Statistics and Probability Letters 82 (2012) 726–732

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On the robustness of two-stage estimators

Mikhail Zhelonkin[a,*], Marc G. Genton[b], Elvezio Ronchetti[a]

[a] Research Center for Statistics and Department of Economics, University of Geneva, Blv. du Pont-d'Arve 40, CH-1211 Geneva 4, Switzerland

[b] Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA

ARTICLE INFO

Article history:
Received 9 August 2011
Received in revised form 18 December 2011
Accepted 18 December 2011
Available online 28 December 2011

Keywords:
Asymptotic variance
Bounded influence function
M-estimator
Change-of-variance function
Two-stage least squares

ABSTRACT

The aim of this note is to provide a general framework for the analysis of the robustness properties of a broad class of two-stage models. 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 and the two-stage least squares estimator.

1. Introduction

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. Robust statistics deals with such problems and develops methods that are more reliable in the presence of such deviations from the model. Standard general books are Huber (1981), Hampel et al. (1986) and Maronna et al. (2006).

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 note 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.

* Corresponding author. Tel.: +41 22 3798214.
E-mail addresses: Mikhail.Zhelonkin@unige.ch (M. Zhelonkin), genton@stat.tamu.edu (M.G. Genton), Elvezio.Ronchetti@unige.ch (E. Ronchetti).

0167-7152/$ – see front matter Crown Copyright © 2011 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2011.12.014

This note is structured as follows. In Section 2 we derive the influence function, change-of-variance function and the asymptotic variance for the two-stage M-estimator and provide our interpretations of them. Section 3 provides some specific examples of applications. Section 4 offers some concluding remarks.

2. Main results

2.1. Two-stage estimators

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 $z_i^{(j)} = (x_{ji}, y_{ji})$, $j = 1, 2$, $i = 1, \dots, N$, 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:

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

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

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.

2.2. Influence function

For a given functional $T(F)$, the influence function (IF) is defined by Hampel (1974) as $\mathrm{IF}(z; T, F) = \lim_{\epsilon\to 0}[T(F_\epsilon) - T(F)]/\epsilon$, 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 (2), the functional $T(F_\epsilon)$ is defined by:

$$\int \Psi_2(z^{(2)}; h(z^{(1)}; S(F_\epsilon)), T(F_\epsilon))dF_\epsilon = 0 \tag{3}$$

and the derivative of (3) with respect to ϵ evaluated at $\epsilon = 0$ is

$$\frac{\partial}{\partial\epsilon}(1-\epsilon)\int \Psi_2(\tilde{z}^{(2)}; h(\tilde{z}^{(1)}; S(F_\epsilon)), T(F_\epsilon))dF(\tilde{z})\Big|_{\epsilon=0} + \frac{\partial}{\partial\epsilon}\epsilon\int \Psi_2(\tilde{z}^{(2)}; h(\tilde{z}^{(1)}; S(F_\epsilon)), T(F_\epsilon))d\Delta_z\Big|_{\epsilon=0} = 0. \tag{4}$$

The second term of (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

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

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. $\frac{\partial}{\partial\epsilon}S(F_\epsilon)|_{\epsilon=0} = \mathrm{IF}(z; S, F)$.

Combining the derivatives of the two terms of (4), we obtain the IF of the two-stage M-estimator:

$$\mathrm{IF}(z; T, F) = M^{-1}\left(\Psi_2(z^{(2)}; h(z^{(1)}; S(F)), T(F)) + \int \frac{\partial}{\partial\theta}\Psi_2(\tilde{z}^{(2)}; \theta, T(F))\frac{\partial}{\partial\eta}h(\tilde{z}^{(1)}; \eta)dF(\tilde{z}) \cdot \mathrm{IF}(z; S, F)\right), \quad (5)$$

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 (5) 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 $\mathrm{IF}(z; S, F)$ drops out and the IF of the estimator of the second stage collapses to $\mathrm{IF}((x_2, y_2); T, F)$, 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 (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 (5) 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.

2.3. 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

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

Denote the components of the IF as follows:

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

Using the expression of IF in (5) and integrating, we obtain the asymptotic variance of $\hat{\beta}_2$:

$$V(T, F) = M^{-1}\int \left(a(z)a(z)^\top + a(z)b(z)^\top + b(z)a(z)^\top + b(z)b(z)^\top\right) dF(z)\, M^{-1}. \quad (6)$$

The form (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 (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.

2.4. 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 $\mathrm{CVF}(z; T, F) = [(\partial/\partial\epsilon)V(T, (1-\epsilon)F + \epsilon\Delta_z)]_{\epsilon=0}$, for all z where this expression exists; see Hampel et al. (1981). 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:

$$\mathrm{CVF}(z; S, T, F) = V(T, F) - M^{-1}\left[\int D dF(z) + \frac{\partial}{\partial\theta}\Psi_2(z^{(2)}; h(z^{(1)}; S(F)), \theta)\right]V(T, F)$$
$$+ M^{-1}\int \left(Aa(z)^\top + Ba(z)^\top + Ab(z)^\top + Bb(z)^\top\right) dF(z)M^{-1}$$

$$+ M^{-1}\int \left(a(z)A^\top + b(z)A^\top + a(z)B^\top + b(z)B^\top\right) dF(z)M^{-1}$$
$$+ M^{-1}\left(a(z)a(z)^\top + a(z)b(z)^\top + b(z)a(z)^\top + b(z)b(z)^\top\right) M^{-1}$$
$$- V(T,F)\left[\int DdF(z) + \frac{\partial}{\partial\theta}\Psi_2(z^{(2)}; h(z^{(1)}; S(F)), \theta)\right] M^{-1}, \quad (7)$$

where D is a matrix with elements

$$D_{ij} = \left(\frac{\partial}{\partial h}\frac{\partial \Psi_{2i}(z^{(2)}; h, \theta)}{\partial \theta_j}\right)^\top \frac{\partial h(z^{(1)}; s)}{\partial s}\mathrm{IF}(z; S, F) + \left(\frac{\partial}{\partial \theta}\frac{\partial \Psi_{2i}(z^{(2)}; h, \theta)}{\partial \theta_j}\right)^\top \mathrm{IF}(z, T, F),$$
$$A = \frac{\partial}{\partial h}\Psi_2(z^{(2)}; h, T(F))\frac{\partial}{\partial s}h(z^{(1)}; s)\mathrm{IF}(z, S, F) + \frac{\partial}{\partial \theta}\Psi_2(z^{(2)}; h(z^{(1)}; S(F)), \theta)\cdot \mathrm{IF}(z; T, F).$$

The matrix B has the following form

$$B = \int R_1 \frac{\partial}{\partial s}h(z^{(1)}; s)dF\mathrm{IF}(z, S, F) + \int \frac{\partial}{\partial h}\Psi_2(z^{(2)}; h, T(F))R_2 dF\mathrm{IF}(z, S, F)$$
$$- \int \frac{\partial}{\partial h}\Psi_2(z^{(2)}; h, T(F))\frac{\partial}{\partial s}h(z^{(1)}; s)dFM_1^{-1}\left[\int DdF + \frac{\partial}{\partial\theta}\Psi_1(z^{(1)}; \theta)\right]\mathrm{IF}(z, S, F)$$
$$+ \int \frac{\partial}{\partial h}\Psi_2(z^{(2)}; h, T(F))\frac{\partial}{\partial s}h(z^{(1)}; s)dFM_1^{-1}\frac{\partial}{\partial\theta}\Psi_1(z^{(1)}; \theta)\mathrm{IF}(z, S, F)$$
$$+ \frac{\partial}{\partial h}\Psi_2(z^{(2)}; h, T(F))\frac{\partial}{\partial s}h(z^{(1)}; s)\cdot \mathrm{IF}(z; S, F),$$

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

$$R^{(1)}_{ij} = \left(\frac{\partial}{\partial h}\frac{\partial \Psi_{2i}(z^{(2)}; h, T(F))}{\partial h_j}\right)^\top \frac{\partial}{\partial s}h(z^{(1)}; s)\mathrm{IF}(z; S, F) + \left(\frac{\partial}{\partial \theta}\frac{\partial \Psi_{2i}(z^{(2)}; h, \theta)}{\partial h_j}\right)^\top \mathrm{IF}(z; T, F),$$

$R^{(2)}$ is the matrix with elements $R^{(2)}_{ij} = \left(\frac{\partial}{\partial s}\frac{\partial h_i(z^{(1)}; s)}{\partial s_j}\right)^\top \mathrm{IF}(z; S, F)$, 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 but is longer.

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 (in press). The boundedness of the CVF function is determined by the boundedness of the IF's.

3. Examples

3.1. Two-stage maximum likelihood estimators

Eq. (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

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

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 (6) then we immediately obtain the result in Murphy and Topel (1985).

3.2. Two-stage least squares estimators

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 $x^\top = (x^{(1)\top}, x^{(2)})$. 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 (1)–(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:

$$\Psi_1((w, x^{(2)}); S(F)) = \left(x^{(2)} - w^\top\alpha\right) w$$

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

Using the general formula (5) we compute the IF for 2SLS as a special case:

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

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

$$\mathrm{IF}(z; T, F) = M^{-1}\left\{\left(y - (x^{(1)})^\top\beta_1 - w^\top\alpha\beta_2\right)\begin{pmatrix} x^{(1)} \\ w^\top\alpha \end{pmatrix} - \left(\int \begin{pmatrix} x^{(1)}\beta_2 \\ w^\top\alpha\beta_2 \end{pmatrix} w^\top dF(z)\right)\cdot \mathrm{IF}(z; S, F)\right\}, \tag{8}$$

where

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

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 (8) we can obtain the asymptotic variance of the 2SLS estimator using formula (6).

3.3. Two-stage least squares simulations

Consider the model described in Section 3.2. 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 $u \sim N(0, 1)$, $x^{(2)} \sim N(0, 1)$, $\mathrm{corr}(x^{(2)}, u) = -0.6$, $\beta_2 = 1$, and an intercept $\beta_0 = 0$. There exists one instrumental variable w, such that $\mathrm{corr}(x^{(2)}, w) = 0.6$ and $\mathrm{corr}(w, u) = 0$. 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 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 (7) depends on the IF of the 2SLS estimator and is unbounded. The unshaded boxplots in Fig. 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 note, the problem of outliers in $x^{(2)}$ can be treated by using a 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 1 and Fig. 1. The grey shaded boxplots are the robust versions of 2SLS based on *MM*-estimators introduced by Yohai (1987).

Table 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

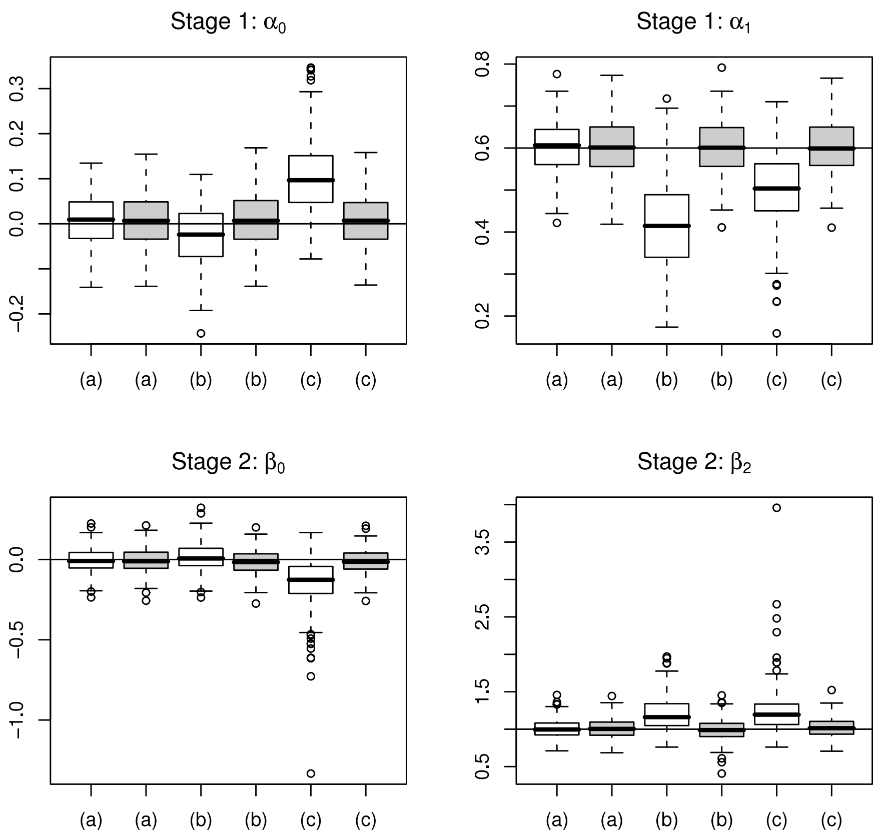

Fig. 1. 2SLS. Unshaded boxplots correspond to the classical 2SLS and shaded boxplots correspond to the robust 2SLS. Case (a) is without contamination, (b) is with contamination of w, and (c) is with contamination of $x^{(2)}$. The top panels correspond to the auxiliary regression, the bottom panels to the regression of interest. Horizontal lines mark the true values of the parameters.

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 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.

4. Discussion

The results of Section 2 provide a general framework for robust estimation and inference in two-stage models. In Section 3 we presented two 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 one important application is in time series when the deterministic and stochastic parts are modeled separately. In this case the IF's of the standard estimators based on MLE or LSE are unbounded, which means that in presence of outliers the estimators of trend and stochastic process can be biased. The same holds for spatial statistics. Also, we should note that many 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. In all these situations the robustness issue can become crucial for estimation and inference.

Acknowledgment

The second author's research was partially supported by NSF grants DMS-1007504 and DMS-1100492, and by Award No. KUS-C1-016-04 made by King Abdullah University of Science and Technology (KAUST).

References

Cohen-Freue, G.V., Ortiz-Molina, H., Zamar, R.H., 2011. A natural robustification of the ordinary instrumental variables estimator. Manuscript.
Dollinger, M.B., Staudte, R.G., 1991. Influence functions of iteratively reweighted least squares estimators. Journal of the American Statistical Association 86, 709–716.
Ferrari, D., La Vecchia, D., 2011. On robust estimation via pseudo-additive information. Biometrika (in press).
Genton, M.G., 2001. Robustness problems in the analysis of spatial data. In: Moore, M. (Ed.), Spatial Statistics: Methodological Aspects and Applications. In: Lecture Notes in Statistics, Springer, pp. 21–37.
Hampel, F., 1974. The influence curve and its role in robust estimation. Journal of the American Statistical Association 69, 383–393.
Hampel, F., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. John Wiley and Sons.
Hampel, F., Rousseeuw, P.J., 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.
Huber, P.J., 1981. Robust Statistics. John Wiley and Sons, New York.
Jorgensen, M.A., 1993. Influence functions for iteratively defined statistics. Biometrika 80, 253–265.
Kim, T.-K., Muller, C., 2007. Two-stage Huber estimation. Journal of Statistical Planning and Inference 137, 405–418.
Maronna, R.A., Martin, G.R., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. John Wiley and Sons.
Murphy, K.M., Topel, R.H., 1985. Estimation and inference in two-step econometric models. Journal of Business and Economic Statistics 3, 370–379.
Newey, W.K., 1984. A method of moments interpretation of sequential estimators. Economics Letters 14, 201–206.
Pagan, A., 1986. Two stage and related estimators and their applications. Review of Economic Studies LIII, 517–538.
Yeap, B.Y., 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.