Erasmus University Rotterdam

Erasmus School of Economics

Department of Econometrics and Operations Research

On the Robustification of the Ordinary Instrumental Estimator's Closed Formula

Bachelor Thesis

5th of July 2020

Supervisor	*Author*
PhD. J. Klooster	Jacky Chu
Second assessor	*Student ID*
dr. W. Wang	484312

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

This paper focuses on the robustification of the Ordinary Instrumental Variables Estimator (OIV) in the form of the Robust Instrumental Variable Estimator (RIV), proposed by Cohen Freue, Ortiz-Molina, and Zamar (2013). The RIV estimator robustifies the OIV's closed formula, using the multivariate robust location and scatter S-estimator. We show that RIV exhibits several attractive properties, such as equivariance, B-robust influence function and a breakdown point of 50% amongst others. Moreover, we conduct an extensive simulation study to evaluate the performance of RIV under different types and level of contamination. Overall, we find that RIV outperforms the OIV estimator for every type and level of contamination. Finally, we showcase the ability of RIV to flag outliers and uncover true relationships on a real data-set regarding earthquakes.

Contents

1 Introduction

> *"A discordant small minority should never be able to override the evidence of the majority of the observations."*

These are the words of Peter J. Huber, well-known for his contribution to the heteroskedastic-consistent Huber-White standard errors in regression-analysis (Huber, 1978). It captures the intuitive idea of what robustness entails.

In statistics, assumptions about the underlying situation are of crucial importance. These assumptions could be about the distribution, independence or properties of parameters, and so on. They are vital in the sense that the assumptions made prior to the analysis of a problem provide a framework in which hypotheses can be tested. Although these assumptions are vital, they are not expected to be certainly true. Essentially, they are mathematical convenient rationalizations of often not so certain beliefs. In addition, literature often suggests that small errors in the mathematical model do not give rise to significant deviations in the (final) conclusion. This is unfortunately not always true in practice. The aforementioned has prompted researchers to develop robust methods and estimators that are not influenced as much by outliers (Ricardo, Maronna, Douglas, Yohai, & Salibián-Barrera, 2019).

When the assumption of exogeneity is violated in the linear regression setting, i.e. the independent variables are uncorrelated with the error terms, the OLS estimator renders inconsistent and biased estimates. A common solution in literature is to use Ordinary Instrumental Variable (OIV) estimators. In essence, OIV estimators use the variability of the endogenous variables that is uncorrelated with the error term to form new independent variables. These variables then render estimates that are consistent when endogeneity is present. However, empirical data sets often contain outliers that can potentially influence the precision of the mathematical model and in turn disturb the accuracy of the OIV estimates (Young, 2020). Therefore, in this paper we investigate the robustification of the ordinary instrumental variable and its robustness measures.

In particular, we follow the framework provided by Cohen Freue et al. (2013) where they propose the Robust Instrumental Variable (RIV) based on the robustification of the OIV closed-formula. Here, we robustify to the estimating equations using multivariate location and scatter S-estimators. The reason why we choose the S-estimator over other robust estimators, is due to its desirable properties under regularity conditions. That is, they are consistent and asymptotically normal, affine equivariant, positive definite, possess a bounded influence function and can achieve a maximal breakdown point of 50% Cohen Freue et al. (2013). These concepts will be revisited and proved in the methodology section.

Besides investigating the robustness measures of the RIV, we also evaluate its practical performance in a simulation study conducted in R (R Core Team, 2017), where the figures and statistical analysis were produced using the riv (Cohen Freue, Ortiz-Molina, & Zamar, 2018) package. Here, we examine how the RIV reacts and performs under different degrees and types of contamination of the data. Moreover, we review the performance of the RIV estimator in practice by comparing the results between OIV and RIV using data provided in Fuller (1987).

We find that our results are consistent with previous research of Cohen Freue et al. (2013). We prove that the RIV estimator follows the above-mentioned desirable properties, that is: (i) equivariant and consistent under weak regularity conditions (ii) the influence function is B-robust; (iii) Breakdown point (BP) is asymptotically 50%; (iv) RIV can be rewritten as a weighted instrumental variables estimator. Moreover, we find that practically for every type and degree of contamination the RIV-estimator outperforms the OIV-estimator. Additionaly, in our extension, we find that riv is an effective method to flag outliers and uncover the true relationship between the variables.

The remainder of the paper is structured as follows. Section 2 provides an extension literature study. In Section 3, the robustness measures are explained and examined and the methodology for the estimation model containing dummy covariates is elaborated. The simulation study is conducted in section 4. Section 5 contains the results, for which a conclusion and discussion is provided in section 6.

2 Literature

This section starts by covering the theoretical generalities that are useful when considering the robustness of a model, these generalities are used throughout the paper. Following is a review of the current literature on robustness and robust instrumental variables.

2.1 Generalities

Huber (Huber, 1964) and Hampel (F. R. Hampel, 1974) have introduced a set of generalities that are useful when considering robust statistics. These general principles are useful in the sense that they allow for a convenient estimation of parameters based on the majority of the data even when the data set contain outliers.

2.1.1 The Functional Approach

The functional T denotes the function that uses either the probability distribution or the empirical distribution as an argument. A parameter can be interpreted as the *functional* of the probability distribution, whereas an estimator is defined by the same functional of the empirical distribution. As an example, consider the mean with the model F_θ parametrized by θ. The functional of the mean is then given by $T(F_\theta) = \int xdF_\theta(x)$. By the Law of Large Numbers, we know that for a sequence of iid random variables $X_1, \ldots, X_n$, we have $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}_\theta[X] := \int xdF_\theta(x)$, as $n \to \infty$. Therefore, asymptotically, we can replace the estimator of the mean by the functional. We can generalize this to practically all other estimators (Huber & Ronchetti, 2009). For this reason, we prefer to use the mathematically more elegant functionals. Functionals allow for a convenient method to evaluate models other than F_θ. Moreover, it allows us to examine the behaviour of the estimator in the neighborhood of the model F_θ, this will be elaborated in the next generality. One notion should be made before moving on to the next generality, that is the notion of Fisher Consistency Functionals. In the field of Robust Statistics, we mainly consider Fisher Consistency Functionals. A functional is said to be Fisher Consistent when $T(F_\theta) = \theta$, the necessity behind it comes from the uselessness when the functional is considered *not* to be Fisher Consistent, i.e. the functional will not give us an reliable estimate based on the distribution.

2.1.2 Studying the Neighborhood

The benefit of using the functional, is that it allows for an elegant way to study the neighborhood of a model F_θ. In the context of Robust Statistics, we are interested in deviations from the main model in the form of outliers. This can be mathematically translated as follows.

$$F_{\theta,\varepsilon} = (1-\varepsilon)F_\theta + \varepsilon G \tag{1}$$

Here, G denotes the alternative distribution that the outliers follow, ε denotes the contamination proportion of the original model and we let $\varepsilon \in [0, 0.5]$, to account for a natural upper limit. Equation 1 gives an elegant way to evaluate functionals when considering contaminated probability distributions. In literature, Equation 1 is often referred to as the *gross-error model* (Huber & Ronchetti, 2009).

2.1.3 Equivariance

Equivariant translation means that a translation of input features results in an equivalent translation of outputs. For robust estimators, and estimators in general, this is a desirable property. Informally, we state that an estimator is equivariant when the function is unchanged after a transformation of the input. Mathematically we say that a function $f(x)$ is said to be equivariant to a function $g(x)$ if $f(g(x)) = g(f(x))$.

2.2 Literature Overview

George Box published his early insights on non-normality and its issues when disregarded in 1953 (Box, 1953). Box is responsible for coining the term "robust"; he was the first to introduce this concept into statistical literature. Since then, it has sparked researchers to come up with advancements in this field. Compared to the 1960's where robust methods were considered to be "dirty" methods, robustness has become a scientific phenomenon that is now seen as a desirable property and almost unmissable property.

Although Box was the first to introduce the term, it was John Tukey who put down fundamental work (Tukey, 1960) where he recognized the extreme sensitivity of certain conventional statistical methods when small deviations occurred from the assumptions that were made.

The paper Tukey published led to more theoretical contributions in the field of robust statistics. Especially, the work of Huber (1964) and F. Hampel (1968) were impactful. They proposed new methods and focussed on optimality properties (Huber, 1978) by the minimax principle to make the so-called "dirty" methods more acceptable.

These contributions have led to the development of robust estimators. Ricardo et al. (2019) discuss several robust estimators in their book; M-estimators, S-estimators L-estimators, all providing their own advantages in application and robustification against outliers. The S-estimator, for example, has the advantage that under weak regularity conditions the estimator is consistent and asymtpotically normal, affine equivariant, positive definite and is B-robust (Rousseeuw & Leroy, 1987).

Robustness is also of importance in the branch of instrumental variables. Young (2020) shows in an extensive review of 31 papers published in the journals of the American Economic Association, that the potential presence of outliers in a simulation study adversely affects the power of the IV estimates. The study thus suggests that the reliability of these estimates is doubtful in the application of real world data - where outliers are imminently present.

Several methods and estimators have been developed the past decades to overcome this problem. Cohen Freue et al. (2013) introduce a robust instrumental variable (RIV) that is easy to compute and has a high breakdown point (asymptotically 50%), among other attractive robust properties. Their estimator is based on the robustification of the closed OIV's formula - using the aforementioned S-estimator, hence robustifying the multivariate location and scatter matrix. Our paper builds forth on this particular robust instrumental variable.

Another approach is taken by Krasker and Welsch (1985). Instead of robustifying the closed OIV's formula, they robustify the estimating equations using a weighted instrumental variable. Similar to Cohen Freue et al. (2013), the robustification can also be done in the two separate stages[1] of the OIV estimation. Wagenvoort and Waldmann (2002) has proposed two of those estimators in his paper: Two-stage generalized M (2SGM) and robust generalized method of moments (RGMM). Although useful in application under the presence of heteroskedasticity and autocorrelated error, it is shown by Krasker and Welsch (1985) that two-stage robust estimators are generally less efficient.

One common problem with the aforementioned RIV, is that it is often unfeasible to compute the RIV when the data contain dummy covariates. The problem lays in the fact that the subsampling procedure needed to compute the S-estimator yields collinear subsamples in the presence of dummy covariates. However, this subsampling algorithm is necessary due to the fact that the S-estimator cannot be computed exactly (Stromberg, 1993). An iterative algorithm that allows for both exogenous categorial and continuous predictors has been proposed that solves this exact issue Maronna and Yohai (2000), which expands the application fields of the RIV.

3 Methodology

The instrumental variable regression with some notation is described first. It is widely known that the IV estimator is sensitive to outliers and as a result often provides unreliable statistics. We review the robustification approach similar to the methods proposed by Cohen Freue et al. (2013). Additionaly, we describe the details of our simulation study to evaluate the performance of RIV as opposed to OIV. Finally, we evaluate the effect in practice of using RIV over OIV using data on earthquakes provided by Fuller (1987).

3.1 Instrumental variable estimator

Consider the multivariate regression model:

$$y_i = \alpha + x_i'\beta + \varepsilon_i \qquad \text{for } i = 1, \ldots, n \tag{2}$$

where β is a vector of p parameters, $x_i = (x_{i1}, ..., x_{ip})'$ is a vector of p explanatory variables and $\varepsilon_1, ..., \varepsilon_p$ represents the error terms, where $E(\varepsilon_i) = 0$. A covariate is considered endogenous when the following expression holds:

$$Cov(x_{ij}, \varepsilon_i) \neq 0 \quad \text{for at least one } x_{ij} \tag{3}$$

In practice, covariates are often found to be endogenous, as a result, OLS estimates are not consistent and conventional results no longer hold valid. This prompted researchers to come with the development of instrumental variables - additional variables that are sufficiently correlated with the endogenous variables but uncorrelated with the error term. Replacing the current explanatory variables set x_i with the instrumental variables, denoted as

[1] The stages refer to the two stages of the OIV estimation process. The first stage regresses x on the instrumental variable z, which gives a exogenous covariate. In the second stage, the exogenous covariates are plugged in the normal OLS function.

$z_i' = (z_{1i}, \dots, z_{qi})$, in a regression leads to estimates that are consistent. Technically, an instrumental variable must satisfy the following conditions:

$$E(z_i \varepsilon_i) = 0, \tag{4}$$

$$\operatorname{rank} E(z_i z_i') = q, \tag{5}$$

$$\operatorname{rank} E(z_i x_i') = p \tag{6}$$

Condition (3) implies that the instruments and the disturbance terms are orthogonal, i.e. uncorrelated. (4) implies the stability condition, i.e. the instruments are informative. Finally, condition (5) requires the number of instruments to be at least a large as the number of regressors ($q \geq p$), this is known in literature as the rank condition. Instrumental variables often consist of all the exogenous variables in the vector x_i and additional instruments that satisfy the conditions above.

The instrumental variable (IV) estimator β_{IV} is essentially a two-stage least squares estimator (2SLS). In the first step, the idea is to replace X by linear combinations of Z that approximate X as well as possible. This best approximation is obtained by regressing every column of X on the instruments Z. The fitted value of this regression are: $\hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X$. In the second stage, the IV estimator is obtained by regressing y on the exogenous covariate $\hat{X}$, resulting in the parameter estimate $b_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$. The validity of these instruments, i.e. the exogeneity of the instruments, can be tested using the methods proposed by Sargan (1958). Essentially, (3) can be tested by checking if z_i is uncorrelated with e_{IVi}. If the instruments are valid, one can proceed to use the Hausman (1978) test for exogeneity of the regressors. The test is based on the difference of OLS and the 2SLS estimators, when these yield the same results, exogeneity can not be rejected. This allows us to use OLS which is in general more efficient in an exogeneous setting and thus more desirable. Real data sets however, often contain outliers that can distort the OIV parameter estimates and its relevant tests. Therefore, it is important to look at estimators analogous to OIV that are robust to outliers in the same setting. This leads us to the rest of this section.

3.2 Robustification of the Ordinary Instrumental Variable

The robustification of the ordinary instrumental variable in this paper is similar to the approach proposed by Cohen Freue et al. (2013). We begin by introducing the robustification process. Relevant statistical theory and key metrics with associated proofs are discussed afterwards, where we use notation similar to Cohen Freue et al. (2013).

3.2.1 The Estimator

As the real data generating process is often unknown we will proceed by using sample data. Let $(\hat{\mu}, \hat{\Sigma})$ be the sample mean and covariance matrix based on a sample $(x_i, z_i, y_i)_{i=1}^n$, where y_i is univariate and x_i, z_i are multivariate with dimensions d_x, d_z. Now, the sample mean and covariance matrix can be decomposed as follows:

$$\hat{\mu} = (\hat{\mu}_x', \hat{\mu}_z', \hat{\mu}_y)' \qquad \text{and} \qquad \hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{xx} & \hat{\Sigma}_{xz} & \hat{\Sigma}_{xy} \\ \hat{\Sigma}_{zx} & \hat{\Sigma}_{zz} & \hat{\Sigma}_{zy} \\ \hat{\Sigma}_{yx} & \hat{\Sigma}_{yz} & \hat{\Sigma}_{yy} \end{bmatrix} \tag{7}$$

OIV can now be written as:

$$\hat{\alpha}_{OIV} = \hat{m}_y - \hat{m}_x \hat{\beta}_{OIV} \qquad \hat{\beta}_{OIV} = [\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx}]^{-1}[\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx}] \tag{8}$$

As can be seen from expression (7), OIV is a function of the sample mean and sample variance-covariance matrix. Because both the sample mean and variance-covariance matrix are non-robust estimates, OIV is extremely sensitive to outliers (Krasker & Welsch, 1985). This is especially the case when the function has an unbounded influence function and zero breakdown point. The two terms will be explained more in depth in the following subsections, as these are important terms in literature regarding RIV estimation.

We robustify the OIV by replacing the sample estimators $(\hat{\mu}, \hat{\Sigma})$ in Equation 8 by a robust multivariate location and S-estimator (m, S). Now, the robust estimator is given by:

$$\hat{\alpha}_{RIV} = m_y - m_x \hat{\beta}_{RIV} \qquad \hat{\beta}_{RIV} = [\hat{S}_{xz}\hat{S}_{zz}^{-1}\hat{S}_{zx}]^{-1}[\hat{S}_{xz}\hat{S}_{zz}^{-1}\hat{S}_{zx}] \tag{9}$$

Now that the form of the robust estimator is given, we continue by discussing and analyzing its most important robustness measures we start off by discussing the equivariance and consistency of the RIV estimator.

3.2.2 Equivariance and Consistency

Equivariant translation means that a translation of input features results in an equivalent translation of outputs. For robust estimators, and estimators in general, this is a desirable property. When the form of structure for the (in)dependent variable changes, we want the estimator to change accordingly. This allows us to make meaningful inferences even when the data were to be transformed. We now proceed by establishing the equivariance of the RIV estimator.

Lemma 3.1 *(Cohen Freue et al., 2013) Let $v_i = (x_i', z_i', y_i')'$ and the transformation of the data denoted as $(\tilde{x}_i', \tilde{z}_i', \tilde{y}_i')' = Av_i + b$. Where*

$$A = \begin{bmatrix} P' & 0 & 0 \\ 0 & Q & 0 \\ \gamma' & 0 & \eta \end{bmatrix} \qquad b = \begin{bmatrix} 0 \\ 0 \\ \delta \end{bmatrix} \tag{10}$$

Moreover, let (m,S) be the multivariate location and scatter S-estimator derived from the sample v. Analogously, we let $(\tilde{m}, \tilde{S})$ be the multivariate location and scatter S-estimator of the transformed sample $\tilde{v}$. Then, we can derive the RIV estimates based on $\tilde{v}$:

$$\tilde{\alpha}_{RIV} = \tilde{m_y} - \tilde{m_x}\tilde{\beta}_{RIV} \tag{11}$$

$$\tilde{\beta}_{RIV} = [\tilde{S}_{xz}\tilde{S}_{zz}^{-1}\tilde{S}_{zx}]^{-1}[\tilde{S}_{xz}\tilde{S}_{zz}^{-1}\tilde{S}_{zy}] \tag{12}$$

Now, given that S-estimators are affine equivariant, it holds that

$$(\tilde{m}, \tilde{S}) = (Am + b, ASA') \tag{13}$$

From (13) *we can derive the following:*

$$\tilde{m}_y = \eta m_y + \gamma' m_x + \delta \tag{14a}$$

$$\tilde{m}_x = P' m_x \tag{14b}$$

$$\tilde{S}_{xz} = P' S_{xz} Q \tag{14c}$$

$$\tilde{S}_{zz} = Q' S_{zz} Q \tag{14d}$$

$$\tilde{S}_{zy} = Q'[S_{zx}\gamma + \eta S_{zy}] \tag{14e}$$

Combining (11), (12) and (14), *we get*

$$\tilde{\alpha}_{RIV} = \eta \hat{\alpha}_{RIV} + \delta$$

$$\tilde{\beta}_{RIV} = (P^{-1})(\eta \hat{\beta}_{RIV} + \gamma)$$

which completes the proof and shows that RIV satisfies the equivariance property.

Davies (1987) shows in his paper, that under certain regularity conditions, the multivariate S-estimator used to compute the RIV estimator is consistent. RIV directly inherits this property. Cohen Freue et al. (2013) show that even under weaker conditions - when possibly the multivariate S-estimator is not consistent - the RIV remains consistent, given that z and ε are independent.

3.2.3 Influence function and Asymptotic Variance

In this section we show that the RIV estimator is B-robust, i.e. its associated influence function (IF) is bounded, consistent and asymptotically normally distributed. Before we proof this statement, we first explain the concept and the importance of B-robustness and the IF. Influence functions can be interpreted as analytical tools used to investigate the effect on a statistic after an adjustment has been made on individual observations. If the influence function happens to be unbounded, then outliers may cause potential problems for the reliability of the estimator. Before giving the functional definition of the IF, it is sensible to first introduce the concept from a sample point of view. To measure the robustness of a given estimator, we add a $x_0 \in \mathbb{R}$ to the sample x = $(x_1, ..., x_n)$. Now, we can define the *sensitivity curve* (SC) for an estimate $\hat{\theta}$:

$$SC(X; x_1, ..., x_n, \hat{\theta}) = \frac{\hat{\theta}(x_1, ..., x_n, x_0) - \hat{\theta}(x_1, ..., x_n)}{\frac{1}{n+1}} \tag{15}$$

The factor $\frac{1}{n+1}$ denotes the standardization, with the purpose that the SC can be plotted and compared against different sample sizes. As the distribution F is assumed to be known approximately, we are interested in how the estimator behaves in the “neighborhood” of a distribution F_0. Ricardo et al. (2019) show that using the following neighborhood function is the easiest way to examine the behaviour of $\hat{\theta}$.

$$\tilde{F}_{\theta,\varepsilon} = (1 - \varepsilon) F_\theta + \varepsilon G \tag{16}$$

ε denotes the so-called, contamination proportion, i.e. the proportion of outliers in the whole data set. Here the outliers are assumed to follow the distribution G. F. R. Hampel (1974) shows that the IF of an estimator is an asymptotic version of the sensitivity curve:

$$IF(x; F_\theta; T) = \lim_{\varepsilon \to 0} \frac{T(\tilde{F}_{\theta,\varepsilon}) - T(F_\theta)}{\varepsilon} \tag{17}$$

For a given functional T and a model F_θ, IF is a function of x. Given the background of the IF, we will now proceed by establishing the influence function of the RIV estimator and its asymptotic variance. We follow the notation used by Cohen Freue et al. (2013).

Theorem 3.2 *(Cohen Freue et al., 2013) Let $(x_i, z_i, \varepsilon_i)$ follow the distribution H, and let (m_H, S_H) denote the functionals of the S-estimator. Now, the functional of RIV, $T(H) = (a(H), b(H))$ can be denoted as:*

$$\begin{aligned} a(H) &= m_{y,H} - m'_{x,H} b(H) \\ b(H) &= [S_{xz,H} S^{-1}_{zz,H} S_{zx,H}]^{-1} [S_{xz,H} S^{-1}_{zz,H} S_{zy,H}] \end{aligned}$$

using definition (17) we can obtain the IF of T at (x, y, z) and H as:

$$\begin{aligned} &IF(x, z, y; a, H) = IF_y - IF'_x b(H) - m'_{x,H} IF(x, z, y; b, H) \\ &IF(x, z, y; b, H) = C^{-1}(H)[IF_{xz} S^{-1}_{zz,H} D(H) - S_{xz,H} S^{-1}_{zz,H} IF_{zz} S^{-1}_{zz,H} D(H) + S_{xz,H} S^{-1}_{zz,H}(IF_{zy} - IF_{zx} b(H))] \end{aligned}$$

Where $C(H) = [S_{xz,H} S^{-1}_{zz,H} S_{zx,H}]$, $D(H) = [S_{zy,H} - S_{zx,H} b(H)]$, $IF_r = IF(x, y, z; m_r, H)$. and $IF_{ij} = IF(x, y, z; S_{ij}, H)$ for $r, i, j \in (x, y, z)$.

The RIV influence function has the desirable property that it is B-robust, when using a bounded influence S-estimator. Using the IF, we now proceed to derive the asymptotic covariance matrix:

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

H_n denotes the empirical joint distribution of $(x_i, z_i, y_i)_{i=1}^{n}$

3.2.4 Breakdown point

The breakdown point (BP) of an estimator $\hat{\beta}$ is the largest amount of contamination, expressed in the proportion against the whole data set, that the data may contain such that it still provides sufficient information about the distribution of the non-outlier points. We establish that, given that the instrumental variables stay valid under contamination, the RIV estimator inherits the breakdown point of the multivariate location and scatter S-estimators.
To fix ideas, without loss of generality, we consider the exactly identified model (i.e. number of observations is equal to the number of parameters; $q = p$). Now, with mathematical manipulation the estimator in Equation 12 can be rewritten as $\beta_{RIV} = S^{-1}_{zx} S_{zy}$. The RIV has two different ways of breaking down from here. First, it could break down due to singularity in S_{zx}. Second, it could break down in either m_x, m_y, or S_{zy} because the vector becomes unbounded.

Wagenvoort and Waldmann (2002) show in their paper that in order to state relevant points on the BP of IV estimator, the used IV should be a valid instrumental variable. The IV estimator breaks down if the contamination of the data results in invalid instrumental variables.

A desirable property of the RIV estimate, when computed by the S-estimator and satisfying condition 4, is that it achieves the maximal BP, i.e. 50%.

3.2.5 Weighted Instrumental Variables Estimator

For the exactly identified model, we prove that RIV can be rewritten as a weighted instrumental variables estimator following notation by Cohen Freue et al. (2013). To acquire the weights we use the Mahalonobis Distance (MD). We illustrate the definition of the MD by an easy example: let $x = (x_1, \dots x_N)$ from a set with mean $\mu = (\mu_1, \dots \mu_N)$ and covariance matrix S. Gives this, the MD is defined as:

$$MD(x) = \sqrt{(x-\mu)' S^{-1} (x-\mu)} \tag{19}$$

Hence, the MD can be interpreted as the Euclidian distance corrected for the correlation between variables. This also shows the usefulness in the context of the RIV estimator due to its correlated nature. The weights for RIV are acquired by the inverse Mahalonobis Distance (MD) of the observation to the location estimate m adjusted for the scatter estimate S. The applied weight for the observations (x, z, y) downweights all outlying observations (x, y, z) and therefore acts as natural outlier flagging tool. We can detect outliers by plotting the weights of every observations produced by the RIV, against the MDs. A cut-off level can be obtained by using a Chi-Squared distribution with degrees of freedom equal to the dimension of the data. We now prove that RIV can be rewritten as a weighted instrumental variables estimator.

Lemma 3.3 *(Cohen Freue et al., 2013) RIV defined using an S-estimator with mean m and covariance S as shown in Equation 11 and Equation 12 can be rewritten as: $(\hat{\alpha}_{RIV}, \hat{\beta}_{RIV}) = (\tilde{Z}'\Omega\tilde{X})^{-1}\tilde{Z}')$. Here $\tilde{X}$ and $\tilde{Z}$ are $(n \times p)$ matrix with ith rows equal to $(1, x_i')$ and $(1, z_i')$, respectively. Moreover, Ω is the diagonal weighting matrix with i^{th} element $\omega(d_i)$, here d_i is the square root of the MD, and $\omega(.)$ is the same as the weighting matrix used by Cohen Freue et al. (2013)*

3.3 Estimation of Models with both Continuous and Dummy Covariates

As mentioned earlier in the literature section, the presence of dummy covariates can possibly make the computation of the S-estimator infeasible. The presence of dummy covariates can be captured in an expanded version of the linear regression model in Equation 2:

$$y_i = \alpha + c_i'\beta_1 + x_i'\beta_2 + \varepsilon \text{for } i = 1,\dots, \text{n} \tag{20}$$

Where c_i is the vector of exogenous dummy variables, x_i is a vector containing both exogenous and endogenous continuous variables. C and X are matrices with row i equal to c_i and x_i, respectively. The source of the problem lays in the fact that the S-estimator can not be computed in an exact manner (Stromberg, 1993), thus subsampling procedures are necessary. However, in the presence of exogenous dummy variables, the subsampling procedures yield collinear samples from which the computation of the scale is infeasible.

We follow the algorithm proposed by Maronna and Yohai (2000) and Cohen Freue et al. (2013). The algorithm used, $L_1 - RIV$, estimates the coefficients of the dummy covariates with a combination of the original RIV estimates and M-estimators. The iterative procedure can be stated as follows:

$$\hat{\beta}_2^{(k)} = g(X, Z, y - C\hat{\beta}_1^{(k-1)}) \tag{21}$$

$$\hat{\beta}_1^{(k)} = L_1(C, y - X\hat{\beta}_2^{(k)}) \qquad \text{for } 1 \leq k \leq K \tag{22}$$

where $g(X, Z, t)$ denotes the application of the RIV estimation for the data (X, Z, t), and $L_1(c, t)$ are the coefficient estimates of the aforementioned L_1 regression of t on C. Following the proposal for the initiation of the algorithm of Maronna and Yohai (2000), the effect of the dummy variables are first removed from all the variables, including the instruments. Then, we proceed by applying RIV on the modified variables to get the initial coefficient $\hat{\beta}_2^{(0)}$. Subsequently, the obtained estimate $\hat{\beta}_2^{(0)}$ is used to recover the dummy variables using the L_1 regression, yielding $\hat{\beta}_1^{(0)}$.

3.4 Simulation

The simulation conducted in this paper is two-fold. We start by comparing the RIV estimates to OIV estimates in the model containing only continuous covariates. Afterwards, we conduct a similar simulation study, but now the model also contains dummy covariates. This allows us to put the L_1 regression in practice.

The simulation we conduct allows for an effective evaluation of the difference between the OIV estimator and the RIV estimator. We contaminate the data with different degrees and types, a detailed description follows in the next subsection. To be specific, we consider the following regression model:

$$y_i = \alpha + x'_{1i}\beta_1 + x'_{2i}\beta_2 + \varepsilon_i \text{for } i = 1, \ldots, \text{n} \tag{23}$$

where x_{1i} and x_{2i} denote respectively the continuous endogenous and exogenous covariates. Moreover, there also exists an instrumental variable z_i.

To conduct our simulation, we generate $R = 1000$ samples of size $n = 250$ for the random vector $(x_1, x_2, z, \varepsilon)$. These samples are drawn using a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. Here $\mu = (0, 0, 0, 0)$ and Σ is constructed such that all diagonal elements and all correlations in Σ are equal to zero. This is to ensure the use of IV based on conditions (3) and (4). The correlation between the endogenous variable and the error term is not equal to zero, and also for the correlation between the endogenous variable and the IV holds the same. We set $\alpha = 1$ and β_1, β_2 both equal to 2, this happens without loss of generality due to the equivariance property of the RIV described earlier.

3.4.1 Data contamination

Four different outliers are generated to make the simulation exhaustive: (i) in the dependent variable y, (ii) in the endogenous variables x_1, (iii) in the exogenous variable x_2, (iv) in the instrumental variables z. For each case, a proportion ϵ is randomly replaced by values that are drawn from either a $\mathcal{N}(\kappa, 0.01)$ for κ=1, 3, 5, 10 or a Cauchy random variable $C(0, 1)$ where ϵ ranges from 0 to 0.3 in 0.05 increments. This can be interpreted as asymmetrical and symmetrical contamination, respectively.

3.4.2 Assessing the estimator's performance

There are different ways to assess the performance of an estimator in literature. Mainly, we are interested in the behaviour of the variance and bias of the estimated coefficients, that are caused by the outliers. As the data contamination happens in two ways, i.e. symmetric and asymmetric, we also distinguish the performance measure for this criteria.

For the symmetric contamination, we assess the estimator performance by using the Monte Monte Carlo median squared errors (MedSE). The MedSE can be expressed as MedSE $=median_r \left\|\hat{\theta}^{(r)} - \theta\right\|^2$,. Intuitively, we evaluate the median of the norm, where the difference between the true parameters $\theta = (\alpha, \beta_1, \beta_2)$ and the estimated coefficients $\hat{\theta}$ are measured. The MedSE then equals the median over all r runs. For the asymmetric contamination, we use a similar method as to the symmetric contamination. Here we use the Maximum MedSE $= max_{k\in(1,3,5,10)} median_r \left\|\hat{\theta}_k^{(r)} - \theta\right\|^2$,. The intuition is similar, but instead of taking the median over only the runs, the maximum MedSE takes the value that is equal to the maximum value over all κ. The $\hat{\theta}_k^{(r)}$ denotes the estimated parameter of the rth simulation run, where the data is contaminated under a $\mathcal{N}(\kappa, 0.01)$.

Moreover, to examine the results of using estimators different than the S-estimator, we include the computation and simulation of RIV computed using the Stahel-Donoho (SD)(Stahel, 1981) and minimum covariance determinant (MCD) (Rousseeuw & Leroy, 1987).

3.5 Application of RIV in Geology

Young (2020) shows that the contamination of data-sets with outliers adversely affect the performance of IV. The size and power of IV estimates are distorted, while the bias of IV relative to OLS increases. Thus, they reduce the power of the initial reasons to choose OIV over OLS in an endogenous setting. He shows this for 31 prominent published papers in the journals of the American Economic Association.

Due to the results of Young (2020), we now test the performance of the RIV estimates compared to the IV estimates of a real data-set under contamination. The data we use in this section is obtained from Fuller (1987). The data-set contains information on 62 Alaskan Earthquakes in the period 1969-1978.

Fuller translates the information to the following regression:

$$y_i = \alpha + \beta x_i + \varepsilon \tag{24}$$

Where y denotes the amplitude of the surface waves of the earthquake and x the logarithm of the seismogram amplitude of longitudial body waves. The real explanatory variable is earthquake strength, but as this is not observable the paper proceeds by using $x = x^* + u$. Here, x^* is the earthquake strength and u represents the measurement error. In this example, the endogenous covariate is therefore x. The logarithm of maximum seismogram trace amplitude (Z), is used as the instrumental variable.

For the application, we begin by detecting the outliers using the method described in Section 3.5. Then, we proceed by comparing the estimated coefficients of RIV to OIV and examine whether there are significant differences. The RIV-estimator down-weights all outliers and thus should help to exhibit the true relationship between the variables.

4 Results

This section covers the results for both the simulation and the application of the RIV based on Fuller (1987). We begin by evaluating the performance of both the RIV and OIV under asymmetric and symmetric contamination. Afterwards, we cover the results of $L1 - RIV$ and the RIV-estimator using different estimators.

4.1 Simulation Results

Figure 1 shows the performance of RIV compared to OIV under asymmetric and different levels of contamination. We find that, as expected, RIV's Maximum MedSE stays fairly constant for contamination levels up to 25%. After the 25% it undergoes a slight increase, but nothing compared to the Maximum MedSE of OIV. The OIV Maximum MedSE increases steeply when the level of contamination increases. It can also be seen that for contamination in the instrumental variable the performance of OIV stays fairly constant, just as RIV. This can be explained by the low level of correlation between the instrument and the endogenous variable, which in turn results in effects that are smaller when the contamination increases.

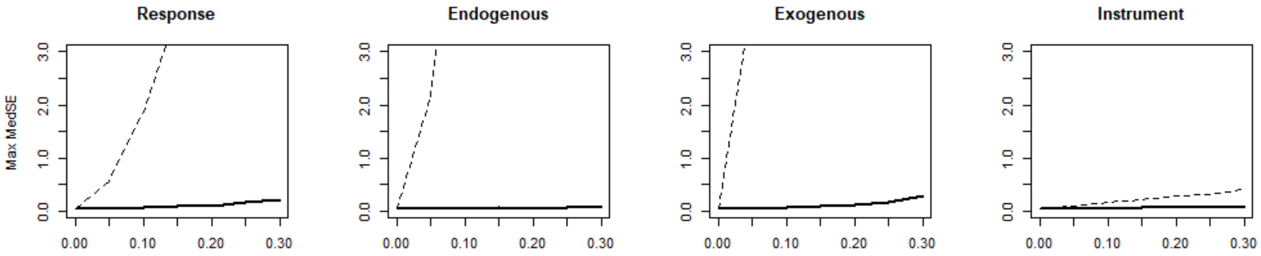

Figure 1: Results of the simulation for asymmetrical contamination. Figure contains the plots of the Response, Endogenous, Exogenous and Instrumental Variables under different level of contaminations for the RIV (Bold Line) and OIV (Dotted Line). On the Y-axis the Maximum MedSE is calculated using contamination under a $\mathcal{N}(\kappa, 0.01)$ distribution for κ=1, 3, 5, 10. On the X-axis, the different levels of contamination are displayed in proportions.

Figure 2 shows the performance of RIV compared to OIV under symmetric and different levels of contamination. We find that, in line with the asymmetric contamination, the OIV's MedSE increases steeply with the level of contamination. The MedSE of RIV stays fairly constant over all levels of contamination up to 25%. Moreover, we also observe that the increase in MedSE for both RIV and OIV is less rapid in the instrument compared to other covariates. The explanation holds the same reasoning as given for the asymmetric case.

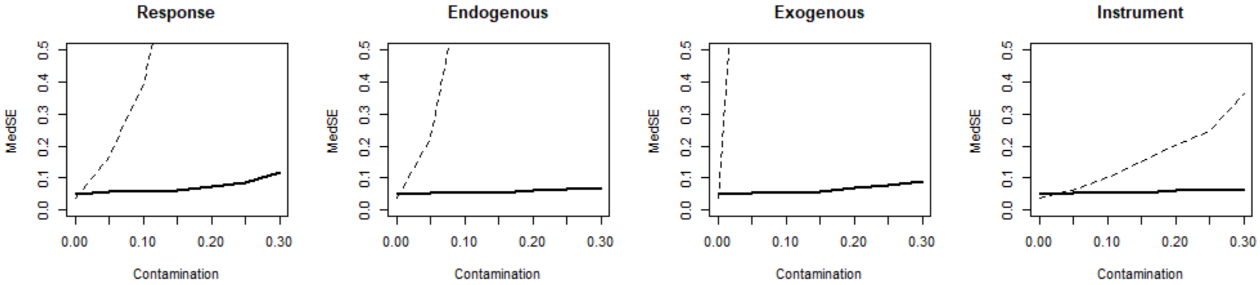

Figure 2: Results of the symmetrical contamination. Figure contains the plots of the Response, Endogenous, Exogenous and Instrumental Variables under different level of contaminations for the RIV (Bold Line) and OIV (Dotted Line). On the Y-axis the MedSE is calculated using contamination under a Cauchy random variable $C(0, 1)$ distribution. On the X-axis, the different level of contamination are displayed in proportions.

Figure 3 shows the performance of RIV under symmetric contamination using different estimators. RIV estimation using S-estimator (solid) exhibits slightly better results compared to MCD (dotted) and SD (dot-dash). Although overall, all estimators used instead of the S-estimator showcase much better results under contamination compared to the OIV estimator.

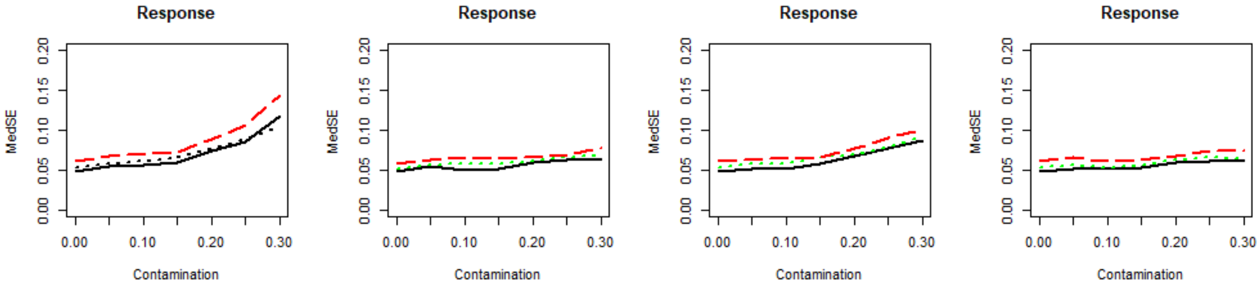

Figure 3: Results of the simulation for symmetrical contamination, where different estimators are used for the RIV estimation. Figure contains the plots of the Response, Endogenous, Exogenous and Instrumental Variables under different level of contaminations for the RIV(Bold Line), MCD(Dotted Line) and SD(dot-dash).

Figure 4 shows the performance of the MedSE of the $L1$-RIV(dashed), OIV (solid), MCD (dotted) and SD (dot-dash). We find that the results are, as expected, very similar to that of the model containing only continuous covariates. Moreover, we conclude that the OIV estimator is very sensitive for low-level contamination as well. Regardless of the estimator used in the computation for RIV, we find that the RIV estimator stays fairly constant for all types against different contamination levels. However, there is a minimal efficiency trade-off when the model is not contaminated.

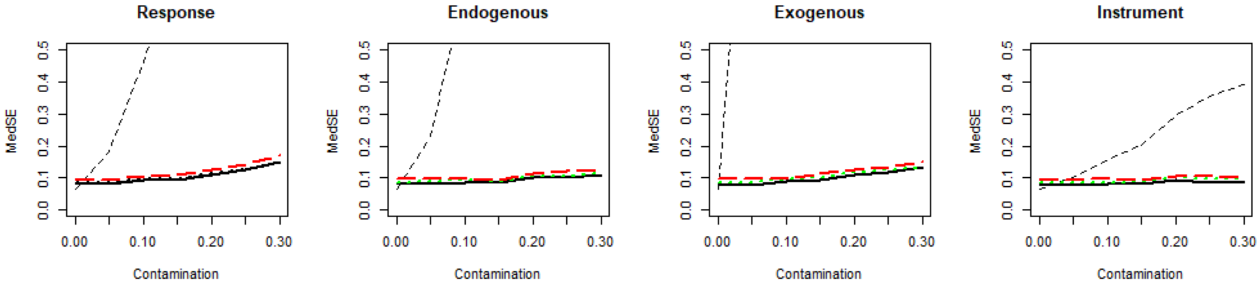

Figure 4: Results of the simulation for symmetrical contamination using $L1$-RIV, which includes the use of dummy variables. Set-up is identical to that of Figure 1.

4.2 Application of RIV

Figure 5 shows the use of RIV for detecting outliers. RIV downweights all points that are far away from the bulk of the data. Using the Chi-Squared distribution with the freedom of the dimension, we can use the data set to obtain a cut-off-level for classifying outliers. This gives RIV the natural ability to flag outliers that are present in the data set and find the true relationship between variables. In Figure 5 the flagged outliers are bolded out, and are the observations 28, 25, 60 and 54. Moreover, in the right subfigure the weights RIV gives to each observation are set out against the observation's distance to the bulk of the data (d_i). Now that we have shown that certain observations are flagged as outliers, we can obtain a more accurate representation of the coefficients downweighting these outliers using RIV estimation.

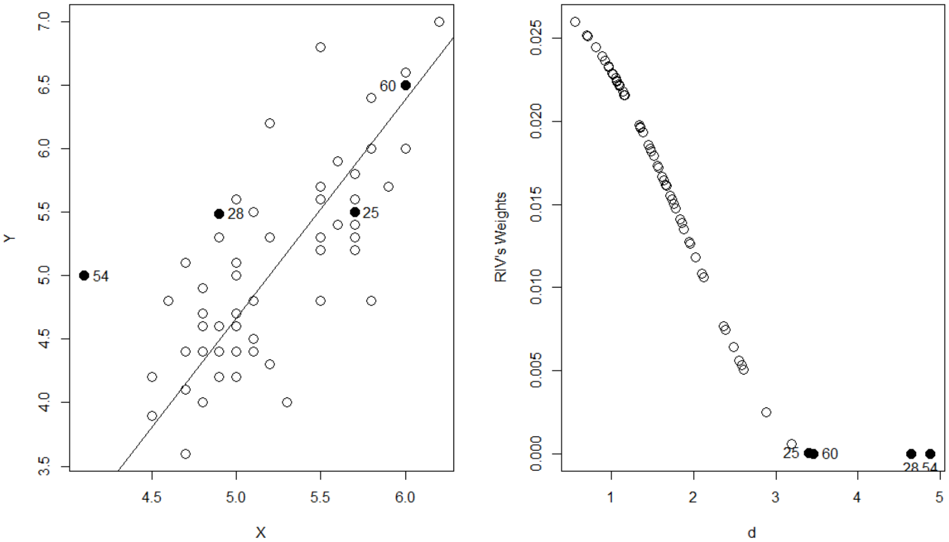

Figure 5: Left subfigure contains the plot of the Y variable (amplitude of the Surface Waves) against the X variable (the logarithm of the seismogram amplitude). The outliers have been bolded out. Right subfigure contains the RIV weights of the observations against the square root of the observation's MD from the bulk of the data d_i.

Table 1: Coefficient table for the estimated regression as described in Section 3.5 for both RIV and OIV.

		Coeff.	Std. Err.	t-value	p-Value
RIV	Intercept	-3.95	0.98	-4.03	1.63e-04
	x	1.72	0.19	9.20	4.48e-13
OIV	Intercept	-4.29	1.11	-3.85	2.89e-04
	x	1.80	0.21	8.43	8.96e-12

Table 1 reports the coefficients for the intercept and endogenous variable, standard error, t-value and p-value, for both the RIV estimation and the OIV estimation, respectively. We find that the significance of the intercept and coefficient does not change. That is, they both stay significant based on the p-Value with $\alpha < 0.5$. However, we find that the coefficient of the endogenous x-variable does change from 1.80 to 1.72, indicating a less strong impact of the earthquake strength on the amplitude of the surface waves. This result can be explained by the fact that RIV downweights the outliers that we have flagged earlier.

5 Conclusion

In this paper, we have researched the variance and bias of the RIV estimator against OIV for different types and levels of contamination. Overall, we conclude that based on theory and on acquired empirical results the RIV outperforms the OIV in the presence of outliers and can thus be useful for researchers who want to robustify their results. Moreover, we have proved that the RIV has several desirable properties: (i) it is equivariant and consistent under weak regularity conditions; (ii) the influence function is B-robust, i.e. it is bounded and under regularity conditions asymptotically normally distributed; (iii) it has a Breakdown point(BP), which reaches 50% asymptotically ;(iv) RIV can be rewritten as a weighted instrumental variables estimator.

Moreover, we conducted a simulation study with $r = 1000$ samples and $n = 250$ observations drawn from a multivariate normal distribution. A proportion ϵ of the data was randomly replaced by values that are drawn from either a $\mathcal{N}(\kappa, 0.01)$ for κ=1, 3, 5, 10, or a Cauchy random variable $C(0,1)$ where ϵ ranges from 0 to 0.3 in 0.05 increments. This can be interpreted as asymmetrical and symmetrical contamination, respectively. We assessed the performance in terms of MedSE for symmetrical contamination and Max. MedSE for the asymmetrical contamination. The results of the simulation study were in line with the theory described in the methodology. As expected, the IV's assessment values increased steeply with the level of contamination, whereas in contrast the RIV remained relatively constant over all contamination levels. Therefore, suggesting to be more robust against outliers.

Finally, we also used the RIV on real earthquake data from Fuller (1987). Here, we found that RIV performs consistently, i.e. the standard errors and estimated parameters stay fairly constant over contamination. Moreover, when contaminating the data set with outliers as with the simulation, we found that the results were consistent with that of the simulation.

For further research we suggests researchers to shed more light on certain aspects of this study. By construction the S-estimator downweights all outliers in the space spanned by the incorporated variables. The caveat here, is that good leverage points also get downweighted, which is generally undesirable as they can reduce the variance of the RIV estimates. Therefore, we suggest for further work to develop an efficient manner to solve this problem. Moreover, as instrumental variables in general are dependent on the validity of the instruments, it is necessary that validity tests that exist for OIV become adjusted for RIV. This ensures for a correct measure when testing for the validity of the instruments in a highly-leveraged environment.

References

Box, G. (1953). Non-normality and tests on variances. *Biometrika*, *40*(6), 318-335.

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

Cohen Freue, G., Ortiz-Molina, H., & Zamar, R. (2018). Retrieved from `https://cran.r-project.org/web/packages/riv/index.html`

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

Fuller, W. A. (1987). Measurement error models. *Wiley*.

Hampel, F. (1968). Contributions to the theory of robust estimation. *Ph.D. Thesis*.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383-393.

Hausman, J. (1978). Specification tests in econometrics. *Econometrica*, *46*(6), 1251-1271.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann Math Stat*, *35*(6), 73-101.

Huber, P. J. (1978). The behavior of maximum likelihood estimates under nonstandard conditions. *Econometrica*, *46*(6), 1251-1271.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics, second edition.* Wiley press.

Krasker, W., & Welsch, R. (1985). Resistant estimation for simultaneous-equations models using weighted instrumental variables. *Econometrica*, *53*(3), 1475-1488.

Maronna, R., & Yohai, V. (2000). Robust regression with both continuous and categorial predictors. *Journal of Statistical Planning and Inference*, *89*, 197-214.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ricardo, A., Maronna, R., Douglas, M., Yohai, V., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with r), 2nd edition.* Wiley press.

Rousseeuw, P., & Leroy, A. (1987). Robust regression and outlier detection. *Wiley*.

Sargan, J. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, *26*(3), 393-415.

Stahel, W. (1981). Robust estimation: Infinitesimal optimality and covariance matrix estimators. *Journal of the American Statistical Association*, *69*(346), 383-393.

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

Tukey, J. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, *46*(6), 448-485.

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

Young, A. (2020). Consistency without inference: Instrumental variables in practical application.