w); S(F)\} = w^T(x_2 - w\alpha).$$

The score function of the second stage is:

$$\begin{aligned}
&\Psi_1[(x_1, x_2, y); h\{S(F)\}, T(F)] \\
&= \left\{ y - x_1^T\beta_1 - \begin{pmatrix} w_1^T\alpha_1 & \dots & w_{p_2}^T\alpha_{p_2} \end{pmatrix} \begin{pmatrix} \beta_{21} \\ \vdots \\ \beta_{2p_2} \end{pmatrix} \right\} \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p_1} \\ w_1^T\alpha_1 \\ \vdots \\ w_{p_2}^T\alpha_{p_2} \end{pmatrix}.
\end{aligned}$$

The derivative of the score function with respect to $h\{S(F)\} = w\alpha$ is a $p \times p_2$ dimensional matrix. The derivative of $h(\cdot)$ with respect to the estimating functional is $\frac{\partial h\{S(F)\}}{\partial S(F)} = w$, which is $p_2 \times q$ dimensional matrix.

Combining all the terms we obtain the expression of IF, which is unbounded in all components of z.

B.3 Asymptotic variance of Heckman's two-stage estimator

In order to derive the expression of the asymptotic variance for the Heckman's two-stage estimator in (4.7) we can use the general expression of the variance for two-stage M-estimator in (3.6). Let us compute the four terms separately.

$$\begin{aligned}
\int a(z)a(z)^T dF &= \int \Psi_2[(x_2, y_2); \lambda\{(x_1, y_1); S(F)\}, T(F)] \\
&\quad \times \Psi_2[(x_2, y_2); \lambda\{(x_1, y_1); S(F)\}, T(F)]^T dF.
\end{aligned}$$

Using the Ψ_2 functions from (A.8) we obtain

$$\begin{aligned}
\int a(z)a(z)^T dF &= \int (y_2 - x_2^T\beta_2 - \lambda\beta_\lambda) \begin{pmatrix} x_2 \\ \lambda \end{pmatrix} y_1 \\
&\quad \times \left\{ (y_2 - x_2^T\beta_2 - \lambda\beta_\lambda) \begin{pmatrix} x_2 \\ \lambda \end{pmatrix} y_1 \right\}^T dF.
\end{aligned}$$

Denote $x = \begin{pmatrix} x_2 \\ \lambda \end{pmatrix}$ and $\beta = \begin{pmatrix} \beta_1 \\ \beta_\lambda \end{pmatrix}$. For the empirical distribution F_N we obtain

$$\int a(z)a(z)^T dF_N = \frac{1}{n}\sum_i^n (y_2 - x^T\beta)xx^T(y_2 - x^T\beta),$$

and using the matrix notation

$$\int a(z)a(z)^T dF_N = X^T Var(Error)X.$$

Taking into account that $Var(Error)$ is heteroscedastic, i.e. $E(v_{2i}^2|x_{2i}, \lambda_i, e_{1i} \geq -x_{1i}\beta_1) = \sigma_2^2\left\{1 + \frac{\beta_\lambda^2}{\sigma_2^2}(x_{1i}\beta_2\lambda_i - \lambda_i^2)\right\}$, see Greene (2008), we obtain:

$$\int a(z)a(z)^T dF_N = \sigma_2^2\left\{X^T\left(I - \frac{\beta_\lambda^2}{\sigma_2^2}\Delta\right)X\right\},$$

where Δ is the diagonal matrix with $\frac{\partial\lambda(x_{1i}\beta_1)}{\partial(x_{1i}\beta_1)}$ on the diagonal. Consider $\int a(z)b(z)^T dF$. This term is the expectation of the multiplication of the score functions of two stages. It means that there is a multiplication of error terms of two equations, and according to the construction of the estimator, these error terms are independent. Hence,

$$\int a(z)b(z)^T dF = 0.$$

By analogy:

$$\int b(z)a(z)^T dF = 0.$$

The last term is

$$\int b(z)b(z)^T dF = \int\left[\int \frac{\partial}{\partial\theta}\Psi_2\{(x_2, y_2); \theta, T(F)\}\frac{\partial}{\partial\eta}\lambda(\eta)dF \cdot IF(z; S, F_t)\right.$$
$$\left.\times\, IF(z; S, F_t)^T\left(\int \frac{\partial}{\partial\theta}\Psi_2\{(x_2, y_2); \theta, T(F)\}\frac{\partial}{\partial\eta}\lambda(\eta)dF\right)^T\right]dF,$$

where the inner integrals are the constants, taking the integral of squared $IF(z, S, F_t)$ we obtain

$$\begin{aligned}\int b(z)b(z)^T dF &= \int \frac{\partial}{\partial\theta}\Psi_2((x_2, y_2); \theta, T(F))\frac{\partial}{\partial\eta}\lambda(\eta)dF Var(S) \\ &\quad\times\left(\int \frac{\partial}{\partial\theta}\Psi_2((x_2, y_2); \theta, T(F))dF\right)^T. \qquad \text{(B.4)}\end{aligned}$$

The sample version of (B.4) is given by

$$\int b(z)b(z)^T dF_N = \beta_\lambda^2 X^T \Delta X_1 \cdot Var(S) \cdot X_1^T \Delta X,$$

β_λ is a scalar, X is $n \times (p+1)$ matrix, Δ is $n \times n$ matrix, X_1 is $n \times q$ matrix, $Var(S)$ is $q \times q$ matrix, hence the final product is $(p+1) \times (p+1)$ matrix. Finally, the asymptotic variance matrix is

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

Appendix C

Additional Monte Carlo Simulations

C.1 Sample selection model

We use the data generating process described in Section 4.3. The level of censoring varies from 25% to 75%. We consider four types of contamination:

> Case A: Contamination of x_1 when the corresponding $y_1 = 1$. The degenerate distribution putting mass 1 at the point $(-6, 1, 1, 1)$. Study of the effect of the leverage outliers when they are transmitted to the main equation.
>
> Case B: Contamination of x_1 when the corresponding $y_1 = 0$. The degenerate distribution putting mass 1 at the point $(6, 0, 1, 1)$. Study of the effect of the leverage outliers when they are not transmitted to the main equation.
>
> Case C: Contamination of x_2. The degenerate distribution putting mass 1 at the point $(1, 1, 6, 1)$. Study of the effect of the leverage outliers in the equation of interest.
>
> Case D: Contamination of y_2. The degenerate distribution putting mass 1 at the point $(1, 1, 1, 6)$. Study of the effect of the outliers in the variable of interest.

We study six estimators with a two-stage structure. They differ by the introduction of robustness into different stages of estimation. In Figures C.4-C.6 they are denoted as follows

(a): Classical Heckman's estimator.

(b): Classical probit MLE and robust linear regression in the second stage.

(c): Robust probit MLE and OLS.

(d): Robust inverse Mills ratio from Section 4.3 and OLS.

(e): Robust two-stage from Section 4.2.

(f): Robust two-stage with robust inverse Mills ratio.

In Figure C.1, Figure C.4, and Figure C.7 we present the results of estimation of the data without contamination. All the estimators perform well, no considerable bias is encountered.

Under contamination the classical estimator breaks down for each type of contamination. The estimator based on classical probit MLE and robust second stage performs well only when the contamination emerges in the second estimation stage, if it appears in the first stage, it breaks down. Combination of robust probit and simple OLS is robust only if the contamination is in the first stage and is not transmitted into the second stage, i.e. the corresponding $y_1 = 0$. Robust probit with robust IMR performs well when the contamination is in the first stage. When there is a contamination in the second stage the robust estimator is needed anyhow. The most robust estimator is the estimator (f). It has the best performance from the robustness point of view, but the loss of efficiency at the model is the largest.

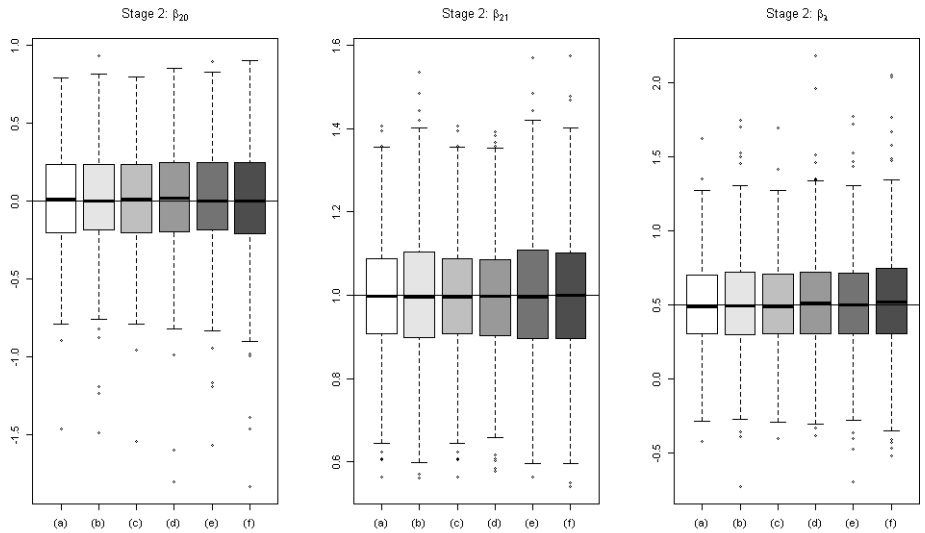

Figure C.1: *Parameter estimation of selection model without contamination. (a) denotes the classical Heckman's estimator. (b) denotes the Classical probit MLE and robust linear regression in the second stage. (c) denotes the robust probit MLE and OLS. (d) denotes the robust inverse Mills ratio from Section 4.3 and OLS. (e) denotes the robust two-stage from Section 4.2. (f) denotes the robust two-stage with robust inverse Mills ratio. The solid line marks the true values of the parameters. The level of censoring is 25%.*

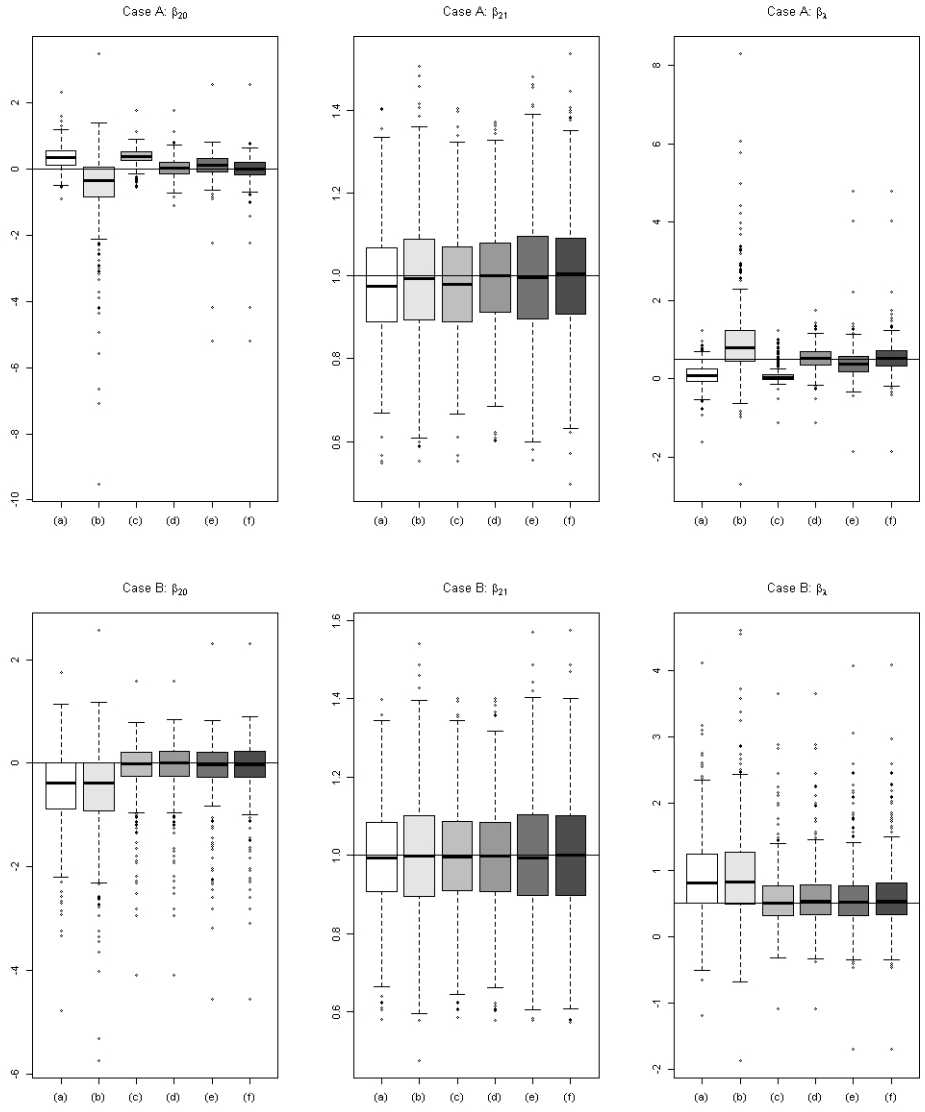

Figure C.2: *Parameter estimation of selection model with contamination. The top panel and the bottom panel correspond to the contamination of Case A and Case B respectively. (a) denotes the classical Heckman's estimator. (b) denotes the probit MLE and robust linear regression in the second stage. (c) denotes the robust probit and OLS. (d) denotes the robust inverse Mills ratio from Section 4.3 and OLS. (e) denotes the robust two-stage from Section 4.2. (f) denotes the robust two-stage with robust inverse Mills ratio. The solid line marks the true values of the parameters. The level of censoring is 25%.*

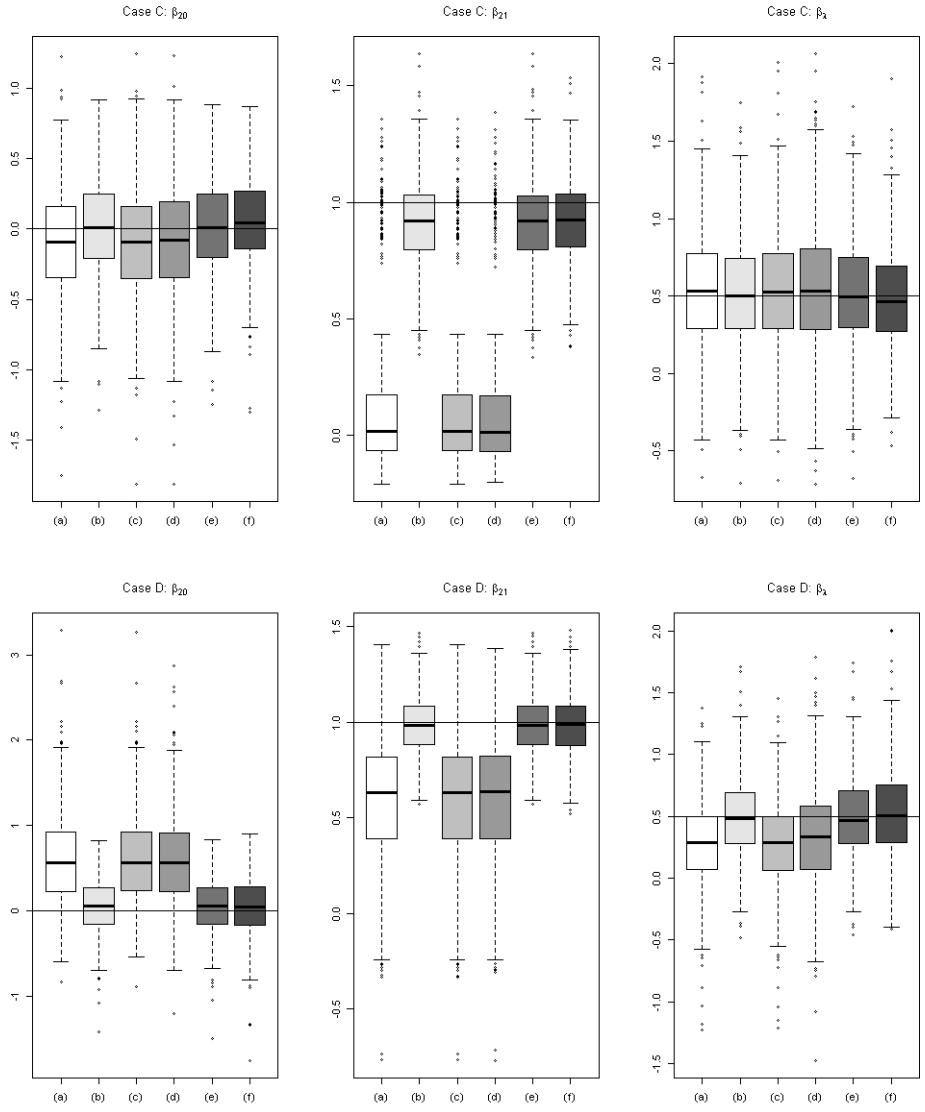

Figure C.3: *Parameter estimation of selection model with contamination. The top panel and the bottom panel correspond to the contamination of Case C and Case D respectively. (a) denotes the classical Heckman's estimator. (b) denotes the probit MLE and robust linear regression in the second stage. (c) denotes the robust probit and OLS. (d) denotes the robust inverse Mills ratio from Section 4.3 and OLS. (e) denotes the robust two-stage from Section 4.2. (f) denotes the robust two-stage with robust inverse Mills ratio. The solid line marks the true values of the parameters. The level of censoring is 25%.*

Table C.1: *Bias, Variance and MSE of the classical and robust two-stage estimators at the model and under two types of contamination. The level of censoring is 25%.*

$N = 200$	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
Heckman's 2S									
β_{20}	0.0047	0.0931	0.0931	0.3543	0.1195	0.2451	−0.5172	0.5635	0.8310
β_{21}	−0.0022	0.0195	0.0195	−0.0219	0.0198	0.0203	−0.0027	0.0196	0.0196
$\beta_{2\lambda}$	0.0017	0.0870	0.0870	−0.4037	0.0789	0.2419	0.4009	0.3991	0.5599
Probit MLE + lmrob									
β_{20}	−0.0024	0.1080	0.1080	−0.5364	1.1420	1.4298	−0.5405	0.7029	0.9951
β_{21}	−0.0025	0.0254	0.0255	−0.0073	0.0245	0.0245	−0.0025	0.0252	0.0252
$\beta_{2\lambda}$	0.0071	0.1079	0.1079	0.4306	0.8336	1.0191	0.4159	0.4955	0.6686
Robust probit + OLS									
β_{20}	0.0053	0.0938	0.0938	0.3751	0.0585	0.1993	−0.0884	0.2785	0.2863
β_{21}	−0.0023	0.0195	0.0195	−0.0194	0.0198	0.0202	−0.0031	0.0196	0.0196
$\beta_{2\lambda}$	0.0009	0.0876	0.0876	−0.4235	0.0421	0.2215	0.0718	0.1966	0.2018
Robust IMR									
β_{20}	0.0086	0.1085	0.1086	0.0231	0.0835	0.0840	−0.0832	0.2983	0.3052
β_{21}	−0.0019	0.0199	0.0199	−0.0017	0.0189	0.0189	−0.0021	0.0200	0.0200
$\beta_{2\lambda}$	0.0214	0.1153	0.1157	0.0071	0.0947	0.0947	0.0877	0.2244	0.2321
Robust 2S									
β_{20}	−0.0019	0.1091	0.1091	0.0808	0.1999	0.2065	−0.0966	0.3147	0.3240
β_{21}	−0.0024	0.0256	0.0256	−0.0056	0.0244	0.0245	−0.0025	0.0254	0.0254
$\beta_{2\lambda}$	0.0070	0.1095	0.1096	−0.0969	0.1597	0.1691	0.0771	0.2333	0.2393
Robust 2S + robust IMR									
β_{20}	0.0002	0.1249	0.1249	0.0124	0.1937	0.1939	−0.0924	0.3306	0.3392
β_{21}	0.0010	0.0258	0.0258	0.0002	0.0239	0.0239	−0.0012	0.0253	0.0253
$\beta_{2\lambda}$	0.0261	0.1336	0.1343	0.0323	0.1771	0.1781	0.0936	0.2552	0.2640

Table C.2: *Bias, Variance and MSE of the classical and robust two-stage estimators under contamination of types C and D. The level of censoring is 25%.*

$N = 200$	x_2 is contaminated			y_2 is contaminated		
	Bias	Var	MSE	Bias	Var	MSE
Heckman's 2S						
β_{20}	−0.1051	0.1515	0.1626	0.6070	0.2926	0.6610
β_{21}	−0.8579	0.1191	0.8552	−0.4151	0.1110	0.2834
$\beta_{2\lambda}$	0.0457	0.1467	0.1488	−0.2201	0.1341	0.1826
Probit MLE + lmrob						
β_{20}	−0.0025	0.1166	0.1166	0.0325	0.1010	0.1020
β_{21}	−0.0882	0.0352	0.0430	−0.0181	0.0247	0.0250
$\beta_{2\lambda}$	0.0095	0.1270	0.1271	−0.0069	0.1032	0.1033
Robust probit + OLS						
β_{20}	−0.1036	0.1526	0.1634	0.6080	0.2923	0.6620
β_{21}	−0.8579	0.1191	0.8552	−0.4151	0.1111	0.2834
$\beta_{2\lambda}$	0.0439	0.1484	0.1503	−0.2215	0.1342	0.1832
Robust IMR						
β_{20}	−0.0893	0.1733	0.1813	0.5863	0.2961	0.6398
β_{21}	−0.8598	0.1188	0.8581	−0.4147	0.1111	0.2832
$\beta_{2\lambda}$	0.0521	0.1916	0.1943	−0.1806	0.1685	0.2012
Robust 2S						
β_{20}	−0.0011	0.1164	0.1164	0.0338	0.1014	0.1025
β_{21}	−0.0880	0.0352	0.0430	−0.0184	0.0244	0.0247
$\beta_{2\lambda}$	0.0072	0.1266	0.1266	−0.0084	0.1031	0.1032
Robust 2S + robust IMR						
β_{20}	0.0460	0.1019	0.1040	0.0331	0.1176	0.1187
β_{21}	−0.0808	0.0316	0.0382	−0.0180	0.0250	0.0253
$\beta_{2\lambda}$	−0.0207	0.1089	0.1093	0.0137	0.1307	0.1309

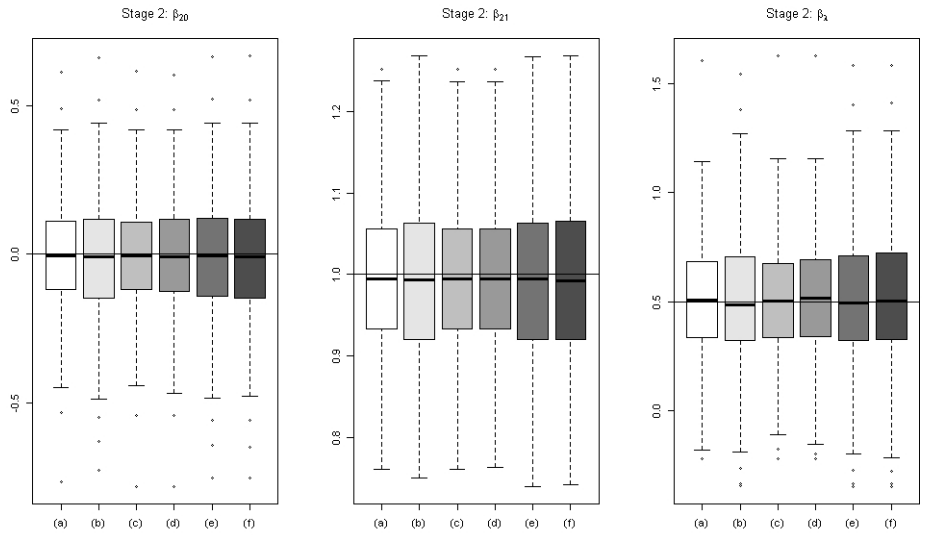

Figure C.4: *Parameter estimation of selection model without contamination. (a) denotes the classical Heckman's estimator. (b) denotes the Classical probit MLE and robust linear regression in the second stage. (c) denotes the robust probit MLE and OLS. (d) denotes the robust inverse Mills ratio from Section 4.3 and OLS. (e) denotes the robust two-stage from Section 4.2. (f) denotes the robust two-stage with robust inverse Mills ratio. The solid line marks the true values of the parameters. The level of censoring is 50%.*

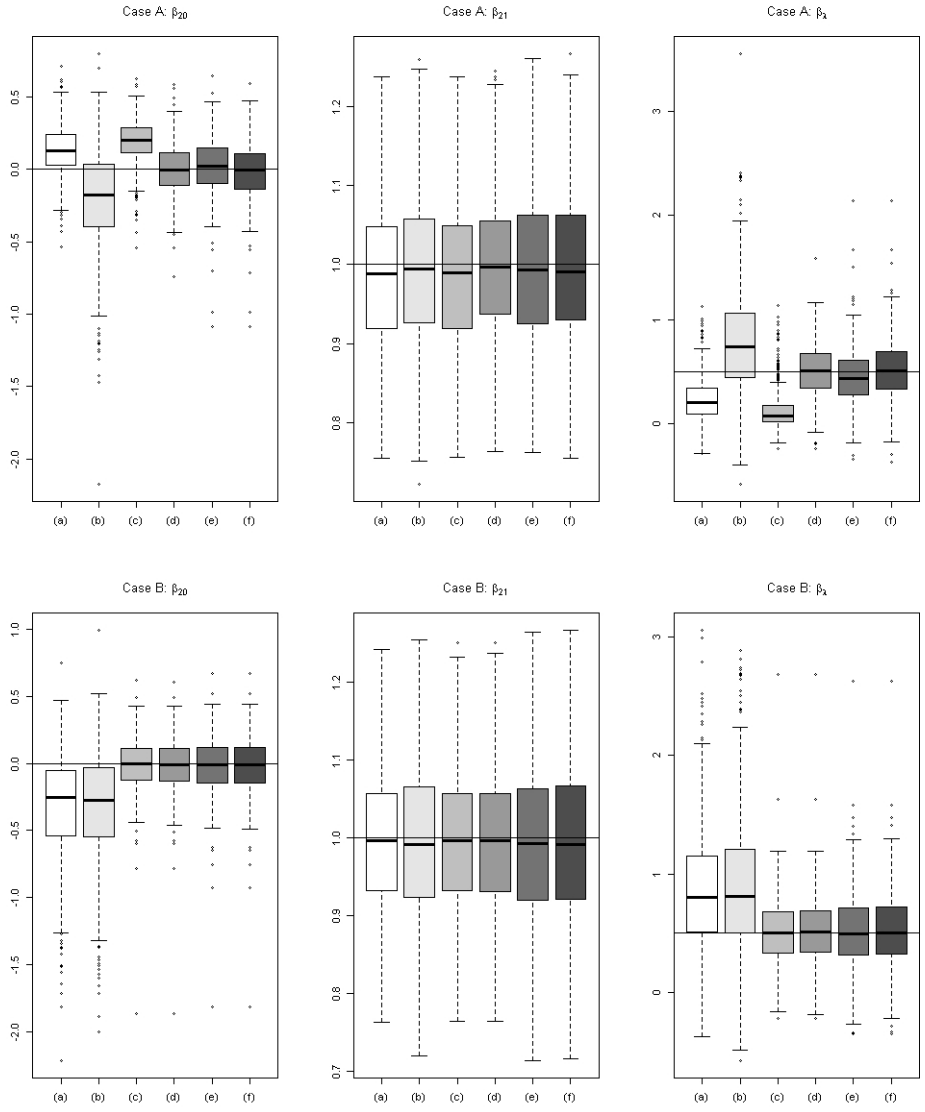

Figure C.5: *Parameter estimation of selection model with contamination. The top panel and the bottom panel correspond to the contamination of Case A and Case B respectively. (a) denotes the classical Heckman's estimator. (b) denotes the probit MLE and robust linear regression in the second stage. (c) denotes the robust probit and OLS. (d) denotes the robust inverse Mills ratio from Section 4.3 and OLS. (e) denotes the robust two-stage from Section 4.2. (f) denotes the robust two-stage with robust inverse Mills ratio. The solid line marks the true values of the parameters. The level of censoring is 50%.*

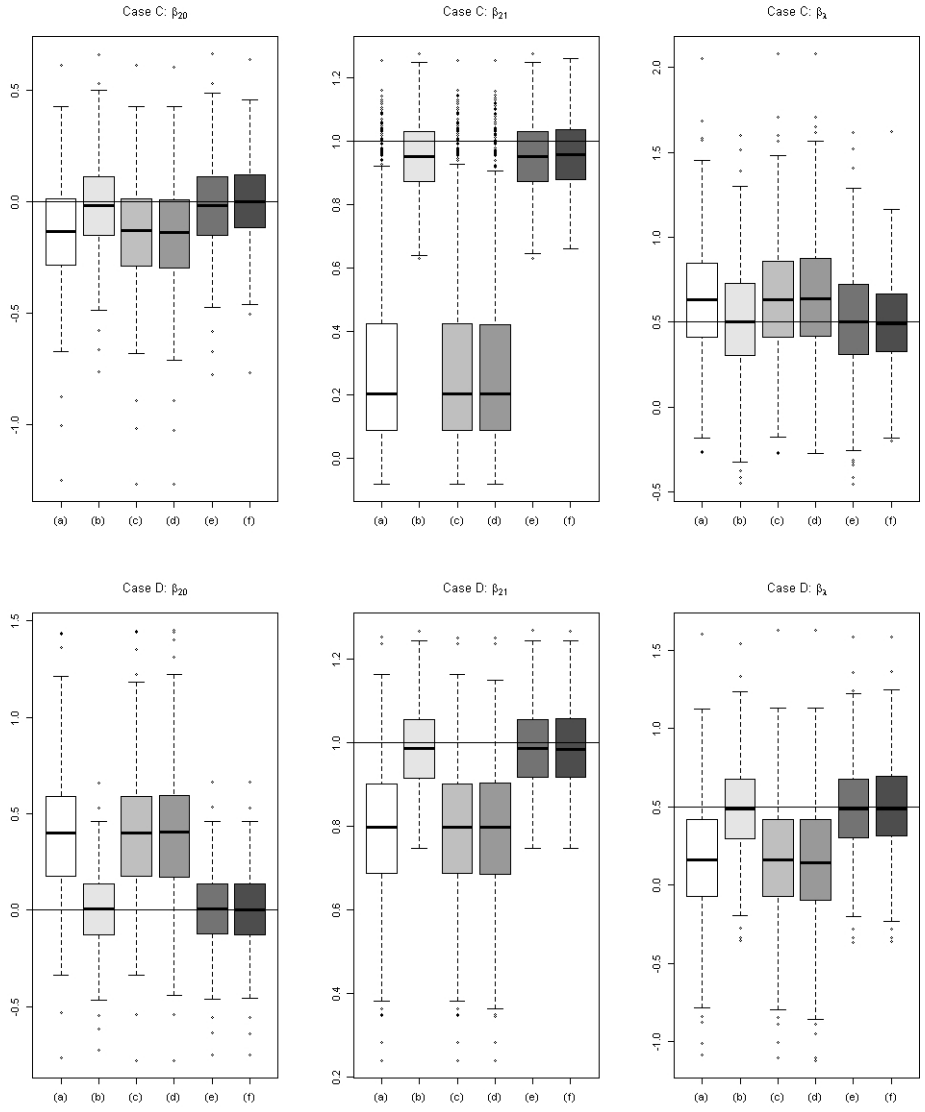

Figure C.6: *Parameter estimation of selection model with contamination. The top panel and the bottom panel correspond to the contamination of Case C and Case D respectively. (a) denotes the classical Heckman's estimator. (b) denotes the probit MLE and robust linear regression in the second stage. (c) denotes the robust probit and OLS. (d) denotes the robust inverse Mills ratio from Section 4.3 and OLS. (e) denotes the robust two-stage from Section 4.2. (f) denotes the robust two-stage with robust inverse Mills ratio. The solid line marks the true values of the parameters. The level of censoring is 50%.*

Table C.3: *Bias, Variance and MSE of the classical and robust two-stage estimators at the model and under two types of contamination. The level of censoring is 50%.*

$N = 200$	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
Heckman's 2S									
β_{20}	−0.0037	0.0298	0.0298	0.1321	0.0278	0.0453	−0.3294	0.1759	0.2844
β_{21}	−0.0034	0.0087	0.0087	−0.0141	0.0088	0.0090	−0.0036	0.0089	0.0089
$\beta_{2\lambda}$	0.0046	0.0623	0.0623	−0.2721	0.0440	0.1181	0.3905	0.2993	0.4518
Probit MLE + lmrob									
β_{20}	−0.0111	0.0360	0.0362	−0.2071	0.1184	0.1613	−0.3358	0.1974	0.3102
β_{21}	−0.0056	0.0103	0.0103	−0.0071	0.0101	0.0102	−0.0055	0.0103	0.0104
$\beta_{2\lambda}$	0.0115	0.0826	0.0828	0.2740	0.2499	0.3250	0.3917	0.3447	0.4982
Robust probit + OLS									
β_{20}	−0.0041	0.0300	0.0300	0.1921	0.0226	0.0595	−0.0096	0.0381	0.0382
β_{21}	−0.0035	0.0087	0.0087	−0.0126	0.0088	0.0090	−0.0034	0.0088	0.0088
$\beta_{2\lambda}$	0.0050	0.0628	0.0628	−0.3618	0.0376	0.1685	0.0107	0.0739	0.0740
Robust IMR									
β_{20}	−0.0066	0.0309	0.0309	−0.0032	0.0285	0.0285	−0.0122	0.0390	0.0392
β_{21}	−0.0035	0.0087	0.0087	−0.0028	0.0084	0.0084	−0.0035	0.0088	0.0088
$\beta_{2\lambda}$	0.0147	0.0654	0.0656	0.0099	0.0625	0.0626	0.0204	0.0765	0.0769
Robust 2S									
β_{20}	−0.0116	0.0362	0.0363	0.0179	0.0364	0.0367	−0.0169	0.0454	0.0457
β_{21}	−0.0054	0.0103	0.0363	−0.0063	0.0101	0.0101	−0.0059	0.0104	0.0104
$\beta_{2\lambda}$	0.0122	0.0830	0.0832	−0.0414	0.0778	0.0798	0.0170	0.0969	0.0972
Robust 2S + robust IMR									
β_{20}	−0.0138	0.0369	0.0371	−0.0147	0.0360	0.0362	−0.0190	0.0459	0.0463
β_{21}	−0.0052	0.0103	0.0103	−0.0041	0.0010	0.0100	−0.0056	0.0104	0.0104
$\beta_{2\lambda}$	0.0110	0.0855	0.0857	0.0238	0.0829	0.0834	0.0256	0.0985	0.0991

Table C.4: *Bias, Variance and MSE of the classical and robust two-stage estimators under contamination of types C and D. The level of censoring is 50%.*

$N = 200$	x_2 is contaminated			y_2 is contaminated		
	Bias	Var	MSE	Bias	Var	MSE
Heckman's 2S						
β_{20}	−0.1395	0.0502	0.0697	0.3873	0.0979	0.2480
β_{21}	−0.6751	0.1069	0.5626	−0.2070	0.0270	0.0699
$\beta_{2\lambda}$	0.1282	0.1004	0.1168	−0.3386	0.1214	0.2361
Probit MLE + lmrob						
β_{20}	−0.0177	0.0397	0.0401	0.0047	0.0361	0.0361
β_{21}	−0.0511	0.0136	0.0162	−0.0130	0.0105	0.0106
$\beta_{2\lambda}$	0.0190	0.0995	0.0998	−0.0019	0.0831	0.0831
Robust probit + OLS						
β_{20}	−0.1396	0.0504	0.0698	0.3872	0.0981	0.2481
β_{21}	−0.6751	0.1068	0.5626	−0.2070	0.0270	0.0699
$\beta_{2\lambda}$	0.1283	0.1008	0.1173	−0.3386	0.1217	0.2364
Robust IMR						
β_{20}	−0.1445	0.0522	0.0731	0.3927	0.1022	0.2565
β_{21}	−0.6752	0.1067	0.5625	−0.2070	0.0270	0.0699
$\beta_{2\lambda}$	0.1443	0.1065	0.1273	−0.3472	0.1313	0.2518
Robust 2S						
β_{20}	−0.0175	0.0398	0.0402	0.0045	0.0363	0.0364
β_{21}	−0.0511	0.0135	0.0162	−0.0129	0.0105	0.0107
$\beta_{2\lambda}$	0.0179	0.1004	0.1008	−0.0019	0.0837	0.0837
Robust 2S + robust IMR						
β_{20}	0.0024	0.0323	0.0323	0.0024	0.0369	0.0369
β_{21}	−0.0448	0.0131	0.0152	−0.0127	0.0105	0.0106
$\beta_{2\lambda}$	−0.0113	0.0759	0.0760	0.0070	0.0855	0.0855

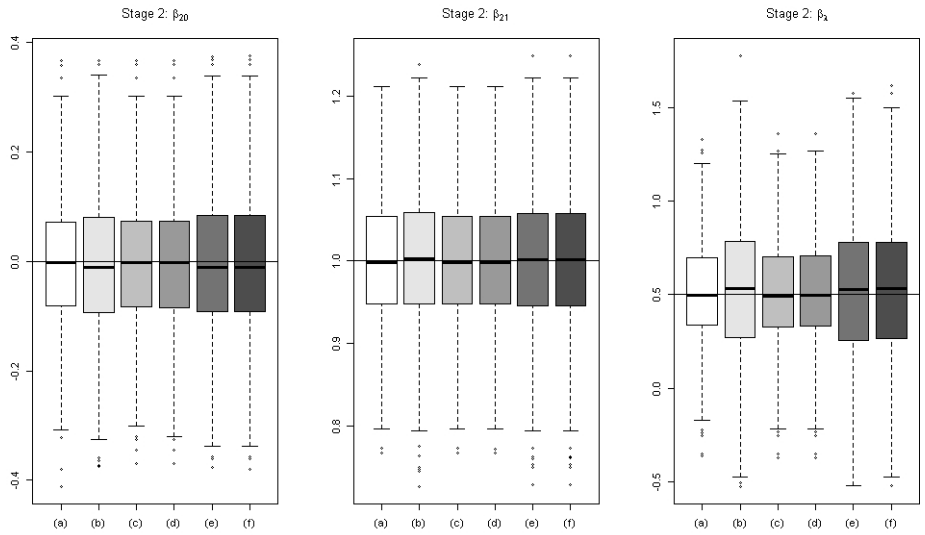

Figure C.7: *Parameter estimation of selection model without contamination. (a) denotes the classical Heckman's estimator. (b) denotes the Classical probit MLE and robust linear regression in the second stage. (c) denotes the robust probit MLE and OLS. (d) denotes the robust inverse Mills ratio from Section 4.3 and OLS. (e) denotes the robust two-stage from Section 4.2. (f) denotes the robust two-stage with robust inverse Mills ratio. The solid line marks the true values of the parameters.*

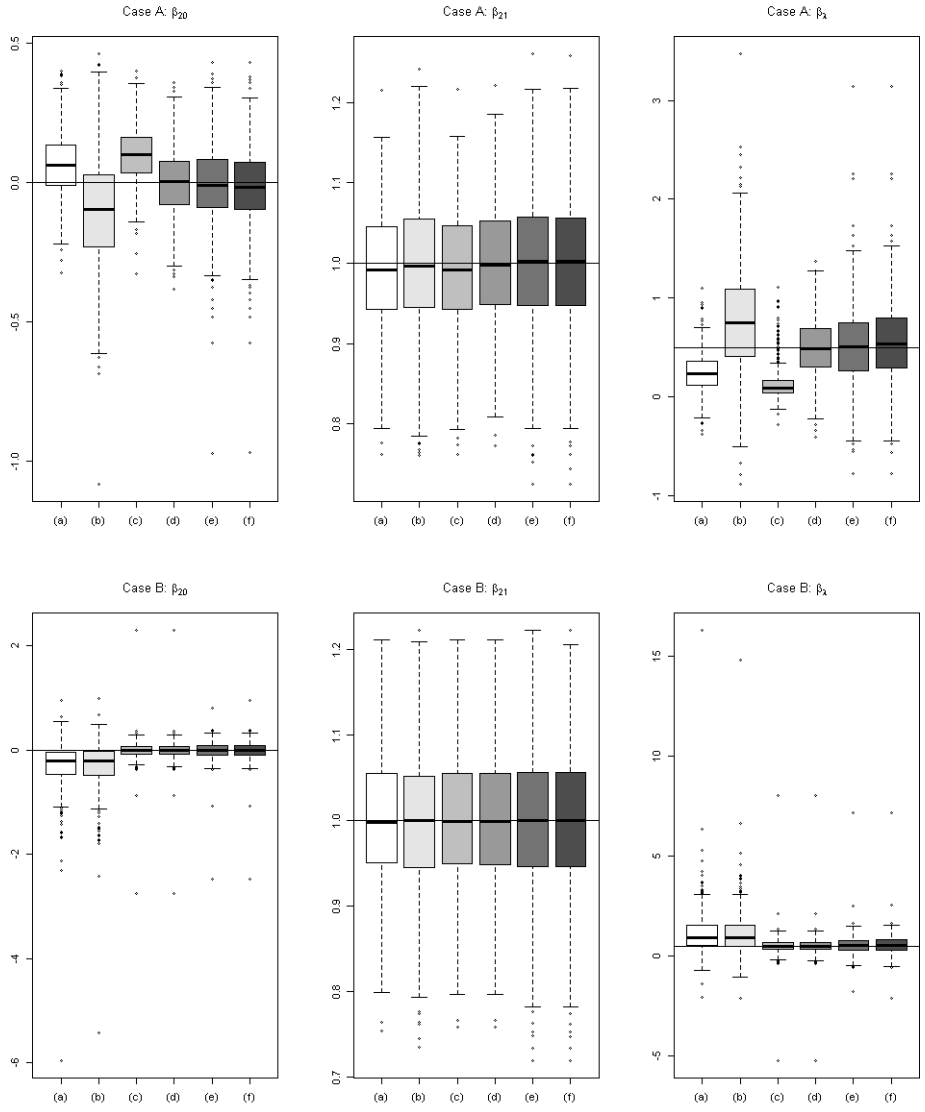

Figure C.8: *Parameter estimation of selection model with contamination. The top panel and the bottom panel correspond to the contamination of Case A and Case B respectively. (a) denotes the classical Heckman's estimator. (b) denotes the probit MLE and robust linear regression in the second stage. (c) denotes the robust probit and OLS. (d) denotes the robust inverse Mills ratio from Section 4.3 and OLS. (e) denotes the robust two-stage from Section 4.2. (f) denotes the robust two-stage with robust inverse Mills ratio. The solid line marks the true values of the parameters.*

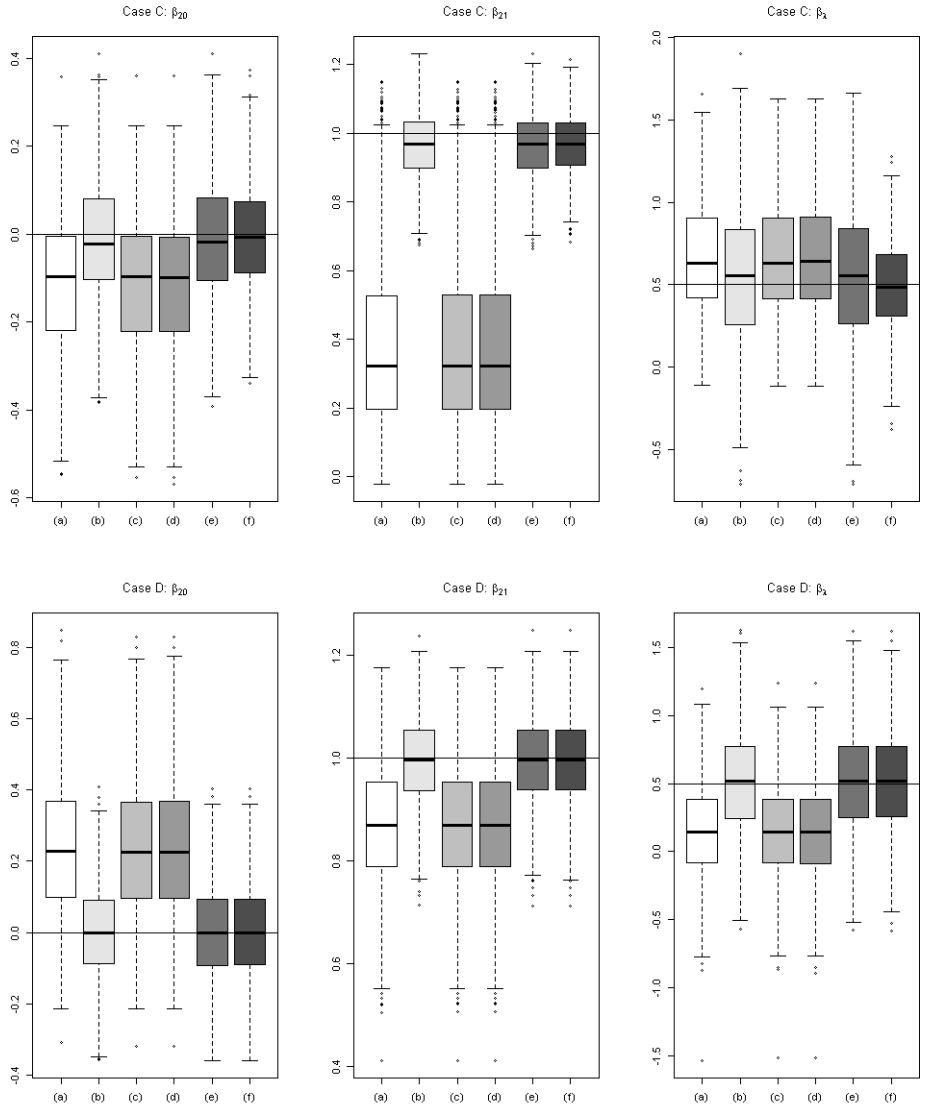

Figure C.9: *Parameter estimation of selection model with contamination. The top panel and the bottom panel correspond to the contamination of Case C and Case D respectively. (a) denotes the classical Heckman's estimator. (b) denotes the probit MLE and robust linear regression in the second stage. (c) denotes the robust probit and OLS. (d) denotes the robust inverse Mills ratio from Section 4.3 and OLS. (e) denotes the robust two-stage from Section 4.2. (f) denotes the robust two-stage with robust inverse Mills ratio. The solid line marks the true values of the parameters.*

Table C.5: *Bias, Variance and MSE of the classical and robust two-stage estimators at the model and under two types of contamination. The level of censoring is 75%.*

$N = 200$	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
Heckman's 2S									
β_{20}	−0.0037	0.0141	0.0141	0.0620	0.0123	0.0161	−0.2946	0.2130	0.2998
β_{21}	−0.0003	0.0056	0.0056	−0.0080	0.0058	0.0058	−0.0010	0.0057	0.0057
$\beta_{2\lambda}$	0.0104	0.0759	0.0760	−0.2551	0.0379	0.1030	0.6381	1.2574	1.6646
Probit MLE + lmrob									
β_{20}	−0.0099	0.0168	0.0169	−0.1065	0.0389	0.0502	−0.2962	0.2157	0.3035
β_{21}	−0.0012	0.0072	0.0072	−0.0031	0.0070	0.0070	−0.0022	0.0070	0.0070
$\beta_{2\lambda}$	0.0264	0.1342	0.1349	0.2743	0.2925	0.3678	0.6254	1.2695	1.6607
Robust probit + OLS									
β_{20}	−0.0037	0.0141	0.0142	0.0997	0.0101	0.0201	−0.0078	0.0416	0.0417
β_{21}	−0.0003	0.0056	0.0056	−0.0072	0.0058	0.0058	−0.0008	0.0056	0.0056
$\beta_{2\lambda}$	0.0110	0.0767	0.0768	−0.3654	0.0288	0.1624	0.0218	0.2616	0.2621
Robust IMR									
β_{20}	−0.0046	0.0141	0.0142	−0.0001	0.0141	0.0141	−0.0087	0.0416	0.0417
β_{21}	−0.0003	0.0056	0.0056	−0.0005	0.0056	0.0056	−0.0008	0.0056	0.0056
$\beta_{2\lambda}$	0.0154	0.0765	0.0767	0.0035	0.0796	0.0796	0.0264	0.2614	0.2621
Robust 2S									
β_{20}	−0.0097	0.0169	0.0170	0.0122	0.0216	0.0218	−0.0161	0.0328	0.0331
β_{21}	−0.0011	0.0072	0.0072	−0.0012	0.0073	0.0073	−0.0019	0.0073	0.0073
$\beta_{2\lambda}$	0.0245	0.1352	0.1358	0.0193	0.1627	0.1631	0.0418	0.2450	0.2467
Robust 2S + robust IMR									
β_{20}	−0.0099	0.0170	0.0171	−0.0195	0.0214	0.0218	−0.0160	0.0333	0.0336
β_{21}	−0.0012	0.0072	0.0072	−0.0007	0.0072	0.0072	−0.0018	0.0073	0.0073
$\beta_{2\lambda}$	0.0257	0.1359	0.1366	0.0511	0.1700	0.1726	0.0427	0.2477	0.2495

Table C.6: *Bias, Variance and MSE of the classical and robust two-stage estimators under contamination of types C and D. The level of censoring is 75%.*

$N = 200$	x_2 is contaminated			y_2 is contaminated		
	Bias	Var	MSE	Bias	Var	MSE
Heckman's 2S						
β_{20}	−0.1116	0.0224	0.0348	0.2377	0.0373	0.0938
β_{21}	−0.5934	0.0804	0.4326	−0.1331	0.0149	0.0326
$\beta_{2\lambda}$	0.1625	0.1094	0.1358	−0.3649	0.1312	0.2644
Probit MLE + lmrob						
β_{20}	−0.0162	0.0185	0.0187	−0.0010	0.0170	0.0170
β_{21}	−0.0353	0.0093	0.0105	−0.0053	0.0073	0.0073
$\beta_{2\lambda}$	0.0401	0.1696	0.1712	0.0086	0.1347	0.1348
Robust probit + OLS						
β_{20}	−0.1119	0.0225	0.0350	0.2367	0.0371	0.0932
β_{21}	−0.5934	0.0804	0.4326	−0.1331	0.0149	0.0326
$\beta_{2\lambda}$	0.1636	0.1109	0.1377	−0.3625	0.1306	0.2620
Robust IMR						
β_{20}	−0.1137	0.0225	0.0355	0.2370	0.0375	0.0937
β_{21}	−0.5933	0.0804	0.4325	−0.1331	0.0149	0.0326
$\beta_{2\lambda}$	0.1714	0.1106	0.1400	−0.3633	0.1327	0.2647
Robust 2S						
β_{20}	−0.0160	0.0186	0.0188	−0.0010	0.0172	0.0172
β_{21}	−0.0355	0.0092	0.0105	−0.0053	0.0072	0.0073
$\beta_{2\lambda}$	0.0383	0.1693	0.1708	0.0085	0.1353	0.1353
Robust 2S + robust IMR						
β_{20}	0.0054	0.0152	0.0152	−0.0012	0.0172	0.0172
β_{21}	−0.0322	0.0083	0.0093	−0.0054	0.0073	0.0073
$\beta_{2\lambda}$	−0.0066	0.0775	0.0775	0.0094	0.1355	0.1356

Bibliography

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.

Arellano-Valle, R. B. and Genton, M. G. (2010), "Multivariate Extended Skew-t Distributions and Related Families," *Metron*, 68, 201–234.

Banasik, J., Crook, J., and Thomas, L. (2003), "Sample Selection Bias in Credit Scoring Models," *Journal of the Operational Research Society*, 54, 822–832.

Berk, R. A. (1983), "An Introduction to Sample Selection Bias in Sociological Data," *American Sociological Review*, 48, 386–398.

Bushway, S., Johnson, B. D., and Slocum, L. A. (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.

Cohen Freue, G. V., Ortiz-Molina, H., and Zamar, R. H. (2011), "A Natural Robustifcation of the Ordinary Instrumental Variables Estimator," *Manuscript*.

Collier, D. and Mahoney, J. (1996), "Insights and Pitfalls: Selection Bias in Qualitative Research," *World Politics*, 49, 56–91.

Das, M., Newey, W. K., and Vella, F. (2003), "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

Davidson, R. and MacKinnon, J. G. (1993), *Estimation and Inference in Econometrics*, Oxford University Press.

de Luna, X. and Genton, M. G. (2001), "Robust Simulation-Based Estimation of ARMA Models," *Journal of Computational and Graphical Statistics*, 10, 370–387.

Di Falco, S., Veronesi, M., and Yesuf, M. (2011), "Does Adaptation to Climate Change Provide Food Security? A Micro-Perspective from Ethiopia," *American Journal of Agricultural Economics*, 93, 829–846.

Dollinger, M. B. and Staudte, R. G. (1991), "Influence Functions of Iteratively Reweighted Least Squares Estimators," *Journal of the American Statistical Association*, 86, 709–716.

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.

Ferrari, D. and La Vecchia, D. (2012), "On Robust Estimation via Pseudo-Additive Information," *Biometrika*, 99, 238–244.

Gallant, R. A. and Nychka, D. W. (1987), "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.

Genton, M. G. (2001), *Robustness Problems in the Analysis of Spatial Data*, Springer Lecture Notes in Statistics, pp. 21–37.

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.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2012), *mvtnorm: Multivariate Normal and t Distributions*, r package version 0.9-9993.

Greene, W. H. (1981), "Sample Selection Bias as a Specification Error: Comment," *Econometrica*, 49, 795–798.

— (2008), *Econometric Analysis*, Upper Saddle River, NJ: Prentice–Hall, 6th ed.

Gronau, R. (1974), "Wage Comparisons - A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.

Hampel, F. (1971), "A General Qualitative Definition of Robustness," *Annals of Mathematical Statistics*, 42, 1887–1896.

— (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.

Hardin, J. W. (2002), "The Robust Variance Estimator for Two-stage Models," *The Stata Journal*, 2, 253–266.

Heckman, J. J. (1974), "Shadow prices, market wages, and labor supply," *Econometrica*, 42, 679–694.

— (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5/4, 475–492.

— (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M.-P. (2009), *Robust Methods in Biostatistics*, Chippenham: John Wiley and Sons.

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.

— (1973), "Robust regression: Asymptotics, Conjectures, and Monte Carlo," *Annals of Statistics*, 1, 799–821.

— (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.

Hubert, M. and Rousseeuw, P. J. (1997), "Robust Regression with Both Continuous and Binary Regressors," *Journal of Statistical Planning and Inference*, 57, 153–163.

Jorgensen, M. A. (1993), "Influence Functions for Iteratively Defined Statistics," *Biometrika*, 80, 253–265.

Jureckova, J. and Picek, J. (2006), *Robust Statistical Methods with R*, Boca Raton, FL: Chapman & Hall/CRC.

Kenny, L. W., Lee, L.-F., Maddala, G., and Trost, R. (1979), "Returns to College Education: An Investigation of Self-Selection Bias Based on the Project Talent Data," *International Economic Review*, 20, 775–789.

Kim, T.-K. and Muller, C. (2007), "Two-stage Huber estimation," *Journal of Statistical Planning and Inference*, 137, 405–418.

Kiriazidou, E. (1997), "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335–1364.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000), *Continuous Multivariate Distributions, Volume 1: Models and Applications*, New York: John Wiley and Sons.

Krasker, W. S. (1986), "Two-Stage Bounded-Influence Estimators for Simultaneous-Equations Models," *Journal of Business and Economic Statistics*, 4, 437–444.

Krishnakumar, J. and Ronchetti, E. (1997), "Robust Estimators for Simultaneous Equations Models," *Journal of Econometrics*, 78, 295–314.

Künsch, H. R. (1984), "Infinitesimal robustness for autoregressive processes," *The Annals of Statistics*, 12, 843–863.

La Vecchia, D., Ronchetti, E., and Trojani, F. (2012), "Higher-Order Infinitesimal Robustness," *Journal of the American Statistical Association*, 107, 1546–1557.

Lee, L.-F., Maddala, G., and Trost, R. (1980), "Asymptotic Covariance Matrices of Two-Stage Probit and Two-Stage Tobit Methods for Simultaneous Equations Models With Selectivity," *Econometrica*, 48, 491–503.

Lehmann, E. L. (1999), *Elements of Large Sample Theory*, New York: Springer.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Hoboken, NJ: John Wiley and Sons, 2nd ed.

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

Mallows, C. L. (1975), "On Some Topics in Robustness," *Bell Telephone Laboratories*.

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.

Maronna, R. A. and Yohai, V. J. (1997), "Robust Estimation in Simultaneous Equations Models," *Journal of Statistical Planning and Inference*, 57, 233–244.

— (2000), "Robust Regression with Both Continuous and Categorical Predictors," *Journal of Statistical Planning and Inference*, 89, 197–214.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall/CRC.

Melino, A. (1982), "Testing for Sample Selection Bias," *Review of Economic Studies*, XLIX, 151–153.

Michel-Kerjan, E., Raschky, P., and Kunreuther, H. (2011), "Corporate Demand for Insurance: an Empirical Analysis of the U.S. Market for Catastrophe and Non-Catastrophe Risks," *NBER Working Paper*, 17403.

Murphy, K. M. and Topel, R. H. (1985), "Estimation and Inference in Two-step Econometric Models," *Journal of Business and Economic Statistics*, 3, 370–379.

Nawata, K. (1994), "Estimation of Sample Selection Bias Models by the Maximum Likelihood Estimator and Heckman's Two-step Estimator," *Economics Letters*, 45, 33–40.

Newey, W. K. (1984), "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters*, 14, 201–206.

— (2009), "Two-step Series Estimation of Sample Selection Models," *Econometrics Journal*, 12, S217–S229.

Ogundimu, E. O. and Hutton, J. L. (2012), "A General Sample Selection Model with Skew-Normal Distribution," *Working Paper*.

Olsen, R. J. (1982), "Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator," *International Economic Review*, 23, 223–240.

Paarsch, H. J. (1984), "A Monte Carlo Comparison of Estimators for Censored Regression Models," *Journal of Econometrics*, 24, 197–213.

Pagan, A. (1986), "Two Stage and Related Estimators and Their Applications," *Review of Economic Studies*, LIII, 517–538.

Peracchi, F. (1990), "Bounded-Influence Estimators for the Tobit Model," *Journal of Econometrics*, 44, 107–126.

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.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2012), *robustbase: Basic Robust Statistics*, r package version 0.9-4.

Rousseeuw, P. J. (1985), *Multivariate Estimation with High Breakdown Point*, Reidel, Dordrecht, pp. 283–297.

Rousseeuw, P. J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223.

Salazar, L. (2008), "A Robustness Study of Heckman's Model," *Master's Thesis, University of Geneva, Switzerland*.

Semykina, A. and Wooldridge, J. M. (2010), "Estimating Panel Data Models in the Presence of Endogeneity and Selection," *Journal of Econometrics*, 157, 375–380.

Stigler, S. M. (1973), "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920," *Journal of the American Statistical Assiciation*, 68, 872–879.

— (2010), "The Changing History of Robustness," *The American Statistician*, 64, 277–281.

Stolzenberg, R. M. and Relles, D. A. (1997), "Tools for Intuition about Sample Selection Bias and Its Correction," *American Sociological Review*, 62, 494–507.

Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24–36.

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.

Vella, F. and Verbeek, M. (1999), "Two-Step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias," *Journal of Econometrics*, 90, 239–263.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, New York: Springer, 4th ed., iSBN 0-387-95457-0.

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.

Wooldridge, J. M. (1995), "Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions," *Journal of Econometrics*, 68, 115–132.

— (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: The MIT Press.

Yeap, B. Y. and Davidian, M. (2001), "Robust Two-Stage Estimation in Hierarchical Nonlinear Models," *Biometrics*, 57, 266–272.

Yohai, V. J. (1987), "High Breakdown-Point and High Efficiency Robust Estimates for Regression," *Annals of Statistics*, 15, 642–656.

Zadrozny, B. (2004), "Learning and Evaluating Classifiers under Sample Selection Bias," *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada.

Zhelonkin, M., Genton, M. G., and Ronchetti, E. (2012), "On the Robustness of Two-Stage Estimators," *Statistics & Probability Letters*, 82, 726–732.

— (2013), "Robust Inference in Sample Selection Models," *Submitted Manuscript*.

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.