ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis

[programme Econometrie en Operationele Research]

On robust instrumental variables estimator performance in contaminated environments.

Name student: Jeroen van Druten

Student ID number: 478791jd

Supervisor: Klooster, J.

Second assessor: dr. (Wendun) W. Wang

Date final version: Sunday, July 5 2020

On robust instrumental variables estimator performance in contaminated environments.

Jeroen van Druten - 478791jd

Abstract

Instrumental variables estimators are used when one or more explanatory variables are correlated with the error term. Outliers could cause the instrumental variables estimator to become inconsistent. Therefore, outlier robust instrumental variables estimators are used to account for outliers in the data. In this paper the performance of the robust instrumental variable estimator of Freue et al. (2013), RIV, is tested against other instrumental variables estimators, both robust and non-robust. We find that RIV performs best in our simulation study when outliers are present. Furthermore, we examine if the outcome of the weak instrument test, first stage F-statistic, can be influenced by outliers. We show that this is indeed the case and show that instruments can therefore be falsely qualified as strong.

Abbreviations

2SGM Two stage generalised M estimator

2SLS Two stage least squares estimator

$\widehat{AV}$ Asymptotic variance covariance matrix

BDP Breakdown point

Dissimilarity 1990 dissimilarity index

Gini Inequality index

IF Influence function

IV Instrumental variables estimator

L_1**-RIV** RIV for regressions where dummy variables are included

LAD Least absolute deviation estimator

Lenper track length per square kilometer

MCD Minimum covariance determinant

MedSE Monte Carlo squared medain errors

OIV Ordinary instrumental variables estimator

OLS Ordinary Least Squares

RDI Railroad division indexes

RIV Robust instrumental variables estimator of Freue et al. (2013)

SD Stahel-Donoho

Tau multivariate τ-estimator for simultaneous equations

WIV Weighted instrumental variables estimator

WLS Weighted least squares

Contents

1 Introduction

The instrumental variables estimator (IV) is used for linear regression models when one or more continuous explanatory variables are endogenous, correlated with the error term. Here the IV depends on a strong instrument, that is, an instrumental variable which is correlated with the endogenous variable and uncorrelated with the error term. This is necessary to obtain consistent estimates. However, the data could be contaminated, that is, the data contains points that are not following the distribution of the bulk of the data (outliers). When outliers occur in the data, that can be in the response variable, the endogenous continuous variable, the exogenous continuous variable or the instrumental variable, the IV estimates are affected. Both the coefficient and the significance of the estimated parameters can become incorrect. Therefore a robustification of the IV could be an useful solution. These robust IV's diminish the influence of outliers on the estimation of the data.

This paper will use a robust IV introduced by Freue et al. (2013) called RIV and compare this estimator with other IV's, both robust and non robust, to test if RIV is able to perform better than other IV estimators when outliers are present in the data. Therefore our research question is: *Does RIV outperform other instrumental variables estimators when contaminated data is present.*

The other IV's used to compare the performance of RIV are, one non robust IV, the ordinary instrumental variables estimator (OIV), and one robust IV, 2SGM proposed by Wagenvoort & Waldmann (2002). RIV is also compared with other different computing estimators, S-estimator (the standard computing estimator), Stahel-Donoho (SD) estimator (Stahel (1981) and Donoho (1982)) and the minimum covariance determinant (MCD) estimator by Leroy & Rousseeuw (1987). Furthermore, we compare the performance of the L_1-RIV, which is an extension from Freue et al. (2013) on RIV for models containing dummy variables, among its computing estimators (same as RIV).

A simulation study is done in the same way as Freue et al. (2013) to compare the performance of the IV's for two different kinds of contamination, symmetric and asymmetric.

As addressed above, IV's rely on a strong instrument, finding such a strong instrument is therefore of great importance. When the assumptions of a strong instrument do not (partly) hold, the instrument is called 'weak'.

Outliers could be the reason that the instrumental variable and the endogenous variable are correlated or that the instrumental variable and the error term are not correlated. So, when these outliers are removed or downweighted, it could be that the instrument is not longer strong. Therefore, we test if the weak instrument test, first stage F-statistic, is influenced when outliers are downweigthed. The paper of Ananat (2011) is used to show if the first stage F-statistic could indeed be influenced by outliers. Where RIV is used as robust IV and the IV from Ananat (2011) is used as non robust IV to show if this influence exists.

This paper could be of interest for one who is interested in modelling contaminated data, where we give insight in which robust IV performs best under different kinds and levels of contamination. Furthermore, this paper is from interest for one who wants to become aware of the influence that outliers can have on the weak instrument test, first stage F-statistic.

In this paper, we will start with a literature study on the subject of (robust) instrumental variables estimators and the importance of strong instruments. Then in Section 3 and 4, the methodology of OIV, RIV and L_1-RIV is given together with the benefits and key points of RIV. This is followed by the simulation study in Section 5, where a contaminated environment is created, which we use to test the performance of the IV's. Hereafter, Section 6 presents the influence of outliers on instrument testing. Finally, the conclusion of this paper is given.

2 Literature

When endogenous variables are used in a model, the Ordinary Least Squares estimator(OLS) becomes biased and inconsistent. Kiviet & Niemczyk (2007) show that when variables are jointly dependent, because of correlation with the error term (endogenous variables) OLS is inconsistent. They provide evidence that IV is indeed consistent when strong instruments are available. Kiviet & Niemczyk (2007) point out that when there are no strong instruments available, and therefore only weak instruments, IV preforms worse than the inconsistent OLS.

Bound et al. (1995) show that when instrumental variables do not explain enough of the variation in the endogenous variable that the instrument is weak as well.

A problem addressed by Ontiveros et al. (2012) is that the conclusions drawn out the weak instrument testing is overoptimistic when good leverage points are present. Ontiveros et al. (2012) show that when good leverage points are present the first stage F-statistic increases strongly. This shows that good leverage points influence the test of weak instruments. Good leverage points are unusual points that do follow the pattern of the bulk of the points (Blatná (2006)). In this paper we look at bad leverage points, points that do not follow the bulk of the points.

To build further on the influence of outliers, Dehon et al. (2015) point out that non-robust IV's do not perform well when outliers are present in the data. Dehon et al. (2015) show that the non-robust IV is influenced strongly by outliers in the data and that robust IV are indeed able to remove (part of) this influence.

Over time different suggestions are made on how to make IV's robust against outliers in the data. Krasker & Welsch (1985) use a weighted instrumental variables estimator (WIV), where WIV assigns weights to observation. When observations get further away from the majority of the observations, the weights assigned to that specific observation gets smaller. This way Krasker & Welsch (1985) were able to lower the influence of the outliers. Their robust IV focuses on staying as efficient as possible.

One could also see OIV as a two stage least squares estimator (2SLS), where each stage of 2SLS could be robustified. Wagenvoort & Waldmann (2002) proposed such method, where they suggest to use a robust estimator in contrast to the least squares used by 2SLS. Wagenvoort & Waldmann (2002) use a generalised M-estimator as robust estimator (2SGM).

The robustification of the IV could also be done using simultaneous equation models. Maronna & Yohai (1997) propose Tau, an estimator that estimates simultaneous equation models using a robust estimator, the multivariate τ-estimator.

Lastly, some methods change the standard quadratic loss function into a robust loss function. Amemiya (1982) takes the least absolute deviation estimator (LAD) instead of the quadratic loss function.

3 Instrumental variables estimator

$$Y = X\beta + \varepsilon \tag{1}$$

OLS can be used to estimate a linear regression as formulated in Formula 1, with Y and ε vectors of $n \times 1$, $X^{n \times p+1} = (\iota, x_1, ..., x_{p-1})$ where ι is a column vector of ones of length n and β a vector of $(p+1) \times 1$. Although, some condition should hold; the errors are independently normal distributed with mean equal to zero and constant variance, no independent variables are linear combinations of one and other and independent variables are not correlated with the error term (exogenous). When the last assumption does not hold, Formula 2, the OLS will become biased and inconsistent.

$$Cov(x_{ij}, \varepsilon_i) \neq 0 \quad \text{for at least one } x_{ij} \text{ in } x_i, \quad \text{where } i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, n \tag{2}$$

IV can be used to avoid inconsistent estimates when there are endogenous variables (Kiviet & Niemczyk (2007)). IV uses replacements for the endogenous continuous variable. These replacements have to be correlated with the endogenous continuous variable and should be uncorrelated with the error term, therefore Formula 3a, the exogenous condition, and Formula 3b and 3c, the rank conditions, should hold for $z_i = (z_{i1}, ..., z_{iq})'$, with $q \geq p$. Such a replacement is called an instrumental variable. When such an instrumental variable is available consistent estimates can be achieved.

$$\text{(a) } E(z_i\varepsilon_i) = 0, \qquad \text{(b) rank } E(z_i z_i') = q, \qquad \text{(c) rank } E(z_i x_i') = p. \tag{3}$$

A widely used instrumental variables estimator is the ordinary instrumental variables estimator (OIV), which is estimated by two-stage least squares. The first stage regresses every column of the endogenous variable, $X^{n \times p+1}$, on the instrumental variable, $Z^{n \times q}$,

$$X = Z\alpha + \tau, \tag{4}$$

with α a vector of q by 1, giving the relation between the instrumental variable(s) and the endogenous variable(s). The error term, τ, is a vector of $n \times 1$ When performing OLS, $\hat{X}$ is as follows:

$$\hat{X} = Z\hat{\alpha} = Z(Z'Z)^{-1}Z'X = P_Z X, \quad \text{where } P_Z = Z(Z'Z)^{-1}Z'. \tag{5}$$

This $\hat{X}$ matrix is now uncorrelated with the error term, ε, when the exogenous condition holds for Z, Formula 3a. The second stage takes the estimated $\hat{X}$ and regresses every column Y on $\hat{X}$:

$$Y = \hat{X}\beta + \varepsilon, \tag{6}$$

which gives the OIV parameter:

$$\hat{\beta}_{OIV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = (X'P_Z X)^{-1}X'P_Z y. \tag{7}$$

The received $\hat{\beta}_{OIV}$ is consistent. A pitfall for OIV is when outliers occur in the data, in Y, X or Z, the estimation of OIV can be distorted. This could lead to wrong coefficients and wrong statistical significance. This influence of outliers can be removed when IV's are robustified. The RIV proposed by Freue et al. (2013) is such a robustification of OIV.

4 RIV

4.1 The RIV estimator

RIV proposed by Freue et al. (2013) uses a robustification of the centred values used in the estimation of OIV. Therefore we rewrite β_{OIV} in Formula 7, which represents all the parameters including the intercept. By separating the intercept (α) and remaining parameters (β) the relation becomes clear between the α and β. It can be written as follows:

$$\hat{\alpha}_{OIV} = \bar{y} - \bar{x}'\hat{\beta}_{OIV}, \tag{8}$$

$$\hat{\beta}_{OIV} = (X'P_Z X)^{-1} X' P_Z y. \tag{9}$$

Here X changes to an $n \times p$ matrix and the i^{th} row of X, Z and Y are $(x_i - \bar{x}), (z_i - \bar{z}), (y_i - \bar{y})$ respectively, for $i = 1, 2, \ldots, n$. It can be seen that centred variables are used for the calculation of α and β. These centred variables (mean, μ and variance, Σ) can be written as follows:

$$\mu = (\mu_x', \mu_z', \mu_y)' \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xz} & \Sigma_{xy} \\ \Sigma_{zx} & \Sigma_{zz} & \Sigma_{zy} \\ \Sigma_{yx} & \Sigma_{yz} & \Sigma_{yy} \end{pmatrix}, \tag{10}$$

where the variables are categorised for there relation to x, z and y.

These centred variables can be replaced with estimators for the mean and variance covariance matrix. Here is where the methods of estimation between OIV and RIV differ. OIV uses the sample mean ($\hat{\mu}$) and sample variance covariance matrix ($\hat{\Sigma}$), as shown in Formula 11 and 12.

$$\hat{\alpha}_{OIV} = \hat{\mu}_y - \hat{\mu}_x'\hat{\beta}_{OIV} \tag{11}$$

$$\hat{\beta}_{OIV} = [\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx}]^{-1}[\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zy}] \tag{12}$$

However when using $\hat{\mu}$ and $\hat{\Sigma}$ the estimates are highly sensitive to outliers. To remove this sensitivity other estimators can be used for the mean and variance covariance matrix. These estimators should have bounded influence functions, which makes the estimator B-robust. Furthermore the estimator should have a breakdown point, which implies that if contamination makes the instrumental variable invalid the IV breaks down. More information over the influence function and breakdown point in Section 4.2.

RIV uses a robust multivariate location (m) and scatter S-estimator (S) as estimators for the mean and variance covariance matrix (Lopuhaa (1989)). This estimator has a bounded influence function and has an high breakdown point, if mild conditions hold. Formula 13 and 14 give the formulation of the estimates, $\hat{\alpha}_{RIV}$ and $\hat{\beta}_{RIV}$, with m and S are categorised as in Formula 10.

$$\hat{\alpha}_{RIV} = m_y - m_x' \hat{\beta}_{OIV} \tag{13}$$

$$\hat{\beta}_{RIV} = [S_{xz} S_{zz}^{-1} S_{zx}]^{-1} [S_{xz} S_{zz}^{-1} S_{zy}] \tag{14}$$

The S-estimator assigns weights to observations in the data, that is, in the response variable, the endogenous variable, the exogenous variable and the instrumental variable. Here the S-estimator down-weights observations which lie away from the bulk of the data. This way the outliers get down-weighted. The S-estimator assigns weights between 0 and 1, were outliers receive weights towards zero, or equal to zero. For a formal definition of the S-estimator one could read the paper of Lopuhaa (1989).

4.2 Benefits and key points of RIV

In this Section the benefits and key points of RIV proposed by Freue et al. (2013) are presented. Firstly, we point out that RIV with S-estimator as computing estimator is both consistent and equivariant under mild conditions. Furthermore, we show that the S-estimator is B-robust. And lastly we show that when condition 3b and 3c hold RIV can have a breakdown point (BDP) of 50%.

Davies et al. (1987) show that when some regularity conditions hold the multivariate S-estimator yields to be consistent. This would be needed to ensure RIV itself is consistent. Although, Croux et al. (2003) state when z and ε are independent RIV could stay consistent when less conditions hold or when the multivariate S-estimator itself is inconsistent. Supporting Theorem can be found in paper of Freue et al. (2013).

The Theorem from Freue et al. (2013) is an extended version of Theorem 1 of Croux et al. (2003). It states that $S_{z\varepsilon} = 0$, that is, the covariance between the instrumental variable, z, and the error term, ε, is zero, when the instrumental variable and the error term are independent of one each other, regardless of the dependency of the explanatory variable, x or the consistency of the S-estimator.

Lemma 1 shows the equivariance of RIV, which means that the result out of the function of transformed parameters is the same as the transformed result out of the function with non-transformed parameters. Which also implies that RIV is a valid formula.

Lemma 1. *[Lemma 1 of Freue et al. (2013)] Let two parameter estimate sets,* $(\hat{\alpha}_{RIV}, \hat{\beta}_{RIV})$ *and* $(\alpha^*_{RIV}, \beta^*_{RIV})$*, be based on the sample* $(x_i, z_i, y_i)_{i=1}^n$ *and* $(x_i^*, z_i^*, y_i^*)_{i=1}^n$ *respectively. If* $y_i^* = \sigma y_i + \tau x_i + \eta$*,* $x_i^* = Q' x_i$ *and* $z_i^* = P' z_i$*, then for every* $\tau, \sigma \in R$*,* $\quad \tau \in R^q$ *and for nonsingular matrices* $Q^{q \times q}$ *and* $P^{p \times p}$ *it holds that:*

$$\alpha^*_{RIV} = \sigma \hat{\alpha}_{RIV} + \eta \qquad \textit{and} \qquad \beta^*_{RIV} = (Q^{-1})(\sigma \hat{\beta}_{RIV} + \tau). \tag{15}$$

This results RIV is equivariance.

To show that the RIV is B-robust, we show that the influence function (IF) stated in Freue et al. (2013) is bounded and that when mild conditions hold, the asymptotic variance is normally distributed. When conditions 3b and 3c hold the IF is bounded with the IF for α given as follows:

$$IF(x, z, y; \alpha, H) = IF(x, z, y; m_y, H) - IF'(x, z, y; m_x, H) b(H) - m'_{x,H} IF(x, z, y; b, H), \tag{16}$$

and the IF for β:

$$\begin{aligned} IF(x, z, y; \beta, H) = [S_{xz,H} S^{-1}_{zz,H} S_{xz,H}]^{-1} \Big[& IF(x, z, y; S_{xz}, H) S^{-1}_{zz,H} [S_{zy,H} - S_{zx,H} b(H)] \\ - \quad & S_{xz,H} S^{-1}_{zz,H} IF(x, z, y; S_{zz}, H) S^{-1}_{zz,H} [S_{zy,H} - S_{zx,H} b(H)] \\ + \quad & S_{xz,H} S^{-1}_{zz,H} (IF(x, z, y; S_{zy}, H) - IF(x, z, y; S_{zx}, H) b(H)) \Big]. \end{aligned} \tag{17}$$

Here the functional $T(H) = (a(H), b(H))$ is given as $a(H) = m_{y,H} - m'_{x,H} b(H)$ and $b(H) = [S_{xz,H} S^{-1}_{zz,H} S_{xz,H}]^{-1} [S_{xz,H} S^{-1}_{zz,H} S_{zy,H}]$. Where m_H and S_H are functionals of the S-estimator with H the distribution of $(x_i, z_i, \varepsilon_i)$.

Davies et al. (1987) together with Lopuhaa (1989) also prove that the IF of the S-estimator is asymptotically normally distributed, when some regularity conditions hold. By using the IF defined in Formula 16 and 17 the asymptotic variance covariance matrix $(\widehat{AV})$ can be derived, which leads to:

$$\widehat{AV}(T, H) = \frac{1}{n} \sum_{i=1}^n [IF(x_i, z_i, y_i; T, H_n) IF'(x_i, z_i, y_i; T, H_n)], \tag{18}$$

with $T(H_n) = (\hat{\alpha}_{RIV}, \hat{\beta}_{RIV})$ and H_n the empirical joint distribution of $(x_i, z_i, y_i)_{i=1}^n$.

The BDP of an estimator is the robustness against contaminated data. When an estimator has a high BDP then the estimator allows for a large portion of contaminated data, which can be up to the maximum of 50%. When an estimator breaks down, it will give incorrect estimations.

The BDP of IV's are depending on the instrumental variable, when the instrumental variable becomes invalid due to the contamination the IV will break down. If under contamination the instrumental variable for RIV remain valid the S-estimator has a high BDP of the maximum, 50%. The RIV can break down if either the S_{zx} becomes singular, which means that the rank conditions 3b and 3c do not hold. The other way RIV can break down is if m_y, m_x or S_{zy} are unbounded. However if the rank conditions 3b and 3c m_y, m_x or S_{zy} stay bounded till the BDP of 50 % is reached. More information on BDP itself can be found in the paper of Donoho & Huber (1983).

4.3 L_1-RIV

When one would be interested in estimating a model where dummy variables are included as well, Formula 19 , the S-estimator is regularly not able to be computed (Stromberg (1993)), which makes RIV unfeasible. To account for this we use an algorithm presented by Freue et al. (2013) which estimates the coefficients of the dummy variables by using a M-estimator.

$$y_i = \alpha + c_i'\beta_1 + x_i'\beta_2 + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, n \tag{19}$$

Here c_i and x_i are the rows of dummy matrix C and continuous variable matrix X, respectively. Where the columns of C and X are linearly independent. The algorithm of Freue et al. (2013) is given as follows:

Step 1: The initial step is estimating the coefficients $\beta_2^{(0)}$ by performing RIV on the model without dummy variables.

Step 2: The partial residuals obtained from $\beta_2^{(0)}$ are then used together with the L_1-estimator to estimate the initial estimates for the dummy variables, $\beta_1^{(0)}$. The L_1-estimator arguments are as follows, $\hat{\beta}_1^{(0)} = L_1(C, y - X\hat{\beta}_2^{(0)})$, which regress $y - X\hat{\beta}_2^{(0)}$ on C.

Step 3: The partial residuals obtained from $\beta_1^{(0)}$ are then used to estimate $\beta_2^{(1)}$ using RIV. Formula 20 gives this estimation for the $\beta_2^{(t)}$. After the estimation of $\beta_2^{(1)}$, *Step 2* will be done for $t = 1, 2, \ldots, T$. Formula 21 gives the estimation of *Step 2* for $\beta_1^{(t)}$.

$$\hat{\beta}_2^{(t)} = g(X, Z, y - C\hat{\beta}_1^{(t-1)}) \tag{20}$$

$$\hat{\beta}_1^{(t)} = L_1(C, y - X\hat{\beta}_2^{(t)}), \quad \text{for } 1 \leq t \leq T \tag{21}$$

Step 4: Repeating Step 3 T times

5 Simulation study

5.1 Data generation from model with continuous variables

To test the performance of the RIV estimator, we estimate the parameters of simulated data. Firstly, data is simulated by use of a basic model (Formula 22) consisting of an intercept (α), two continuous variables (x_1 is exogenous and x_2 is endogenous), parameters β_1 and β_2 and an error term, ε.

$$y_i = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i, \quad \text{for } i = 1, 2, ..., n \tag{22}$$

The simulation will be performed similarly to that from Freue et al. (2013). We program all simulation and estimators in R Core Team (2013). The data will consist of 1000 samples generated by the model. Each sample consists of 5 variables, y, x_1, x_2, z and ε, of 250 elements. x_1, x_2, z and ε will be generated by using a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, where $\mu = (0, 0, 0, 0)$. Furthermore, we set the diagonal of Σ equal to one and non diagonal elements equal to zero, excluding the elements $\Sigma_{x_1\varepsilon}$, $\Sigma_{\varepsilon x_1}$, $\Sigma_{x_1 z}$ and $\Sigma_{z x_1}$. This is done to satisfy the rank conditions (Formula 3). The parameter values used for the simulation for the basic model are: $\alpha = 1$ and $\beta_1 = \beta_2 = 2$. y will then be generated by putting the generated variables, x_1, x_2 and ε together with the parameters, α, β_1 and β_2 in Formula 22.

5.2 Contamination

To test the robustness of IV estimators against outliers, a percentage of the data points are changed into contaminated points, which is done randomly. This percentage will change from 0 to 30 percent with increments of 5 percent.

We consider two types of contamination, symmetric and asymmetric contamination. Where the symmetric contamination is generated by use of a Cauchy random variable $\mathcal{C}(0, 1)$. For the asymmetric contamination, we consider four different means for a Normal random variable $\mathcal{N}(\kappa, 0.01)$, where $\kappa = 1, 3, 5, 10$. The contamination can be present in y, x_1, x_2 and z, which is done separately to show the impact of contamination in each of the variables individually.

5.3 Performance measurements

To capture the performance difference the Monte Carlo squared median errors (MedSE) is used. Here we prefer the lowest MedSE, where a lower MedSE implies the parameter estimate to be closer to the specified parameter values. For symmetric contamination the MedSE is used as follows:

$$MedSE \ = \ median_r||\hat{\theta}^{(r)} - \theta||^2, \tag{23}$$

with θ^r being the parameter space of the r^{th} sample. For asymmetric contamination the maximum is taken of the MedSE for each different mean, $1, 3, 5$ and 10. This can be written as follows:

$$Maximum\ MedSE \ = max_{\kappa \in 1,3,5,10}\ median_r||\hat{\theta}^{(r)} - \theta||^2 \tag{24}$$

The maximum MedSE therefore captures the MedSE wherefore the estimator performs the worst.

5.4 Performance of RIV's computing estimators

Firstly, the performance of the different RIV computing estimators, S-estimator, SD-estimator (Stahel (1981) and Donoho (1982)) and MCD-estimator (Leroy & Rousseeuw (1987)), is compared.

Figure 1 shows the MedSE of the three different computing estimators of RIV for symmetric contamination. It can be seen that S-estimator has the lowest MedSE for all levels of contamination in X_1 (endogenous variable) and Z. For Y and X_2 (exogenous variable) the S-estimator has the lowest MedSE for contamination under 25 and 30 percent respectively. When higher contamination occurs the MCD-estimator has a lower MedSE. The SD-estimator performs overall the worst.

Figure 2 gives the result of the maximum MedSE of the RIV computing estimators for asymmetric contamination. For asymmetric contamination the SD-estimator seems to be the overall worst. An interesting observation can be done on the results of X_2, where the S-estimator radically increases, while the other computational estimators do not increase heavily. This observation was from asymmetric contamination, with $\kappa = 3$. Appendix Figure 8 to 11 show all different asymmetric contaminations for the computing estimators.

An overall note is that the results imply that asymmetric contamination seems to have more impact on the performance of the RIV than symmetric contamination. This can be seen when looking at the scale of both Figure 1 and 2.

The higher impact of the asymmetric contamination on performance could be explained by the characteristics of asymmetry and symmetry. That is, asymmetric contamination has a different mean and will therefore be a cluster of points around a different mean. In contrast, symmetric contamination has the same mean, but an different distribution around this mean. Because these contaminated points will be around the mean the impact is less than that of the clusters around a different mean made by asymmetric contamination.

We can conclude that SD-estimator performs the worst, except from the strange behaviour of the S-estimator when high asymmetric contamination is present in X_2. The S- and MCD-estimator are closer together in the results of the simulated data. However, specially for contamination under the 25 percent, the S-estimator outperforms the MCD-estimator. Therefore we will use the S-estimator as computational estimator for RIV to compare RIV against other robust instrumental variables estimators.

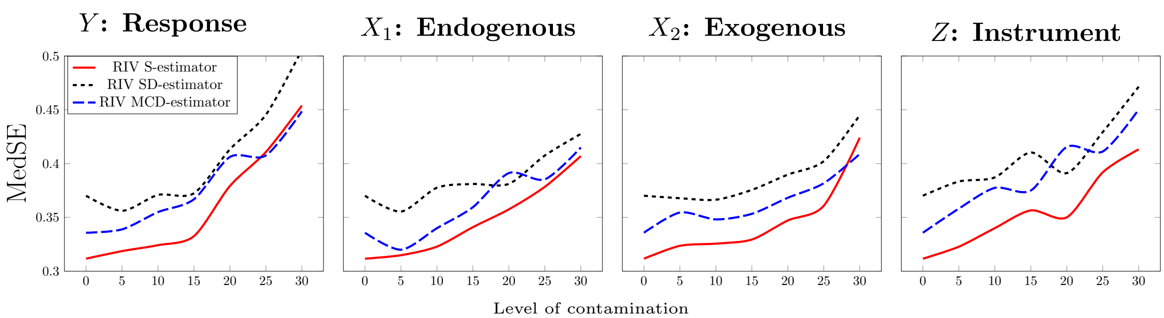

Figure 1: Symmetric contamination for RIV computational estimators.

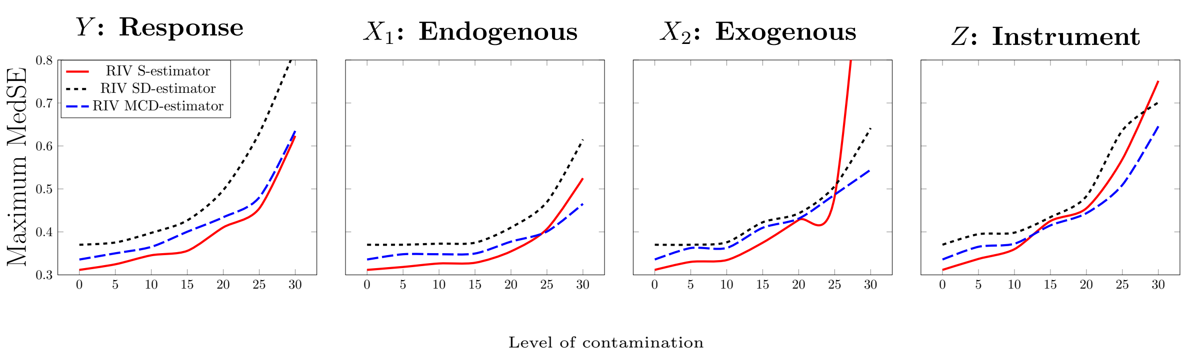

Figure 2: maximum asymmetric contamination for RIV computational estimators.

5.5 Comparing RIV with other instrumental variables estimators

Now, we test performance of the RIV estimator with the S-estimator against other IV's. The comparing IV's are the robust IV, 2SGM from Wagenvoort & Waldmann (2002), which is briefly explained in Section 2. And the non-robust IV, OIV, described in Sections 3 and 4.

Figure 3 shows that the RIV outperforms the other IV's when symmetric contamination is present. When no contamination is present, the OIV has lower MedSE for every variable. However, when symmetric contamination is included, the MedSE of OIV increases strongly. OIV it is the worst performing estimator for Y, X_1 and X_2. However for the instrumental variable Z, 2SGM performs worse for all levels of symmetric contamination except from 25%.

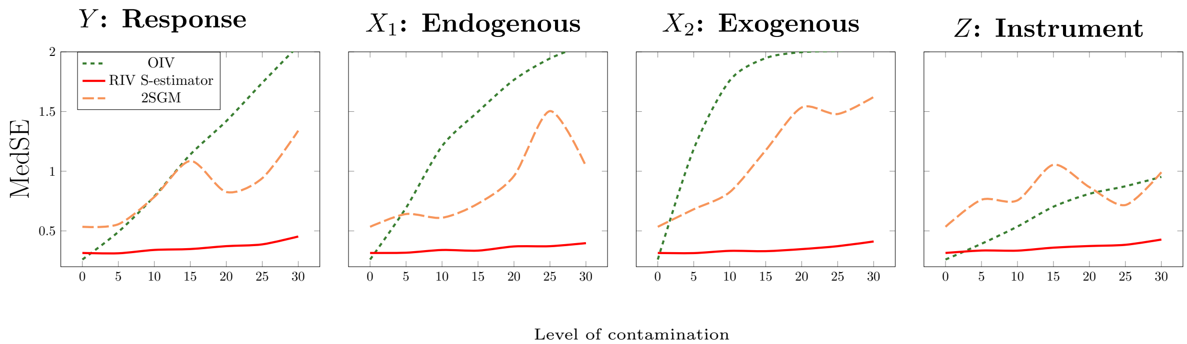

Figure 3: Symmetric contamination for RIV, 2SGM and OIV.

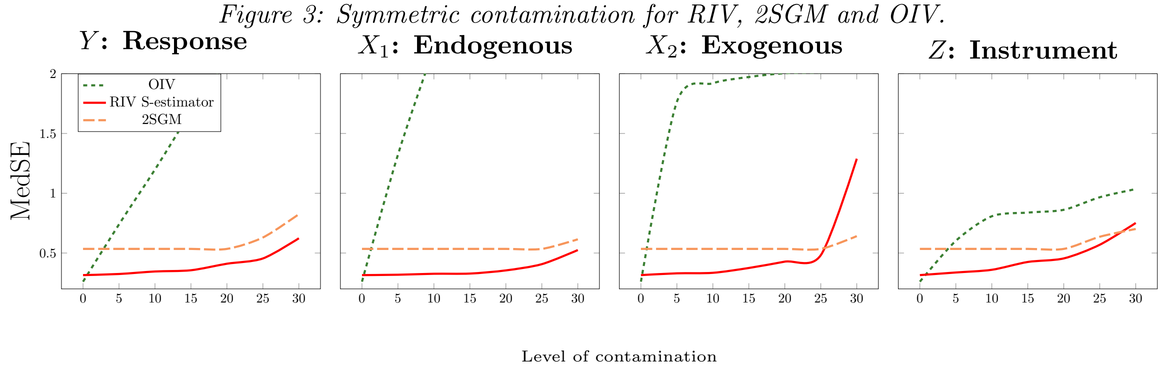

Figure 4: maximum asymmetric contamination for RIV, 2SGM and OIV.

When examining Figure 4, 2SGM performs closer to RIV when asymmetric contamination is present, in contrast with the performance difference by symmetric contamination. 2SGM even outperforms RIV for high asymmetric contamination (30%) in the exogenous variable, X_2, and the instrumental variable Z.

The spike of RIV for 30% of asymmetric contamination was already observed in Figure 2, where the computing estimators are compared. It seems that RIV computed with the S-estimator is the only IV that struggles to deal with this kind of contamination in the endogenous variable.

The results imply that OIV indeed performs poorly when contamination is present in data. However, when contamination is not present the OIV outperforms the other IV's. The overall results are in favour for RIV, which clearly performed better with symmetric contamination and performed better for almost every level of asymmetric contamination.

5.6 Data generation from model with continuous and dummy variables

We extend the basic model with continuous variables by including dummy variables in the model, which results in the model:

$$y_i = \alpha + \beta_{11}c_{1i} + \beta_{21}c_{2i} + \beta_{31}c_{3i} + \beta_{12}x_{1i} + \beta_{22}x_{2i} + \varepsilon_i, \quad \text{for } i = 1, \ldots, n \tag{25}$$

There are three dummy variables, c_{1i}, c_{2i} and c_{3i}, included in the model. All dummy variables are generated independently from a binomial distribution, with three different success probabilities, $0.5, 1/3$ and 0.25. The remaining variables, x_{1i}, x_{2i}, z_i and ε_i are generated in the same way as in the basic model with continuous variables. y will then be generated by putting the generated variables, c_1, c_2, c_3, x_1, x_2 and ε together with the parameters, $\alpha, \beta_{11}, \beta_{21}, \beta_{31}, \beta_{12}$ and β_{22} in Formula 25.

5.7 Performance of L_1-RIV computing estimators

The performance of the L_1-RIV computing estimators is compared for data generated in the previous Section, Section 5.6. This is done to show the performance of L_1-RIV when dummy variables present in a model.

Figure 5 shows the results of the computing estimators for symmetric contamination. The results of the simulation show that the difference in MedSE for the computing estimators are minimal. However, it can be seen that the S-estimator always has the lowest MedSE and that the SD-estimator has the highest MedSE.

When examining the results of the maximum MedSE for asymmetric contamination the same pattern occurs (Figure 6). Therefore, the S-estimator is the best performing computing estimator despite the minimal difference. The Appendix Figures 16 to 19 contain the result for all different means of asymmetric contamination. These Figures also include the OIV, which performs much worser than the RIV.

When comparing the impact of symmetric and asymmetric contamination, the MedSE is about two times as high for asymmetric contamination than for symmetric contamination, which is in line with the results of Section 5.4. Another observation is that the MedSE shows a high increase when small portions of symmetric contamination are present, whereafter the MedSE increases slowly. Asymmetric contamination shows a different pattern, where it follows a more linear trend.

An explanation for the pattern of the symmetric contamination could be the following. When a small percentage of symmetric contamination is present, it could be detected as asymmetric. This is because the outliers could be drawn such that they are all on one side of the mean of the main distribution. However, when a higher percentage of symmetric contamination is present it will become more likely that this will be symmetric around the mean. This could lead to the present pattern, because asymmetric has more impact as noted above. Therefore, it could be that there is a greater increase for smaller percentages of symmetric contamination. Note that this pattern is not occurring in our research for the model without dummy variables. Therefore the algorithm used for L_1-RIV could also be the reason for this pattern.

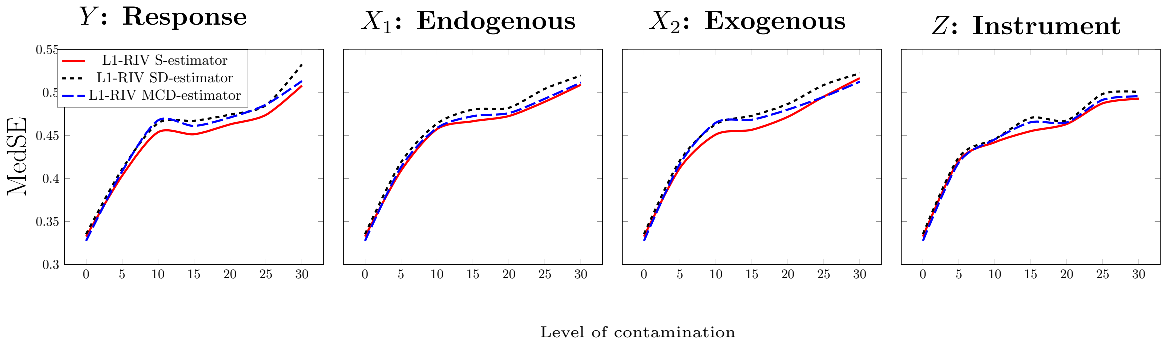

Figure 5: Symmetric contamination for L1-RIV computational estimators.

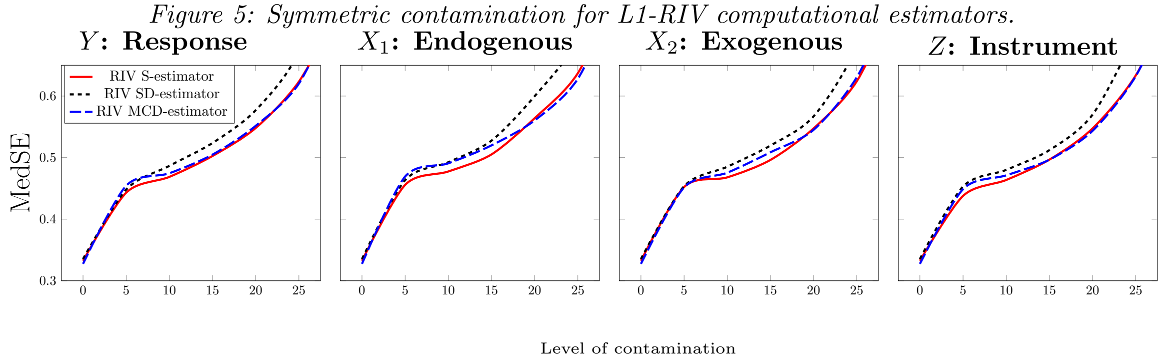

Figure 6: maximum asymmetric contamination for L1-RIV computational estimators.

6 Influence of outliers on weak instruments

6.1 First stage F-statistic

As explained in Section 1, strong instruments are from great importance for correct estimates of IV. In the same line that outliers can influence the estimation of the coefficients, outliers can also cause weak instrument tests to be incorrect. It can be the case that an instrument is only strong due to one or more outliers which increase the correlation between the instrumental variable and the endogenous variable or remove the dependency between the instrumental variable and the error term. In this Section the robustness against outliers of the weak instrument test, first stage F-statistic, is examined.

The first stage F-statistic tests if the coefficient of the instrumental variable is different from zero in the first stage of IV, An instrument is called weak if the first stage F-statistic is beneath 10, proposed by Staiger & Stock (1994).

The first stage F-statistic can be derived from the Wald test (Sanderson & Windmeijer (2016)).

Which has null hypothesis: $\alpha = 0$. The Wald test can be written as follows:

$$W_\alpha = \frac{\hat{\alpha} Z' Z \hat{\alpha}}{\hat{\sigma}_v^2}. \tag{26}$$

, where $\hat{\alpha}$ is the first stage OLS estimate (Section 3), Z the instrumental variable matrix of $n \times q$ and $\hat{\sigma}_v^2$ is equal to $(x'(I - Z(Z'Z)^{-1}Z')x)/n$, with x the explanatory variable $(n \times p)$ and $I^{n \times n}$ the diagonal matrix of ones. Under the null hypothesis, Wald becomes asymptotically equal to χ_q^2. F is equal to W_α/q, where q the number of instrumental variables.

To examine the first stage F-statistic we disprove some results of the paper of Ananat (2011), *"The Wrong Side(s) of the Tracks: The Causal Effects of Racial Segregation on Urban Poverty and Inequality"*. In this paper IV is used to discover which impact segregation has on metropolitan rates of black and white poverty and inequality in income. Together with the IV estimates they also provide the OLS estimates to show the influence that endogenous variables have on there estimates.

The paper assumes that there are no outlying observations in there data. Therefore, they do not consider a robust estimator. Conversely, we use a robust estimator, RIV with S-estimator, to estimate their regressions. This is done to show the difference in test results of the first stage F-statistic, when outliers are downweighted and when not. Furthermore, the impact of the RIV results on the use fullness of the coefficients and statistical significance is given. Firstly, the use of the data of Ananat (2011) is explained. Then the first stage regression of RIV is explained. Whereafter the results are given of both the paper and the RIV results.

6.2 Data

Their main regressions, presented in Table 2 of Ananat (2011), give insight on the effects of segregation on inequality and poverty for whites and blacks. Here the first two regressions, Formula 27 and 28, regress the 1990 dissimilarity index (Dissimilarity), which accounts for segregation, on the within-race poverty and inequality index (Gini) for whites and blacks separately. The third and fourth regression regress Dissimilarity on the poverty rate for whites and blacks separately (Formula 29 and 30). All four regressions are extended with an exogenous variable, track length per square kilometer (Lenper), when IV or in our case RIV is used. Lenper is used to account for area difference in the RDI observations. $(+\gamma \times Lenper)$

$$Gini_{white} = \alpha + \beta \times Dissimilarity \tag{27}$$

$$Gini_{black} = \alpha + \beta \times Dissimilarity \tag{28}$$

$$Poverty_{white} = \alpha + \beta \times Dissimilarity \tag{29}$$

$$Poverty_{black} = \alpha + \beta \times Dissimilarity \tag{30}$$

For the IV Ananat (2011) uses the instrumental variable railroad division indexes (RDI) for the endogenous variable, Dissimilarity. When running the first stage regression for there IV, the received results in the paper imply that RDI is a strong instrument for Dissimilarity. Therefore the instrumental variable is used in every IV regression which they perform with robust standard errors. They run their program with IV and OLS regressions in Statacprp (n.d.), both with robust standard errors.

Figure 7 presents the pair plots of the four main regressions together, to investigate if outliers are present in the used data. It can be seen that the extra variable for IV and RIV estimation, Lenper, has one big outlier. This outlier could have a big influence on the estimation. Furthermore, some clusters seem to be outside the bulk of the observations when looking at all pair plots made with Lenper.

When examining the pair plot of the endogenous variable, Dissimilarity, and the instrument variable, RDI, it shows that there are indeed some observations that are far from the majority of the observations. The same result can be seen for the other pair plots made with RDI and the response variables. The pair plots made with Dissimilarity and the response variable shows also some observations, which could be flagged as outliers.

Using Figure 7, we can conclude that the used data is indeed contaminated. The results from Section 5 show that when contamination is present in the data robust IV should be used. Therefore, RIV with the S-estimator will be used as robust IV for the main regressions of Ananat (2011).

Because the data is contaminated, the first stage F-statistic could have wrongly tested the instrument. Therefore, it could be that the instrumental variable, RDI, is not strong when the outliers are downweighted using RIV.

In Section 6.4 we present the findings of the first stage F-statistic and the effects of the results.

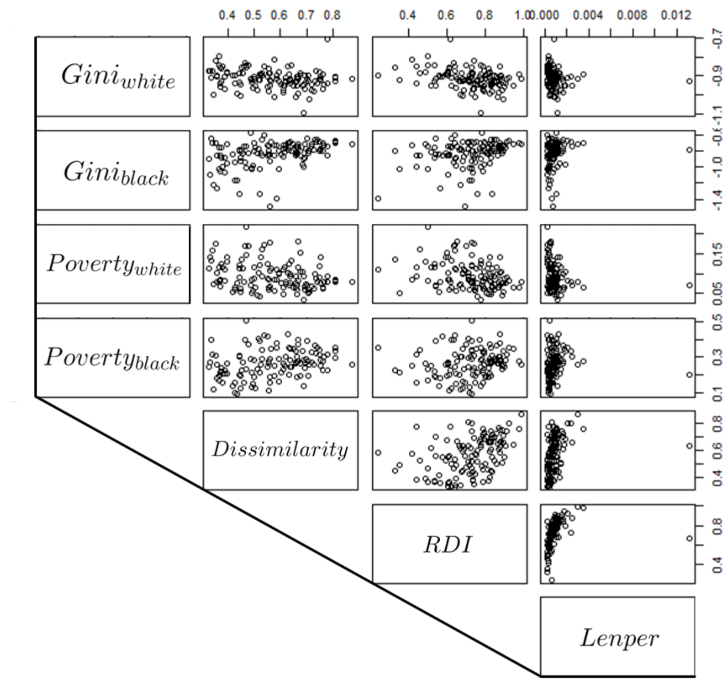

Figure 7: Pairs plot of for main regressions together.

6.3 First stage of RIV

IV's first stage regressions follow the regression as explained in Section 3. OIV uses OLS to estimate the first stage regression. RIV uses weighted least squares estimator (WLS), where the weights are calculated by the S-estimator.

As explained in Section 4, the S-estimator, which is the computing estimator of RIV, assigns weights to observations. The weights decrease when the observations are further away from the majority of the points, the bulk. Therefore, when outliers are present the S-estimator will down-weight them, which removes (part of) the influence of the outliers. So when testing the instrument on weakness with RIV, the influence of the outliers on the outcome of the test is removed.

A note to make is that due to the way RIV calculates the weights, the weights used in the first stage depend not only on the instrumental variable and endogenous variable, but also on the exogenous and response variables. Because of this, the estimates of the first stage as well as the first stage F-statistic differ for different exogenous and response variables included in the regression. Therefore, it could be such that one regression with certain exogenous and response variables could have a first stage F-statistic, which implies the instrument is weak. Where as a regression with other exogenous and response variables but the same instrumental variable and endogenous variable could have a first stage F-statistic, which implies a strong instrument is used.

Because of this characteristic of RIV, we estimate the first stage regression with WLS, where the weights are still calculated by RIV. However, the exogenous variable, Lenper, is now removed from the weights determination. Which means that the weights are only depending on the response variable, the endogenous variable and instrumental variable. This is done to show the impact that outliers in an exogenous variable can have on the first stage F-statistic. We refer to this estimation as, RIV*. Note that the exogenous variable is not removed from the first stage regression itself.

6.4 Comparing results

Ananat (2011) presents their first stage results without a first stage F-statistic, instead the t-test is used to determine the significance of the instrument in the first stage. When performing a first stage F-statistic on their IV regression, it results in an F-statistic of 16.57 (> 10), which implies RDI to be a strong instrument.

*Table 1: Main regression results. IV and RIV use RDI as instrumental variable and include an exogenous variable: Lenper. p-value: * >0.10, ** >0.05 & ***>0.01*

	Second stage coefficient of Dissimilarity			First stage coefficient of RDI		First stage F-statistic [1]		
Regression	OLS	IV	RIV	IV	RIV	IV	RIV	RIV*
Formula 27: $\text{Gini}_{\text{white}}$	-0.079**	-0.334***	-0.340*	0.357	0.368	16.57	5.72	46.75
Formula 28: $\text{Gini}_{\text{black}}$	0.459***	0.875**	-0.086	0.357	0.237	16.57	2.52	41.71
Formula 29: $\text{Poverty}_{\text{white}}$	-0.073***	-0.196***	-0.182	0.357	0.291	16.57	3.59	54.59
Formula 30: $\text{Poverty}_{\text{black}}$	0.182***	0.258**	-0.006	0.357	0.264	16.57	3.01	44.23

Table 1 gives the results of the main regressions together with the first stage results of the IV, RIV and RIV*. The first stage results show that when we use RIV to estimate the coefficients, the

[1]Note: the first statistic of RIV and RIV* are estimated in R Core Team (2013) with the R package of RIV (Cohen-Freue & Cubranic (2018)) using a weighted least squares regression with the weights of calculated in RIV regression. Whereafter the F-test is done on RDI=0. The RIV regression is the full regression for RIV and the reduced regression for RIV*. The WLS regression is the first stage regression, Dissimilarity = c + β_1 * RDI + β_2 * Lenper.

instrumental variable RDI is tested weak, for all for regressions the first stage F-statistic is below 10. This result supports the hypothesis that the first stage F-statistic is influenced by outliers in the data. A note should be made that the outliers in the response and exogenous variables could have influenced the outcome due to the way RIV calculates it weights. Therefore, it could be that RDI would be a strong instrument when only outliers in RDI and Dissimilarity are removed.

For RIV*, the influence of the exogenous variable, Lenper, on the weights is removed. This leads to first stage F-statistics above 10 and even above IV's first stage F-statistics. The impact of Lenper on the first stage F-statistic is strong. This shows that outliers outside the endogenous variable and instrumental variable, so in the exogenous variable and response variable, can indeed influence the outcome of the first stage F-statistic.

Thus, without downweighting the exogenous variable, the relationship between Dissimilarity and RDI is strong. However, in the regression proposed by Ananat (2011), Lenper is used in the main regressions when applying IV. Therefore, Lenper should be taken into account by the weighting of the RIV. One should investigate where these outliers come from and what they represent, especially the big outlier. An option could be removing the outlier(s). However, this could remove valuable information. Another option, could be changing Lenper.

However, for the main regressions of Ananat (2011) Lenper is included in the regression and results in a weak instrument when using RIV. This causes RIV's estimates to be possibly worse than the OLS estimates and possibly inconsistent as noted in Section 2 and further investigated by Kiviet & Niemczyk (2007). The problem that occurs, is that the estimates of OLS are biased and inconsistent due to an endogenous variable used in the regressions, and the IV estimates are inconsistent due to outliers in the data. On top of that RIV does not feature a strong instrument in the regressions either. This causes that none of the estimates can be interpreted correctly. In this case the used data should be investigated and reconsidered.

The results that are drawn from our research support the concern of the first stage F-statistic to be influenced by outlying observations in the data. Further work is needed to come up with a weak instrument test which is robust against outlying observations. One could think of a test which downweights observations before performing the first stage F-statistic or an other weak instrument test. However, using a robust IV will give already the information if the instrument is strong under contamination, because as shown above robust IV will remove the influence of the outliers.

7 Conclusion

Non-robust IV's become inconsistent when contamination is present in the data. Robust IV's are a solution for this problem. We focused our research on the robust IV, RIV of Freue et al. (2013).

In a simulated environment we have shown that RIV with the S-estimator outperforms the other computing estimators, SD- and MCD-estimator. Furthermore, RIV performs better than the robust IV, 2SGM proposed by Wagenvoort & Waldmann (2002), and outperforms the non-robust IV, OIV, when contamination was present in the simulation. Lastly, we compared the performance of the computing estimators of L_1-RIV, the RIV extension which allows RIV to estimate models containing dummy variables. The results were close between the three computing estimators. However, the S-estimator outperformed the SD- and MCD-estimator slightly.

We limited our research on some cases of symmetric and asymmetric contamination, therefore it could be from interest to investigate how other types or levels of contamination affect the estimates. Another limitation is that our research uses Monte Carlo median squared errors as performance measurement, other performance measurements could reveal other results. Further study could also be done towards more robust IV's, for instance WIV proposed by Krasker & Welsch (1985) or Tau proposed by Maronna & Yohai (1997).

After the simulation study, the testing of instruments for IV's in contaminated environments is criticised. This is done based on the paper of Ananat (2011). Ananat (2011) uses IV to estimate which impact segregation has on metropolitan rates of black and white poverty and inequality in income. Their IV uses a strong tested instrument, when running the first stage regression using IV. However, when downweighting outlying observations by use of RIV the first stage F-statistic implies that the instrument is weak. This supports that the instruments could be tested incorrectly strong due to outliers in the data. This result shows that one should be aware of the influences of outliers on the instruments. The outliers in all variables can influence these results.

To extend on these findings on weakness of the instrument, one could further investigate if there are differences between robust IV's and if a robust weak instrument test could be made. Furthermore, for the weak instrument tests, we only investigate the first stage F-statistic. Other weak instrument test, such as the Hausman test statistic (Hausman (1978)) could be investigated as well.

References

Amemiya, T. (1982). Two stage least absolute deviations estimators. *Econometrica: Journal of the Econometric Society*, 689–711.

Ananat, E. O. (2011). The wrong side (s) of the tracks: The causal effects of racial segregation on urban poverty and inequality. *American Economic Journal: Applied Economics*, *3*(2), 34–66.

Blatná, D. (2006). Outliers in regression. *Trutnov*, *30*, 1–6.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, *90*(430), 443–450.

Cohen-Freue, G., & Cubranic, D. (2018). Riv: Robust instrumental variables estimator package [Computer software manual]. (R package version 2.0-5)

Croux, C., Van Aelst, S., & Dehon, C. (2003). Bounded influence regression using high breakdown scatter matrices. *Annals of the Institute of Statistical Mathematics*, *55*(2), 265–285.

Davies, P. L., et al. (1987). Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, *15*(3), 1269–1292.

Dehon, C., Desbordes, R., & Verardi, V. (2015). The pitfalls of ignoring outliers in instrumental variables estimations: an application to the deep determinants of development. In *Empirical economic and financial research* (pp. 195–213). Springer.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators* (Tech. Rep.). Technical report, Harvard University, Boston. URL http://www-stat. stanford ....

Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. *A festschrift for Erich L. Lehmann*, *157184*.

Freue, G. V. C., Ortiz-Molina, H., & Zamar, R. H. (2013). A natural robustification of the ordinary instrumental variables estimator. *Biometrics*, *69*(3), 641–650.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, 1251–1271.

Kiviet, J. F., & Niemczyk, J. (2007). The asymptotic and finite sample distributions of ols and simple iv in simultaneous equations. *Computational statistics & data analysis*, *51*(7), 3296–3318.

Krasker, W. S., & Welsch, R. E. (1985). Resistant estimation for simultaneous-equations models using weighted instrumental variables. *Econometrica: Journal of the Econometric Society*, 1475–1488.

Leroy, A. M., & Rousseeuw, P. J. (1987). Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*.

Lopuhaa, H. P. (1989). On the relation between s-estimators and m-estimators of multivariate location and covariance. *The Annals of Statistics*, 1662–1683.

Maronna, R. A., & Yohai, V. J. (1997). Robust estimation in simultaneous equations models. *Journal of Statistical Planning and Inference*, *57*(2), 233–244.

Ontiveros, D. U., Verardi, V., et al. (2012). *Supposedly strong instruments and good leverage points* (Tech. Rep.).

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Sanderson, E., & Windmeijer, F. (2016). A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of Econometrics*, *190*(2), 212–221.

Stahel, W. A. (1981). *Breakdown of covariance estimators*. Section for Statistics, Swiss Federal Technical University.

Staiger, D., & Stock, J. H. (1994). *Instrumental variables regression with weak instruments* (Tech. Rep.). National Bureau of Economic Research.

Statacprp. (n.d.). Stata [Computer software manual]. Retrieved from `http://www.stata.com`

Stromberg, A. J. (1993). Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM Journal on Scientific Computing*, *14*(6), 1289–1299.

Wagenvoort, R., & Waldmann, R. (2002). On b-robust instrumental variable estimation of the linear model with panel data. *Journal of Econometrics*, *106*(2), 297–324.

Appendix

Supporting Figures for result of comparison RIV computing estimators.

Here the supporting Figures for the maximum asymmetric contamination Figure, Figure 2. The Figures are given for asymmetric contamination with κ equal to 1, 3, 5 and 10. Here the scales are independent for every Figure.

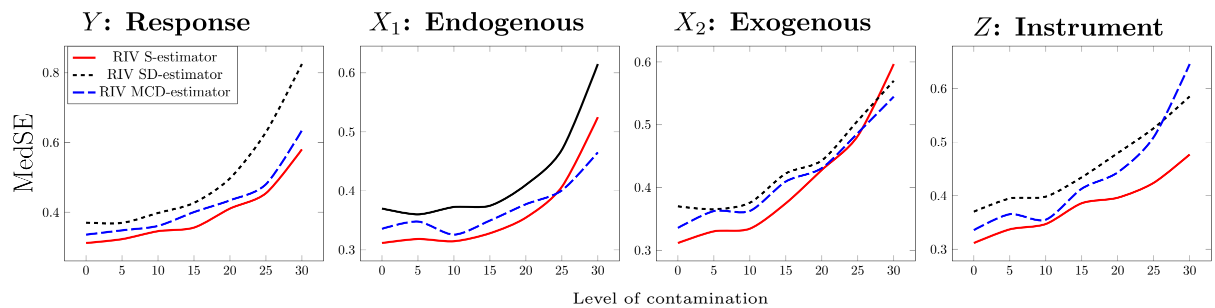

Figure 8: Asymmetric contamination, with $\kappa = 1$ for RIV computational estimators.

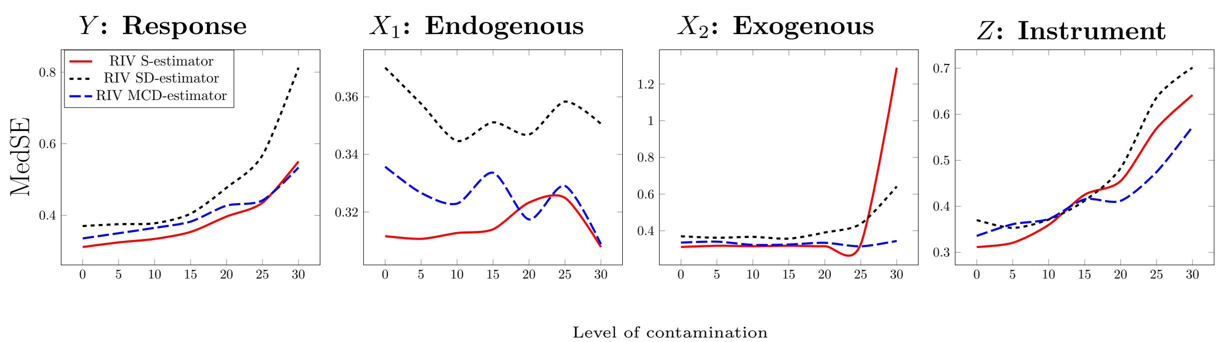

Figure 9: asymmetric contamination, with $\kappa = 3$ for RIV computational estimators.

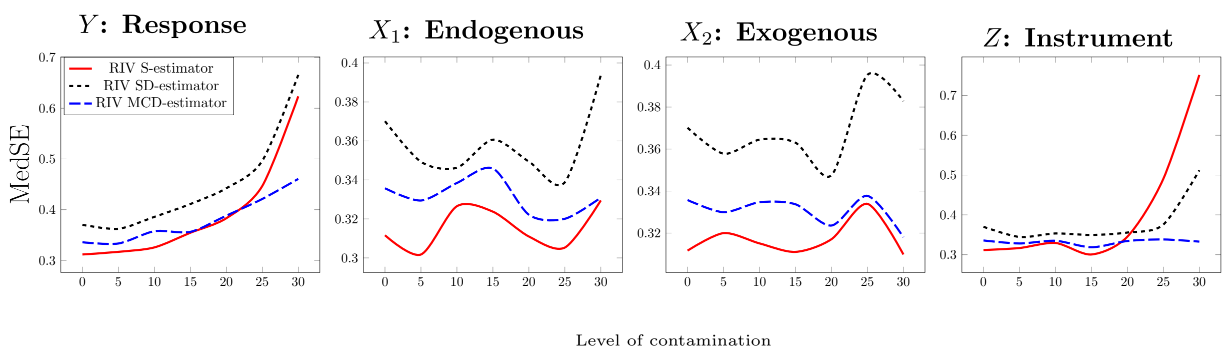

Figure 10: asymmetric contamination, with $\kappa = 5$ for RIV computational estimators.

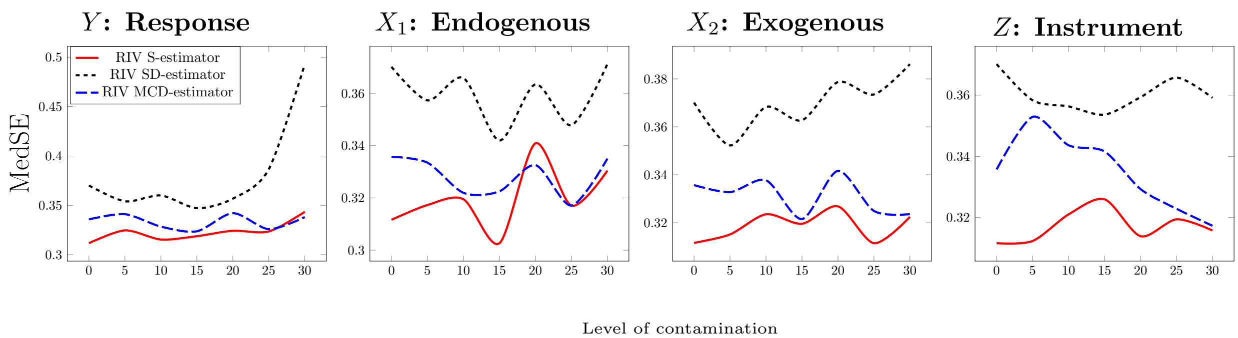

Figure 11: asymmetric contamination, with $\kappa = 10$ for RIV computational estimators.

Supporting Figures for result of comparison RIV against other IV's.

Here the supporting Figures for the maximum asymmetric contamination Figure, Figure 4. The Figures are given for asymmetric contamination with κ equal to 1, 3, 5 and 10. Here the scales are independent for every Figure.

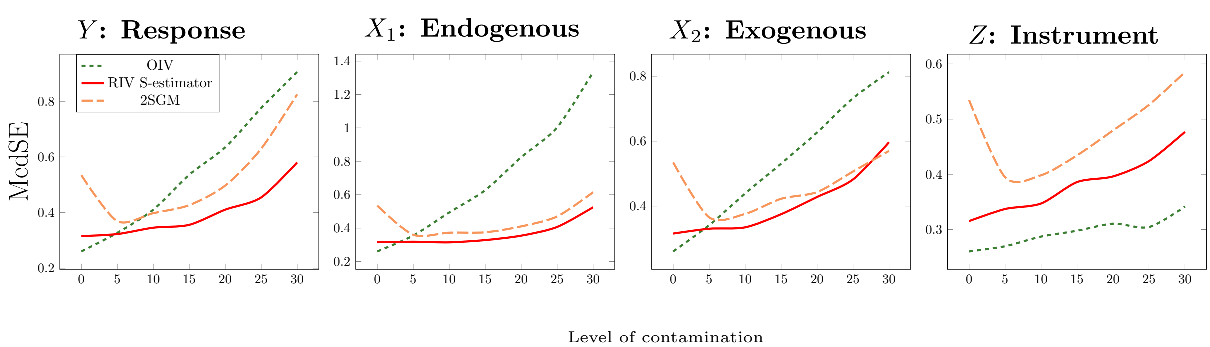

Figure 12: Asymmetric contamination, with $\kappa = 1$ for RIV, 2SGM and OIV.

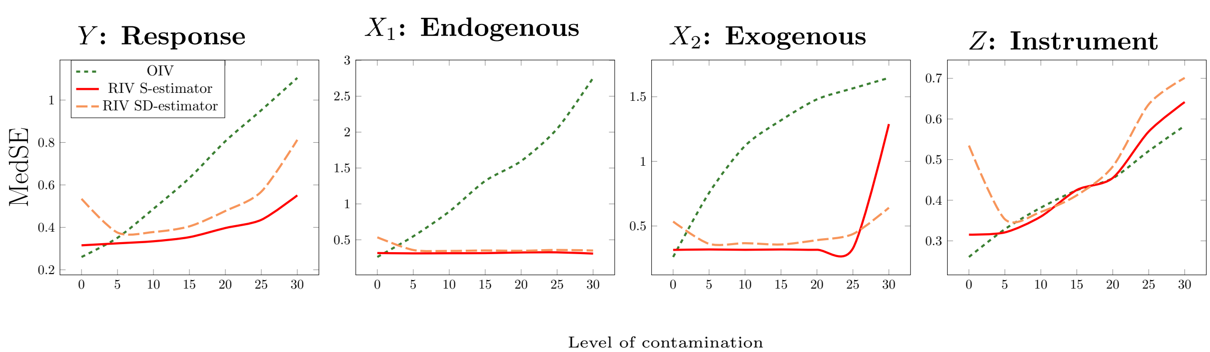

Figure 13: asymmetric contamination, with $\kappa = 3$ for RIV, 2SGM and OIV.

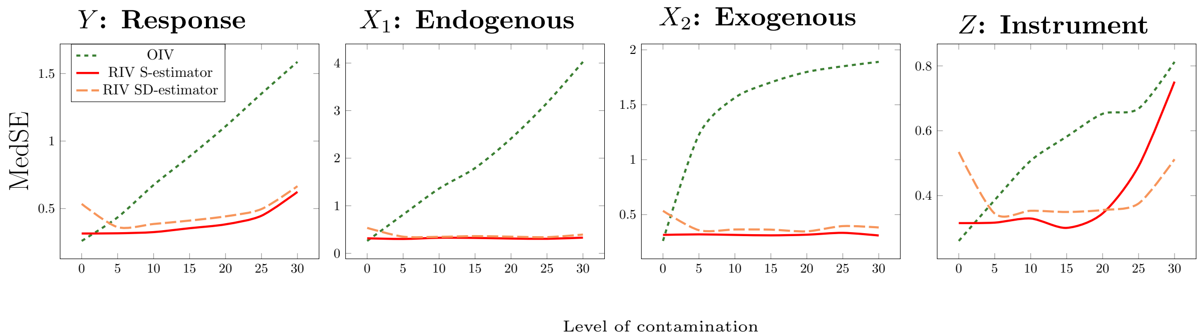

Figure 14: asymmetric contamination, with $\kappa = 5$ for RIV, 2SGM and OIV.

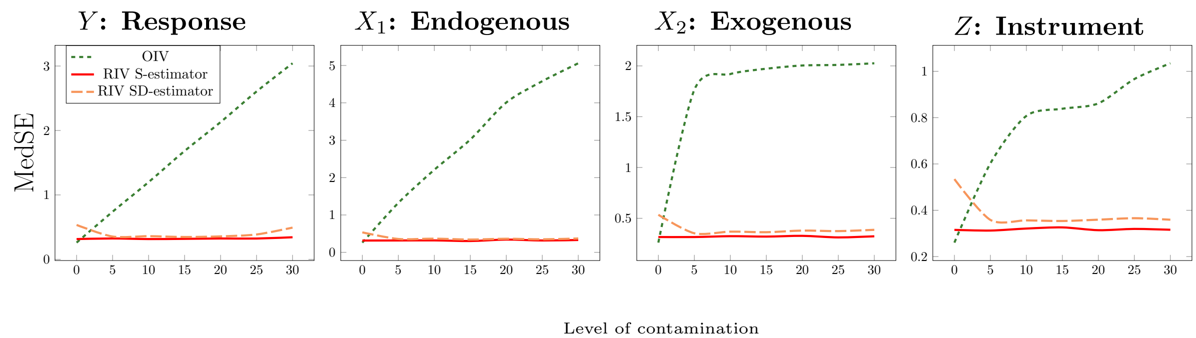

Figure 15: asymmetric contamination, with $\kappa = 10$ for RIV, 2SGM and OIV.

Supporting Figures for result of comparison L_1-RIV computing estimators.

Here the supporting Figures for the maximum asymmetric contamination Figure, Figure 6. The Figures are given for asymmetric contamination with κ equal to 1, 3, 5 and 10. Furthermore, the Figures below contain the OIV results to show as comparison. Here the scales are independent for every Figure.

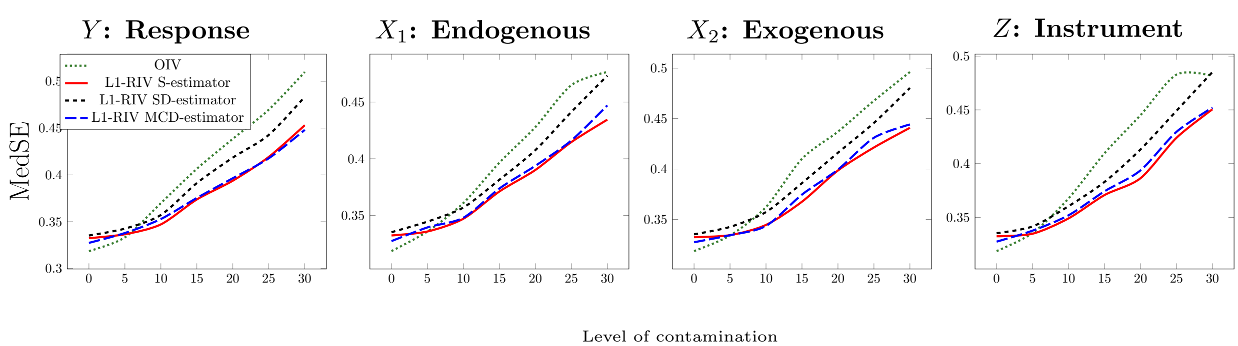

Figure 16: Asymmetric contamination, with $\kappa = 1$ for L1-RIV computational estimators.

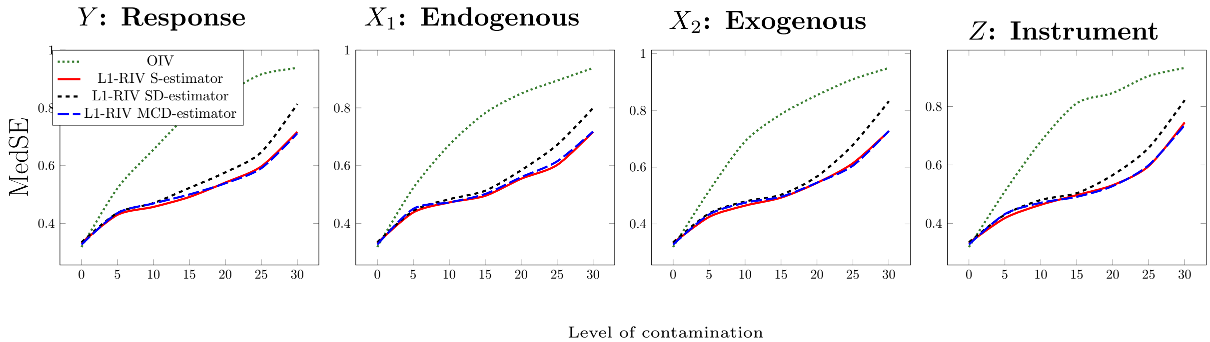

Figure 17: asymmetric contamination, with $\kappa = 3$ for L1-RIV computational estimators.

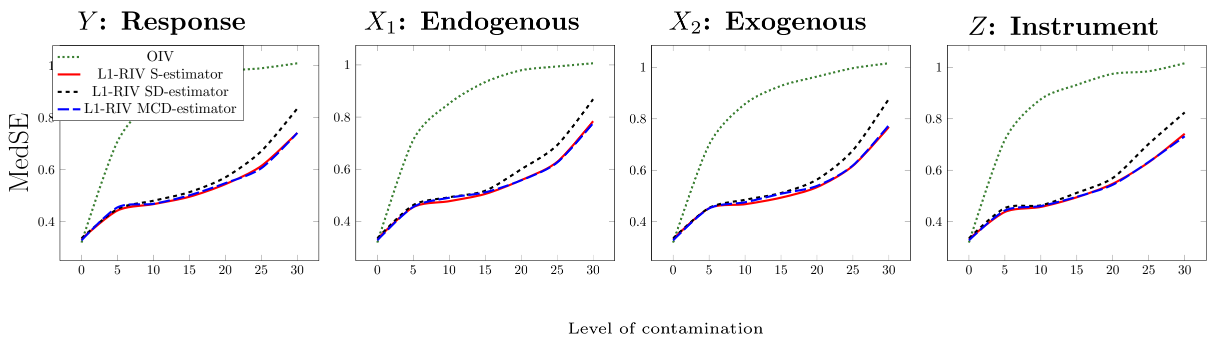

Figure 18: asymmetric contamination, with $\kappa = 5$ for L1-RIV computational estimators.

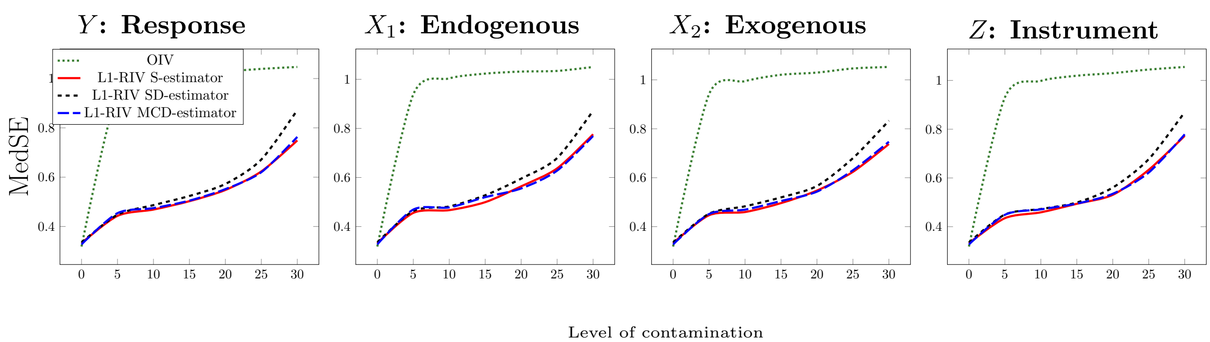

Figure 19: asymmetric contamination, with $\kappa = 10$ for L1-RIV computational estimators.